No cover image available

The Oxford Handbook of Digital Ethics

Carissa Véliz (ed.)

https://doi.org/10.1093/oxfordhb/9780198857815.001.0001

Published: 2021 **Online ISBN:** 9780191890437 **Print ISBN:** 9780198857815

CHAPTER

Ethical Issues with Artificial Ethics Assistants

Elizabeth O'Neill, Michał Klincewicz, Michiel Kemmer

https://doi.org/10.1093/oxfordhb/9780198857815.013.17 Pages C17.S1–C17.N26

Published: 20 October 2022

Abstract

This chapter examines the possibility of using artificial intelligence (AI) technologies to improve human moral reasoning and decision-making. The authors characterize such technologies as artificial ethics assistants (AEAs). The authors focus on just one part of the AI-aided moral improvement question: the case of the individual who wants to improve their morality, where what constitutes an improvement is evaluated by the individual's own values. The authors distinguish three broad areas in which an individual might think their own moral reasoning and decision-making could be improved: one's actions, character, or other attributes fall short of one's values and moral beliefs; one sometimes misjudges or is uncertain about what the right thing to do is, given one's values; or one is uncertain about some fundamental moral questions or recognizes a possibility that some of one's core moral beliefs and values are mistaken. The authors sketch why one might think AI tools could be used to support moral improvement in those areas and distinguish two types of assistance: preparatory assistance, including advice and training supplied in advance of moral deliberation, and on-the-spot assistance, including on-the-spot advice and facilitation of moral functioning over the course of moral deliberation. Then, the authors turn to ethical issues that AEAs might raise, looking in particular at three under-appreciated problems posed by the use of AI for moral self-improvement: namely, reliance on sensitive moral data, the inescapability of outside influences on AEAs, and AEA usage prompting the user to adopt beliefs and make decisions without adequate reasons.

Keywords: artificial intelligence, artificial ethics assistant, artificial ethics advisor, moral cognition, moral enhancement, moral decision-making

Subject: Moral Philosophy, Philosophy

Series: Oxford Handbooks

Introduction

Flora is in a restaurant, chatting with a group of friends, when the conversation turns to climate change, animal welfare, and diet. By the time the waiter asks her what she'd like for lunch, Flora finds herself at a loss, unsure what to order. The locally sourced roast beef? The chicken, likely from a factory farm? Salmon? The vegan quinoa? Usually, she orders whatever sounds most appetizing. But she's just heard a lot of strongly held, conflicting views on what food choices are best for the environment and the well-being of animals. She'd like to do the right thing, and she cares about animals and the environment. However, she isn't sure what option would be best. For one thing, she's missing a lot of information that seems relevant for her decision. She wonders whether fish feel pain, how the chickens were treated, and which dish took the most resources to produce. She'd also like to know what the best arguments are for and against positions like vegetarianism. After listening to some of her friends, she's even begun to question some of her deeper values and normative beliefs like her assumption that ecosystems have only instrumental value.

What to have for lunch is just one of the many ordinary purchasing and consumption decisions we face every day. In many of these situations, we are missing information that we think might be relevant, and we haven't had a chance to fully think through the question, yet we have to make a choice. With the advent of big data and recent developments in artificial intelligence (AI) research, is there some way that computer systems could help us with this kind of decision?

Recently, a number of philosophers and computer scientists have proposed that technologies employing AI could be used to improve human moral reasoning and decision-making (Borenstein and Arkin 2016; Giubilini and Savulescu 2018; Klincewicz 2016; Savulescu and Maslen 2015; Formosa and Ryan 2021; Lara 2021; Sinnott-Armstrong and Skorburg 2021).[1] We will focus on just one part of this AI-aided moral improvement question: the case of the individual who wants to improve their morality, where what constitutes an improvement is evaluated by the individual's own values.[2] We will refer to the technology of interest in this chapter as the 'artificial ethics assistant' (AEA)—an AI system that is designed and used with the aim of improving human moral cognition.[3] We look primarily at the use of AEAs to influence purchasing and consumption decisions made by individuals; we set aside discussion of the many other possible uses of AI technologies to aid moral decision-making.[4]

If Flora were presented with an AEA, with the promise that it would help her make more moral consumer decisions, or perhaps more specifically, decisions that better align with her core moral values, would she have reason to use it? In comparison with other approaches to moral improvement, attempts to use AI technology to improve moral cognition make several potential ethical questions and problems salient. This chapter examines a selection of such problems. We begin by distinguishing three broad areas in which an individual might think their own moral reasoning and decision-making could be improved. We then sketch why one might think that AI tools could be used to support moral improvement in those areas. Finally, we discuss some of the ethical issues that AEAs might raise, looking in particular at three under-appreciated problems posed by the use of AI for moral self-improvement, namely, reliance on sensitive moral data, the inescapability of outside influences on AEAs, and AEA usage prompting the user to adopt beliefs and make decisions without adequate reasons.

Moral self-improvement

When considering the topic of artificial ethics advisors, there is an essential preliminary question to ask: what would it mean to improve one's own moral reasoning and moral decision-making capacities?[5] Given conflicting conceptions of morality, different people have ideas of moral improvement that diverge in substance—one person thinks moral improvement involves increasing compassion; another person thinks it involves increasing devotion to family or loyalty to the country.[6] One can distinguish several aspects of moral cognition that an individual with nearly any set of values might believe could be improved.[7] One may think:

(1) One's actions, character, or other evaluable states or attributes[8] fall short of one's values and what one believes one ought to do. One admits one is sometimes hypocritical or that one suffers from weak self-discipline. For instance, even if Flora concluded that she ought to be a vegetarian, she may be so tempted by the thought of the delicious roast beef sandwich that she orders this option despite her moral belief.

(2) One sometimes misjudges or is uncertain about what the right thing to do is, given one's values, in a particular situation. Pressure or lack of time can lead to logical errors or to a failure to consider all the information that one ordinarily believes is relevant. One might be missing information or one might have false beliefs about the non-normative features of a particular situation (perhaps Flora incorrectly thinks that the locally sourced beef has a smaller carbon footprint than imported salmon) or about relevant non-normative features of the world more generally (such as whether fish can suffer as much as farm animals). It could be due to thought patterns that one considers to be biased —Flora may tend to have more sympathy for cows than chickens yet view this tendency as a bias. Even if one is confident in one's moral beliefs and values, there is room for things to go wrong when applying them to a particular case and forming a judgement about what to do.

(3) One may also be uncertain about some fundamental moral questions or one may recognize a possibility that one is mistaken in some of one's core moral beliefs and values (or in one's

understanding of them).[9] Individuals sometimes look back and think that they have been mistaken about what is right and wrong; people's moral priorities sometimes shift over time. Listening to her friend describe the plight of factory farm animals, Flora might come to think that she has made a moral mistake in the past by weighting human welfare as so much more important than animal welfare.

If one sets out to improve one's morality, then, one may do so with the aim of improving in any of these three areas—roughly, one may aim to align actions or character with moral judgements, aim to align moral judgements about particular cases with one's core moral commitments, or aim to either improve or better understand one's core moral commitments. There are many traditional tools for moral self-improvement that address aims (1)–(3). These include reflection, prayer and other religious practices, the study of texts, soliciting and following the advice of others, imitating people we admire, building up habits, structuring our environments to avoid temptations, and so on. More recently, philosophers have suggested the possibility of moral bioenhancement—the use of biomedical technologies, such as pharmacological, genetic, or neural interventions, for moral enhancement (Persson and Savulescu 2012). An AEA, if it performed as promised, would add a different kind of tool for moral self-improvement. At the same time, the use of AI for moral improvement poses a distinct set of potential ethical problems. In this chapter, we focus primarily on AEAs that could be used to advance aim 2, leaving discussion of the other aims for future work.

How could AI technology be used to improve moral cognition?

AI, interpreted broadly, is already being used to support reasoning and decision-making in a variety of contexts: for instance, to guide decisions about whether to admit patients into intensive care; to assist predictions about how likely a defendant is to re-offend; and to assist predictions about whether one investment or another is more likely to produce a profit (Phillips-Wren 2012). Furthermore, AI is used in e-commerce, where product rankings or chatbots help customers in making decisions. AI is also used as part of expert systems that automate decisions for pilots, doctors, and in other domains (Liebowitz 2019). Given these already existing systems, it is not implausible that the use of AI to supplement reasoning and decision-making will gradually encroach on more canonically moral domains in the near future. It is also possible that, as some philosophers have been advocating, efforts will be made to develop AI systems explicitly designed to support human moral capacities (see Whitby 2011).

Perhaps the most technologically feasible version of a personal AEA in the near term would be a recommender system that takes into account factors that the user considers moral. Recommender systems are at the heart of popular music streaming services, such as Spotify, and online shopping platforms, such as Amazon. One can easily imagine similar systems that generate purchasing or consumption recommendations in a process that incorporates information about users' values.[10] For instance, if one wants to avoid listening to White supremacist bands or to songs with misogynistic lyrics, one can imagine a music streaming recommendation system accommodating this preference.[11]

There are also a number of systems that provide users with information pertinent to advancing particular kinds of values within specific contexts, such as apps that supply information about restaurant options,[12] household products,[13] and clothing companies (Hansson 2017). And there are robo-advisor systems that make investment recommendations that take into account whether the user wants to avoid investing in particular industries such as in weapons or tobacco. However, no existing systems allow the user to personalize recommendations or information on the basis of moral values in a fine-grained way.

With regard to more advanced systems that could supply personalized ethics assistance, there are many technical questions about how the systems could acquire information about users' values. Here are some possibilities:

- the user completes surveys in which they indicate what they take to be their priorities, which considerations they believe are morally relevant in particular situations (c.f. Sinnott-Armstrong and Skorburg 2021: 9–14), or what their moral judgements are about particular scenarios (Giubilini and Savulescu 2018: 174–175). One might think of these surveys as much more elaborate versions of the Moral Foundations Questionnaire (Graham et al. 2013) or the questions asked in the Moral Machine experiment (Awad et al. 2018), which uses thought experiments to find patterns underlying moral judgements;[14]

- the AI system asks the person specific questions to fill in gaps in the system's model of the person's moral views;
- the AI system observes the person's choices and behaviour and infers desires, preferences, and values on the basis of them (Etzioni and Etzioni 2016);
- the AI system observes the person's choices and behaviour together with the choices and behaviour of others and infers desires, preferences, and values on the basis of correlations at the population level (Giubilini and Savulescu 2018).
- the person gives the AI system feedback on its judgements and recommendations, for example, as in some forms of inverse reinforcement learning (Ng and Russell 2000; Hadfield-Menell et al. 2016) or reward modelling (Leike et al. 2018);

Some of these are quite futuristic; current efforts to do this sort of thing are very rudimentary. Nonetheless, it may eventually be possible to represent much of an individual's moral world view outside of that person's mind.

Suppose, then, that systems have been created that are capable of acquiring a picture of the user's moral functioning and world view (or some subdomain within these) for the purpose of helping the user to improve their moral reasoning and decision-making. Such systems could conceivably influence an individual's moral cognition in a number of ways. One distinction is between preparatory assistance, on the one hand, which is any influence on the user in advance of their consideration of the case or moral question of interest, and, on the other hand, on-the-spot content-specific assistance with particular moral questions or cases.[15]

At the *preparatory* stage of assistance, the AEA might play an advisory role or a training role. When giving *on-the-spot* assistance, during the duration of the user's deliberation on a particular question, the system could play an advisory role or it could play a function facilitation role. In an *advisory role*, the AEA provides information and recommendations, offered for consideration within the user's reasoning process. In advance, the advising AEA could give information about the process of moral reasoning, for instance, and recommendations on how to prepare to make better decisions when one faces a case or moral question; on the spot, the advising AEA could give information and recommendations specific to particular cases or particular moral questions such as about what to value. A *trainer* supplies training that permits the user to improve some moral capacity; this role is specific to preparatory assistance.[16] *Function facilitation* occurs on the spot, over the course of one's consideration of a particular moral question, when the system helps one exercise some capacity such as the detection of morally relevant features of a situation or generation of possible options for action.

Preparatory assistance: Advice and training

One type of preparatory assistance is procedural assistance, which aims to improve one's moral deliberation processes (Paulo 2018).[17] Schaefer and Savulescu (2019: 73) propose that increasing 'logical competence, conceptual understanding, empirical competence, openness, empathy and bias' would be a way to improve the quality of moral decisions of individuals no matter their antecedent values. Another potential target for procedural assistance would be affective perspective-taking—the ability to recognize or infer the emotional states of others (Klincewicz et al. 2018). Via advice and training, the system could also help one develop dispositions for reflection (Giubilini and Savulescu 2018: 179) or for engaging in various stoic practices (Klincewicz 2019). The AEA could also help train one's own capacities for recognizing situations that (one believes) raise moral issues and for more quickly or reliably noticing features of a situation that one considers after reflection to be morally relevant (e.g. by training affective perspective-taking).

In addition to procedural assistance, an AEA might assist in advance of particular cases by helping us better understand our own moral views. For instance, the AI system might find patterns in our judgements about cases so as to identify the underlying norms and principles we subscribe to and the core values we are most concerned with. The advisor might also identify conflicts between an individual's moral world view and their actions or identify apparent conflicts within an individual's moral belief system. Making these things explicit may prompt reflection on whether these are the norms, principles, and values one wants to endorse. We call this the AEA's *moral psychologist function*.

The AEA could also prompt the user to increase the completeness of their moral world view—for instance, by giving them new cases to consider, such as cases they might be likely to encounter in the future in our profession, so that the user thinks about them in advance with the idea that this puts the user in a better position to analyse and develop views about similar cases when they are encountered in reality. In addition, the AEA could simply provide facts that can become a part of background information relevant to future moral decision-making.

On-the-spot assistance: On-the-spot advice and facilitation

During the time when the individual is considering a particular moral question or case, the AEA could supply support by facilitating the functioning of component parts of one's moral deliberation process. If the user has already formed a tentative moral judgement on a particular question, the AEA may weigh in with an assessment of whether the judgement is consistent with one's deeper values and other commitments.[18] Savulescu and Maslen's proposed *moral reasoning prompter* would help the individual think through a moral decision by prompting the individual through some (personalizable and possibly adapting) reasoning procedure. They list some example questions for this process, such as 'Would the act involve crossing a line you promised yourself or another you wouldn't cross?' and 'Do you think you will feel shame or remorse if you go ahead with the act?' (Savulescu and Maslen 2015: 87).[19] With information about the principles or considerations underlying one's judgements (or other people's judgements), the AEA could help with the moral reasoning process by collecting and presenting to the user a range of arguments or considerations that might be offered for and against particular options for action. Lara and Deckers propose a 'Socratic Assistant', with which the user would 'deliberate in dialogue' to subject the user's beliefs to 'conditions of empirical, logical and ethical rigour' (Lara and Deckers 2019: 8);[20] Seville and Field suggest that AI systems could play a devil's advocate role in moral reasoning (Seville and Field 2000). For Flora, who wants to decide whether to become a vegetarian or a vegan or continue eating meat, the AEA could provide arguments for each of these positions, generating them on the basis of her antecedent values. If Flora is a utilitarian, the AEA could provide utilitarian arguments for these diets.

An AEA could remind one of the facts or analogous cases that one usually considers relevant for a given kind of moral question but that one may fail to think about at the moment. It could draw one's attention to features of the situation that one believes are relevant but that one does not always attend to or adequately consider. Suppose Flora is visiting her grandmother and wondering about whether to eat the grilled chicken that her grandmother has lovingly prepared: it might be helpful for the ethics adviser to remind Flora to think about how her grandmother may feel if she refuses the meal, given that on previous, similar occasions Flora has regretted failing to consider the cook's feelings.

Another function that is relevant when the individual is in the midst of considering a case is that of the 'moral environment monitor' (Savulescu and Maslen 2015), which monitors general factors that cause the individual to reason better or worse. These might include things like stress level, sleep deprivation, hormone levels, possible influence by unwanted emotions, etc. On the spot, the system could alert the individual if they are making a decision in a less optimal condition—for example, so that the individual could postpone the decision or take steps to counteract the effects of the non-optimal environment. (If an AEA were to supply information about these patterns in *advance* of the individual considering the particular case or question, then it would be playing a preparatory role—for example, if it recommended that the user ensure that particular conditions are in place before making decisions. And, if the effect of the environment monitor over the long term was that the user learned when they were better at moral decision-making and adjusted their moral deliberation process accordingly, the result would be an enduring procedural enhancement.)

More directly, the AEA might supply advice on what means would fulfil one's ends. Supposing Flora has decided to minimize the number of meat products she consumes, the AI system might assist her lunch decision by screening all the meal options based on whether they include meat. If Flora were thinking more long term, making a decision about where to live, she might be interested in which areas are most conducive to a vegetarian lifestyle; the system might then advise her in a way that allows her to change her social and environmental context so that it is easier to maintain her vegetarianism. Along these lines, Savulescu and Maslen propose the *moral organizer* (Savulescu and Maslen 2015: 86), which may identify ways in which the person could meet their goal or alert them to the fact that they have not yet met the goal—thus aiding them on the spot in the ongoing process of reasoning about how one can meet one's goal.

More fundamentally, AEAs might provide advice—recommendations or information—meant to prompt changes to people's ends or their interpretations of their ends—their core moral beliefs and values. For example, working from premises that the user already accepts or would be likely to accept if they considered them, the AI system could identify, for the user's consideration, a range of arguments for and against possible core moral beliefs and values (Klincewicz 2016: 180; Lara and Deckers 2019: 9). The AI system might have gathered those arguments from other human beings (e.g. from philosophical texts) or it may have generated them itself. The arguments could proceed by appealing to non-moral values (e.g. epistemic values such as coherence) that the individual holds; or they could appeal to metaethical or conceptual arguments (e.g. an argument appealing to the nature of action or obligation).

What kind of arguments would work for this function would depend on what beliefs the user already has. For instance, one type of information potentially relevant for changing one's ends would be information about the morality of others—their judgements about cases, their concepts, the types of considerations they believe to be relevant, the weight they put on different considerations, and so on. Depending on the question, one might be interested in the beliefs of people in one's profession, one's parents and friends, religious leaders, or the entire population. In the moral domain, the reason this information would be of interest is that the user may consider other people's moral beliefs to be information that bears in some way on what they ought to believe (c.f. Chituc and Sinnott-Armstrong 2020: 273) or ought to take into account. To that end, the AEA could deliver Flora a compendium of commonly held moral beliefs about a vegetarian diet among people with whom she shares other moral views, for example.

In addition to offering direct advice on one's ends, the AEA might facilitate one's reasoning about one's ends—for instance, the system could guide the user through the reasoning or discernment processes recommended by particular traditions. If the user believes that a particular approach to ethical reasoning and decision-making is right (e.g. Jewish tradition, or virtue ethics, or something else), the system could help guide the individual through reasoning and decision-making processes from those traditions to help them modify or reinterpret their core values.[21]

Ethical risks and problems

Given the wide variety of moral functions that could be handed off to an AEA, identifying all the potential ethical issues associated with the use of ostensible AEAs would require much more than one chapter. In addition, which ethical problems may arise depends substantially on what form AEAs might take. Here, we focus on three problems that are worth considering for a broad range of AEAs for moral self-improvement, all of which are tied to their use of AI: (a) the risks associated with AEA reliance on information about the user's own moral psychology; (b) the risks associated with outside influences on AEAs and the possibility that an AEA will not accomplish its function; and (c) the risks associated with taking the AEA's advice without having adequate reasons. After discussing these problems, we briefly discuss some of the problems that have come up in the literature on moral bioenhancement that also appear applicable to the AEA case.[22]

The sensitivity of moral data

Given the complexity of human moral reasoning and psychology, for an AEA to be effective it would likely require substantial information about the world, the cases the user encounters, and the user. In particular, in order to advise the user in a personalized way, an AEA would need *moral data* about that individual: information about the individual's values and beliefs, and possibly even quite an in-depth model of that individual's moral psychology. Information about an individual's values (moral, political, personal, etc.), let alone information about the underlying patterns in a person's moral psychology, which the individual is not even aware of themselves, can be highly sensitive (cf. Christen et al. 2015 on 'morality mining'). Thus, one of the distinguishing ethical issues associated with AEAs, in contrast to other moral improvement interventions, is the risk introduced by its likely reliance on large quantities of detailed, sensitive moral data.

Each of the AEA's various interactions with the individual's moral data—elicitation, aggregation, storage, processing, creation (based on inferences from other data), possible sharing, and use—introduce risks for the user.[23] Furthermore, the creation of AEA technology, long before it reaches the user, may require a process in which AI systems learn moral concepts and construct models of moral world views; such a

process would itself likely require the collection and processing of moral data from many people. Then, during use, if the AEA system is updated based on information about others or if one of its functions is to provide the user information about the values and opinions of others, risks arise for all the individuals whose moral information is collected and transferred—such information may flow or be used in ways that violate rights or are otherwise problematic.[24] Additional risks arise from the identification of patterns in moral views across individuals that enable further inferences, including inferences about groups and uninvolved individuals (on the issue of group privacy, see Floridi 2014; Taylor et al. 2016; Loi and Christen 2019; Veliz 2020, ch. 3; on privacy interdependencies, see Humbert et al. 2019 and Barocas and Levy 2020). The use of AEAs by some individuals could expose (probabilistic) information about non-users that enables predictions about non-users' likely moral views or dispositions based on other available information about them. This opens the door for dangerous discrimination against both historically marginalized groups and previously unrecognized groups that become salient as a result of newly discovered moral psychological patterns (see Taylor 2016 on algorithmic groupings).

Why is moral data so sensitive? Information about the individual AEA user—which might include information about the user's intentions, past actions, judgements about controversial cases, and so on—would be of great interest to many people, such as prospective employers, spouses, landlords, security services, loan providers, governments, etc. Imagine a landlord interested to know whether a prospective renter has little regard for property rights; a person interested in whether their fiancée secretly thinks that occasional cheating is not so bad; a parent concerned with whether a babysitter will be a bad influence. Such parties might even claim to have a legitimate interest in accessing this data under some circumstances. Morality is central to identity (Strohminger and Nichols 2014), and the topics that people tend to consider questions of morality are frequently high-stakes. Importantly, people tend to be less tolerant of those they disagree with on moral matters than of those they disagree with on other matters (Skitka et al. 2005; Wright et al. 2008). Individuals' moral data would also be of great interest to people concerned with rule violations and punishment, such as divorce attorneys, law enforcement, juries, etc. In addition, any system that relies on valuable personal data runs a risk of leakages and attacks. We can imagine AEA user data being repurposed for a variety of illegal and malicious ends—moral data used for blackmail, exposing political dissidents, manipulating targets, etc.

Historically, information about individuals' moral values has only been accessible to others in a piecemeal way, for example, via self-disclosure, rumour, behaviour, and other indirect indicators.[25] We have laws and norms that regulate our interactions with such information—in many countries, for instance, closely related data, such as religious affiliation and voting record, is subject to various protections. Digitalizing moral data at the scale an AEA would require introduces substantial hazards—it would harvest some of the most significant information about individuals that there is. It is difficult to anticipate all the ways in which storehouses of presumed-reliable information about individuals' moral world views might alter our social world. Think of friendships made untenable by one person's repugnance at the moral views of the other, moral purity tests for employment, deeper polarization and segregation between groups on the basis of moral views, and moral inquisitions.

Some of the risks associated with moral data can perhaps be addressed by engineering solutions that prevent repurposing or hacking, other technological tools, new social norms, and legal instruments that regulate moral data and the development of purported AEAs. But pending substantial developments along these lines, for the sake of social functioning there is reason to be wary of highly personalized AEA, given their likely reliance on massive amounts of moral data.

The inescapability of outside influences on AEAs

A worry for any ostensible moral enhancement intervention is that it will fail to perform its purported function and, on top of that, that the user might not recognize the failure. Let us distinguish two possible characterizations of the AEA's function. One is the *individual value alignment function*: the system helps the user better align their moral judgements and decisions with their core moral values (cf. Christian 2020; Boddington 2021). However, if the user's core moral values are completely misguided, it could be that efforts to increase alignment will not result in moral improvement and may indeed result in moral worsening. Thus, there is another characterization of the AEA's function to consider, which we will label the *moral improvement function*: the system helps the person improve their moral judgements and decisions, simpliciter. For a user whose core moral values are roughly on track, performing the individual value alignment function may have the result that the system also performs the moral improvement function. For the user whose values are completely misguided, the AEA would presumably need to influence the user's core moral values (or their interpretation) if it is to fulfil the moral improvement function. Since the moral improvement function lies beyond the scope of this chapter, we are concerned here with the possible function failure of an intervention designed to perform individual value alignment and in particular with the low probability that the user would be able to assess whether the system will perform its purported function.

In practice, an AEA cannot be fully personalized: the values of others will invariably influence the design of the system in ways that no user could fully reconstruct (cf. Serafimova 2020). Although the system may take the user's moral views as input, and though it could even be designed to change over time on the basis of the user's input, the initial design of the system and its parts (unless the user were to design the entire system from scratch) comes from elsewhere. For an AEA to come into being, some set of individuals must design it, construct it, advertise it, etc. The creation and availability of such technology is influenced by economic and political structures as well as many other individuals' interests and values; whatever advice or guidance the system provides the user will be partly determined by the choices of the designers of the system, the institution that built it, standards organizations, and others (Frank and Klincewicz 2016).

At minimum, the variety of ostensible AEAs that the interested user may encounter is constrained. As with any computer system, those options that do become available will be the product of a vast number of value-laden design decisions, many of which the user will not be aware of or able to inspect (Friedman and Nissenbaum 1996; Nissenbaum 2001). Among other things, there is the unavoidable challenge of the system needing to select some format in which to present options and information, given the existence of order effects and other presentation factors that affect reasoning. The result may be that the AEA systematically biases or restricts the scope of morally relevant actions that the user considers; it might 'imprison us within a certain zone of agency' (Danaher 2018: 18). There is also the possibility that the use of an AEA, though not intended (by designers or users) to change one's core values, nonetheless does. Using an AEA could then be a 'transformative experience' (Paul 2014), producing a shift in one's values.[26] As many philosophers have observed, even ordinary technology can mediate values and change the values of its users without their awareness (Bergen and Verbeek 2021). The reverse problem is also a potential worry—that a user who interacts with the system will change less over time than they would have had they conducted their moral deliberation without the system; that the system will constrict the moral growth of the user.

In addition to unintended influence, there are critical concerns related to intentional bias, manipulation, paternalism, and social control. If such systems are created by people with an agenda or partisan interests, there will be incentives to design them in ways that nudge or otherwise influence users' choices or even their core moral views. Hidden influence is a risk associated with decision support systems generally (Susser 2019) and would be especially pressing in a context in which the system is used with the aim of moral improvement.

There are serious reasons to doubt that a user could ever adequately confirm that an AEA is successfully accomplishing the individual value alignment function. This is so despite the fact that, for the most part, the user of the AEA system in the scenario we have imagined is an active rather than passive participant in the moral improvement process. In an active moral enhancement intervention, at least as discussed by Focquaert and Schermer (2015), the intervention requires conscious mental participation on the part of the subject. The AEA that we have considered intervenes by offering information and recommendations, which the user may accept or disregard; training, which presumably requires the user's conscious participation; and function facilitation, which we will assume here occurs only with the user's active involvement. The user has an idea of how the AEA is influencing them; and the user may cease using the AEA at any time.

Nonetheless, even for the best-case scenario that we have sketched, in practice there remains a high probability that the system will promote values other than the user's and that the user will not be able to identify the ways in which the design decisions or interference of others is exerting an influence on their moral reasoning. Among the causes of this epistemic difficulty is the complexity and opacity of the system, which we discuss in the next section. Thus, there is an ethical risk associated with the purported AEA failing to perform its individual value alignment function and causing the user whose moral convictions are mostly on track to morally worsen.

Moral advice without adequate reasons

One of the worries about moral bioenhancement is that, even if it changes some aspects of an individual's morality for the better (whether their dispositions, actions, beliefs, or something else), if the user did not make that change on the basis of adequate reasons, that change may be less than praiseworthy or may not constitute a moral improvement overall. This kind of issue shows up in a particularly important form for the AEA, given that it would likely employ machine learning, which tends to be highly opaque (Burrell 2016; Goodman and Flaxman 2017; Arrieta et al. 2020). The problem arises most directly when the AEA acts as an advisor providing recommendations.

Suppose the AEA system were to recommend a course of action that would increase the alignment of the individual's reasoning and decision-making with their core moral values. The user does this but lacks adequate reasons for the change (e.g. it may be that the user's only reason in favour of the change is the fact that it is AEA-recommended). The user is unable to independently identify adequate reasons for their behaviour, and the AEA is not able to supply the user with adequate reasons in support of its recommendation. The result of this attempt to use an AEA for moral improvement would be that the user obtains beliefs and reaches decisions without being in possession of or being capable of formulating reasons for those beliefs and decisions. If, as a number of philosophers think, there is inherent or instrumental value in the human agent being in possession of or able to supply reasons for their judgements, actions, and so on, this is a problem.

For instance, Nickel (2001) argues that there is a 'recognition requirement' for moral action. On his view, to act in a moral way one must act on the basis of a recognition of what morality requires: 'morality requires one to act from an understanding of moral claims, and therefore to have an understanding of moral claims that are relevant to action' (257). Having a full understanding of a moral claim is 'a matter of having a grasp of the relevant reasons bearing on action, or in other words, having a grasp of the justificatory basis of the claim' (259). On this view, then, if an AEA cannot supply the user with adequate reasons for its recommendation, there is a risk that one will be left without an understanding of the moral claim so that, even if one acts on the basis of it, one's action will not be moral.

Suppose that in the situation in which Flora is trying to decide whether to eat the meat dish her grandmother has prepared, the best action to take, given the circumstances, Flora's core values, and so on, would be to eat the dish. If Flora takes this action solely on the basis of the AEA's recommendation, without recognition of the morally relevant features of the situation (without recognizing, say, that she should be appreciative of her grandmother's efforts or that accepting the dish would convey respect, etc.), then her action might not be a moral action. Furthermore, lacking a full understanding of the moral conclusion that she has obtained from the AEA puts Flora in a less than optimal position going forward; facing similar situations in the future, Flora will be no more capable of reasoning about what to do than she was before she consulted the AEA about how to respond to her grandmother.

There may be some circumstances in which it is permissible to defer to the moral testimony of another human being—perhaps in low-stakes circumstances, in situations in which one obtains advice about how to implement shared values, or when 'we know that we're unable to resolve a difficult moral issue and reasonably expect that others can do better' (Mogensen 2017: 263). However, deferring to the advice of an AEA may be different—the AEA, or our relationship with it, may be missing some element on which our usual deference relationships depend. This could be so, for instance, if the permissibility of deference to human testimony rests on the fact that human beings can be held responsible for their testimony. The AEA might not be the sort of thing that can be punished or adequately held to account for its claims or recommendations (Floridi 2016; Nyholm 2018).

By contrast, if the AEA could help the user to acquire adequate reasons for making the change that the system recommends, the user would not be epistemically dependent on the AEA's advice itself; the AEA might have played a critical causal role in facilitating the user's acquisition of reasons (and user dependence on that facilitation could raise its own ethical problems), but the user's beliefs or decisions would rest on the reasons that the user has acquired over the course of their interaction with the AEA.

What are the prospects that the AEA could not only advise but could also help the user acquire adequate reasons for accepting the AEA's claim or following its recommendation? A natural place to start would be for the AEA to share information about what prompted its advice. Unfortunately, the computational processes of an AI system that could function as an AEA would invariably be complex, likely so complex that no human being would be capable of understanding them. Even for existing machine-learning systems, the factors that prompt the system to generate outputs are not easily translated into familiar human concepts, let alone moral concepts that might provide the building blocks for articulating reasons. Substantial efforts are currently underway to create explainable AI systems (see the discussion in, e.g. Mittelstadt et al. 2016; Goldenfein 2019; Martin 2019), but some are pessimistic about how far such techniques can take us (see, e.g. Rudin 2019).

In sum, one concern about AEAs that provide moral advice is that they may prompt the user to make changes—acquire beliefs, take actions, and so on—without adequate reasons. This may undermine the moral worth of the user's actions. Although a user like Flora might have had noble motivation for using an AEA in the first place (namely, acting more morally), if she is left without reasons for the system's recommendations, following its advice may leave her less morally worthy.

Additional problems

Many of the potential ethical problems associated with moral bioenhancement efforts may also be applicable for AEAs. One general issue has to do with the fact that using new technologies to interfere with a complex system like human moral tradition and practice risks causing significant bad consequences we cannot predict beforehand. Despite the flowering of empirical research on moral cognition in the past few decades, we still know only a fraction of what there is to know about how moral cognition works—including how hormones and other chemicals, situational factors, and so on, affect reasoning; and how moral world views develop and change. Furthermore, both the functioning and development of moral cognition depends crucially on one's environment and one's interactions with others. Supposing that we could successfully create AI systems that could perform the individual value alignment function, there are still numerous risks associated with use of such a computer system.

For instance, suppose we were using an AEA to reduce inconsistencies in our own moral belief systems—it may well be that some inconsistencies in one's moral beliefs are useful somehow; maybe becoming more consistent would reduce flexibility, creativity, or other qualities that undergird well-functioning moral cognition. For instance, it could be that being as free as we currently are to fall short of our moral standards permits us to keep our standards high. If an AEA were to point out the inconsistencies in our actions and beliefs too frequently, we might respond in a potentially perverse way by lowering our standards. Giubilini and Savulescu (2018: 174) mention this issue, discussing the possibility that one might start out as a strict utilitarian, but in the face of the AI system's recommended actions, such as donating half one's income to charity, say, one might begin to think that strict utilitarianism is too demanding a theory. This weakening of standards might not constitute an improvement.

Another class of unintended consequence has to do with the potential for problematic dependence (Danaher 2018). One incarnation of this problem is the potential for moral deskilling (Vallor 2015). We would want to anticipate whether the use of AEAs would cause us to lose our own abilities to perform essential moral cognitive tasks. We would also want to know whether our own ability to perform that task is instrumentally valuable, such as for reasons of robustness, or intrinsically valuable. Similar concerns already exist in domains of human activity where the introduction of computing tools permanently changes people's cognition. These unintended consequences that AEAs introduce may be difficult to undo. Of course, it is an empirical question whether AEAs would worsen our ability to independently perform whatever cognitive tasks it takes over, but when it comes to moral decision-making, such a technology should not be unleashed until we have good evidence that it does not.

However, preparatory assistance, including procedural assistance, is less likely to raise this worry than some of the other assistance that the AEA might provide. Preparatory assistance aims to help us improve skills and obtain information (with the idea that the user retains that information for future use or develops judgements on the basis of that information that are available for future use).

We should also worry about effects such as general alienation from other people and a loss of the social aspects of moral practice—for instance, threats to moral collaboration and discussion due to greater reliance on technology during the process of moral reasoning and decision-making. Among other things, some people think that coming to an ethical agreement via direct social interaction is an intrinsically valuable part of human morality. Social groups are often distinguished by their moral values, and their cohesion often depends on norms that regulate and punish transgressions. The widespread use of personalized AEAs for moral self-improvement may effectively dampen the societal role of moral practice.

Depending on how we interact with AEAs—for example, which moral cognition tasks we outsource entirely, which we perform using AEAs as cognitive extensions (Hernández-Orallo and Vold 2019)—and how reliant we become on them, AEAs may threaten some core component of what we take to be human nature; this is a concern that some have also had about some forms of moral bioenhancement (Cohen 2006; Sandel 2004). There is a worry about whether these interventions would undermine our self-understanding as human beings, as some authors argue moral bioenhancement may do (Elliott 2014; Danaher 2019; Kraemer 2011), or undermine human dignity (Fukuyama 2003; Cohen 2006; Giubilini and Sanyal 2015). A closely related potential problem is the possibility that the AEA could undermine freedom of will or action or, more generally, freedom of mind: some have argued that pharmacological, genetic, or direct neural interventions would impose a limit on these capacities (Bublitz 2016: 94) and that these interventions interfere with our freedom to morally fall (Harris 2011; DeGrazia 2014; Pugh 2019). An AEA may similarly challenge this freedom, especially if such devices could cause permanent changes in our moral world view or ensure, in a way that circumvents our agency, that we always act in accordance with our values.

We may also have certain obligations related to moral reasoning that the use of an AEA would prevent us from discharging. For instance, there may be something morally wrong in Flora failing to spontaneously, independently consider the well-being of her grandmother when making a decision about whether to accept the food she offers. One may argue that this is ultimately a function that she is obliged to perform herself, without the prompting of a computer system.

Another recurring type of worry about enhancement interventions has to do with problems that arise if a substantial portion of the population uses the technology. In the context of moral cognition, widespread changes to our reasoning and decision-making process could produce homogenization at the group level, which might be a problem, for example, if it limits the potential for moral progress at the societal level (Schaefer 2015). For the personalized AEA that we have discussed, one might worry more about polarization and radicalization than homogenization, but the latter is also a risk, particularly if there are common patterns of influence in the way that AEAs affect users' moral cognition.

Conclusion

In this chapter, we have analysed three of the risks raised by individual use of purported AEAs, each of which applies to a range of systems designed to provide personalized assistance for individual moral reasoning in the near term. We focused on the hazards associated with reliance on moral data, outside influence on the design of AEAs, and the possibility of the AEA user adopting beliefs and making decisions without sufficient reasons. Each of these is a problem that is difficult to avoid with AEAs. To provide specific, personalized, moral assistance with a broad enough scope to be interesting, an AI system presumably requires vast quantities of moral data. Likewise, at least some outside influence on the substance of AEA advice, via design decisions and other constraints, cannot be avoided. Finally, it remains to be seen to what extent the opacity of complex machine-learning systems can be overcome and to what extent such systems will be able to assist users in acquiring adequate reasons for beliefs and decisions influenced by the AI system. Thus, we consider these three problems to be rather serious obstacles for the hope that AI systems might be harnessed for the purpose of helping individuals who wish to improve their own moral functioning.

Acknowledgements

Thanks to audiences at a Cornell Tech Digital Life Initiative seminar, a TU/e Center for Humans and Technology Research Meeting on Artificial Moral Agents, and CEPE 2019 for comments on earlier versions of this paper. E.O.'s research on this paper has been supported by the Netherlands Organisation for Scientific Research under grant number 016.Veni.195.513; the Ethics of Socially Disruptive Technologies research programme, funded through the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organisation for Scientific Research under grant number 024.004.031; a fellowship at the Cornell Tech Digital Life Initiative NSF Grant #SES-1650589 (PI: Helen Nissenbaum); and a visit to the Simons Institute at the University of California, Berkeley.

References

Alfano, Mark, Ebrahimi Fard, Amir, Carter, J. Adam, Clutton, Peter, and Klein, Colin (2021), 'Technologically Scaffolded Atypical Cognition: The Case of YouTube's Recommender System', *Synthese* 199(1–2), 835–858. doi: https://doi.org/10.1007/s11229-020-02724-x.
Google Scholar WorldCat

Arrieta, Alejandro Barredo, Díaz-Rodríguez, Natalia, Del Ser, Javier, Bennetot, Adrien, Tabik, Siham, Barbado, Alberto, García, Salvador, Gil-López, Sergio, Molina, Daniel, Benjamins, Richard, Chatila, Raja, and Herrera, Francisco (2020), 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI', *Information Fusion* 58, 82–115. doi: https://doi.org/10.48550/arXiv.1910.10045.
Google Scholar WorldCat

Awad, Edmond, Dsouza, Sohan, Kim, Richard, Schulz, Jonathan, Henrich, Joseph, Shariff, Azim, Bonnefon, Jean-François, and Rahwan, Iyad (2018),'The Moral Machine Experiment', *Nature* 563(7729), 59–64. doi: https://doi.org/10.1038/s41586-018-0637-6.
Google Scholar WorldCat

Barocas, Solon, and Levy, Karen (2020), 'Privacy Dependencies', *Washington Law Review* 95(2), 555–616.
Google Scholar WorldCat

Beck, Birgit (2015), 'Conceptual and Practical Problems of Moral Enhancement', *Bioethics* 29(4), 233–240. doi: https://doi.org/10.1111/bioe.12090.
Google Scholar WorldCat

Bergen, Jan Peter, and Verbeek, Peter-Paul (2021), 'To-Do Is to Be: Foucault, Levinas, and Technologically Mediated Subjectivation', *Philosophy and Technology* 34(2), 325–348. doi: https://doi.org/10.1007/s13347-019-00390-7.
Google Scholar WorldCat

Boddington, Paula (2021), 'AI and Moral Thinking: How Can We Live Well with Machines to Enhance Our Moral Agency?', *AI and Ethics* 1(2), 109–111.
Google Scholar WorldCat

Borenstein, Jason, and Arkin, Ron (2016), 'Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being', *Science and Engineering Ethics* 22(1), 31–46.
Google Scholar WorldCat

Bublitz, Christoph (2016), 'Moral Enhancement and Mental Freedom', *Journal of Applied Philosophy* 33(1), 88–106. doi: https://doi.org/10.1111/japp.12108.
Google Scholar WorldCat

Burrell, Jenna (2016), 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms', *Big Data and Society* 3(1). doi: https://doi.org/10.1177/2053951715622512.
Google Scholar WorldCat

Cave, Stephen, Nyrup, Rune, Vold, Karina, and Weller, Adrian (2019), 'Motivations and Risks of Machine Ethics', *Proceedings of the IEEE* 107(3), 562–574. doi: https://10.1109/JPROC.2018.2865996.
Google Scholar WorldCat

Chituc, Vladimir, and Sinnott-Armstrong, Walter (2020), 'Moral Conformity and Its Philosophical Lessons', *Philosophical Psychology* 33(2), 262–282. doi: https://doi.org/10.1080/09515089.2020.1719395.
Google Scholar WorldCat

Christen, Markus, Alfano, Mark, Bangerter, Endre, and Lapsley, Daniel (2015), 'Ethical Issues of "Morality Mining": Moral Identity as a Focus of Data Mining', in *Human Rights and Ethics: Concepts, Methodologies, Tools, and Applications*, (Hershey, PA: IGI Global), 1146–1166.
Google Scholar Google Preview WorldCat COPAC

Christian, Brian (2020), *The Alignment Problem: Machine Learning and Human Values* (New York: WW Norton & Company).
Google Scholar Google Preview WorldCat COPAC

Cohen, Eric (2006), 'Conservative Bioethics and the Search for Wisdom', *Hastings Center Report* 36(1), 44–56. doi: https://doi.org/10.1353/hcr.2006.0004.
Google Scholar WorldCat

Conitzer, Vincent, Sinnott-Armstrong, Walter, Borg, Jana Schiach, Deng, Yuan, and Kramer, Max (2017), 'Moral Decision Making Frameworks for Artificial Intelligence', Proceedings of the 31st Association for the Advancement of Artificial Intelligence (AAAI-17) Conference on Artificial Intelligence, 4831–4835.

Danaher, John (2018), 'Toward an Ethics of AI Assistants: An Initial Framework', *Philosophy and Technology* 31(4), 629–653. doi: https://doi.org/10.1007/s13347-018-0317-3.
Google Scholar WorldCat

Danaher, John (2019), 'Why Internal Moral Enhancement Might Be Politically Better Than External Moral Enhancement', *Neuroethics* 12(1), 39–54. doi: https://doi.org/10.1007/s12152-016-9273-8.
Google Scholar WorldCat

de Sio, Filippo Santoni, and van den Hoven, Jeroen (2018), 'Meaningful Human Control over Autonomous Systems: A Philosophical Account', *Frontiers in Robotics and AI*, 28 February. doi: https://doi.org/10.3389/frobt.2018.00015.

DeGrazia, David (2014), 'Moral Enhancement, Freedom, and What We (Should) Value in Moral Behaviour', *Journal of Medical Ethics* 40(6), 361–368. doi: https://doi.org/10.1136/medethics-2012-101157.
Google Scholar WorldCat

Earp, Brian D. (2018), 'Psychedelic Moral Enhancement', *Royal Institute of Philosophy Supplements* 83, 415–439. doi: https://doi.org/10.1017/s1358246118000474.
Google Scholar WorldCat

Elliott, Carl (2014), *A Philosophical Disease: Bioethics, Culture, and Identity* (New York: Routledge). doi: https://doi.org/10.4324/9781315822150.
Google Scholar Google Preview WorldCat COPAC

Etzioni, Amitai, and Etzioni, Oren (2016), 'AI Assisted Ethics', *Ethics and Information Technology* 18(2), 149–156. doi: https://doi.org/10.1007/s10676-016-9400-6.
Google Scholar WorldCat

Floridi, Luciano (2014), 'Open Data, Data Protection, and Group Privacy', *Philosophy & Technology* 27(1), 1–3.
Google Scholar WorldCat

Floridi, Luciano (2016), 'Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions', *Philosophical Transactions of the Royal Society A, Mathematical, Physical and Engineering Sciences* 374(2083). doi: https://doi.org/10.1098/rsta.2016.0112.
Google Scholar WorldCat

Focquaert, Farah, and Schermer, Maartje (2015), 'Moral Enhancement: Do Means Matter Morally?', *Neuroethics* 8(2), 139–151. doi: https://doi.org/10.1007/s12152-015-9230-y.
Google Scholar WorldCat

Formosa, Paul, and Ryan, Malcolm (2021), 'Making Moral Machines: Why We Need Artificial Moral Agents', *AI and Society* 36(3), 839–851.
Google Scholar WorldCat

Frank, Lily, and Klincewicz, Michał (2016), 'Metaethics in Context of Engineering Ethical and Moral Systems', Association for the Advancement of Artificial Intelligence (AAAI-16) Spring Symposium Series.

Freedman, Rachel, Borg, Jana Schiach, Sinnott-Armstrong, Walter, Dickerson, John P., and Conitzer, Vincent (2020), 'Adapting a Kidney Exchange Algorithm to Align with Human Values', *Artificial Intelligence* 283. doi: https://doi.org/10.1016/j.artint.2020.103261.
Google Scholar WorldCat

Friedman, Batya, and Nissenbaum, Helen (1996), 'Bias in Computer Systems', *ACM Transactions in Information Systems* 14(3), 330–347. doi: https://doi.org/10.1145/230538.230561.
Google Scholar WorldCat

Fukuyama, Francis (2003), *Our Posthuman Future: Consequences of the Biotechnology Revolution* (New York: Farrar, Straus and Giroux).
Google Scholar Google Preview WorldCat COPAC

Giubilini, Alberto, and Sanyal, Sagar (2015), 'The Ethics of Human Enhancement', *Philosophy Compass* 10(4), 233–243. doi:

https://doi.org/10.1111/phc3.12208.
Google Scholar WorldCat

Giubilini, Alberto, and Savulescu, Julian (2018), 'The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence', *Philosophy & Technology* 31(2), 169–188. doi: https://doi.org/10.1007/s13347-017-0285-z.
Google Scholar WorldCat

Goldenfein, Jake (2019), 'Algorithmic Transparency and Decision-Making Accountability: Thoughts for Buying Machine Learning Algorithms' in Cliff Bertram, Adriana Nugent, and Asher Gibson, eds, *Closer to the Machine: Technical, Social, and Legal Aspects of AI* (Melbourne: Office of the Victorian Information Commissioner), 41–61.
Google Scholar Google Preview WorldCat COPAC

Goodman, Bryce, and Flaxman, Seth (2017), 'European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"', *AI Magazine* 38(3), 50–57. doi: https://doi.org/10.1609/aimag.v38i3.2741.
Google Scholar WorldCat

Graham, Jesse, Haidt, Jonathan, Koleva, Sena, Motyl, Matt, Iyer, Ravi, Wojcik, Sean P., and Ditto, Peter H. (2013), 'Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism', *Advances in Experimental Social Psychology* 47, 55–130. doi: https://doi.org/10.1016/B978-0-12-407236-7.00002-4.
Google Scholar WorldCat

Hadfield-Menell, Dylan, Dragan, Anca, Abbeel, Pieter, and Russell, Stuart J. (2016), 'Cooperative Inverse Reinforcement Learning', *Advances in Neural Information Processing Systems* 29, 1–9.
Google Scholar WorldCat

Hansson, Lena (2017), 'Promoting Ethical Consumption: The Construction of Smartphone Apps as "Ethical" Choice Prescribers', in Franck Cochoy, Johan Hagberg, Magdalena Petersson McIntyre, and Niklas Söram, eds, *Digitalizing Consumption: How Devices Shape Consumer Culture* (London: Routledge), 270. doi: https://doi.org/10.4324/9781315647883.
Google Scholar Google Preview WorldCat COPAC

Harris, John (2011), 'Moral Enhancement and Freedom', *Bioethics* 25(2), 102–111.
Google Scholar WorldCat

Hern, Alex (2018), 'Facebook, Apply, YouTube and Spotify ban Infowars' Alex Jones', *The Guardian*. https://www.theguardian.com/technology/2018/aug/06/apple-removes-podcasts-infowars-alex-jones, accessed 24 April 2022.

Hernández-Orallo, José, and Vold, Karina (2019), 'AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI', *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 507–513. doi: https://doi.org/10.1145/3306618.3314238.

Humbert, Mathias, Trubert, Benjamin, and Huguenin, Kévin (2019), 'A Survey on Interdependent Privacy', *ACM Computing Surveys (CSUR)* 52(6), 1–40.
Google Scholar WorldCat

Klincewicz, Michał (2016), 'Artificial Intelligence as a Means to Moral Enhancement', *Studies in Logic, Grammar and Rhetoric* 48(61). doi: https://doi.org/10.1515/slgr-2016-0061.
Google Scholar WorldCat

Klincewicz, Michał (2019), 'Robotic Nudges for Moral Improvement', *Techné: Research in Philosophy and Technology* 23(3), 425–455. doi: https://doi.org/10.5840/techne2019122109.
Google Scholar WorldCat

Klincewicz, Michał, Frank, Lily E., and Sokólska, Marta (2018), 'Drugs and Hugs: Stimulating Moral Dispositions as a Method of Moral Enhancement', *Royal Institute of Philosophy Supplements* 83, 329–350. doi: https://doi.org/10.1017/s1358246118000437.
Google Scholar WorldCat

Kosinski, Michal, Stillwell, David, and Graepel, Thore (2013), 'Private Traits and Attributes Are Predictable from Digital Records of Human Behavior', *Proceedings of the National Academy of Sciences of the United States of America* 110(15), 5802–5805. doi: https://doi.org/10.1073/pnas.1218772110.
Google Scholar WorldCat

Kraemer, Felicitas (2011), 'Authenticity Anyone? The Enhancement of Emotions via Neuro- Psychopharmacology', *Neuroethics* 4, 51–64. doi: https://doi.org/10.1007/s12152-010-9075-3.
Google Scholar WorldCat

Kreitmair, Karola V. (2019), 'Dimensions of Ethical Direct-to-Consumer Neurotechnologies', *AJOB Neuroscience* 10(4), 152–166.
Google Scholar WorldCat

Lara, Francisco (2021), 'Why a Virtual Assistant for Moral Enhancement When We Could Have a Socrates?', *Science and Engineering Ethics* 27(4), 1–27.
Google Scholar WorldCat

Lara, Francisco, and Deckers, Jan (2019), 'Artificial Intelligence as a Socratic Assistant for Moral Enhancement', *Neuroethics* 13, 275–287. doi: https://doi.org/10.1007/s12152-019-09401-y.
Google Scholar WorldCat

Leike, Jan, Krueger, David, Everitt, Tom, Martic, Miljan, Maini, Vishal, and Legg, Shane (2018), 'Scalable Agent Alignment via Reward Modeling: A Research Direction', arXiv Preprint, arXiv1811.07871.

Liebowitz, Jay (2019), *The Handbook of Applied Expert Systems* (Boca Raton: cRc Press).
Google Scholar Google Preview WorldCat COPAC

Loi, Michele, and Christen, Markus (2019), 'Two Concepts of Group Privacy', *Philosophy & Technology* 33(2), 207–224.
Google Scholar WorldCat

Manolios, Sandy, Hanjalic, Alan, and Liem, Cynthia C.S. (2019), 'The Influence of Personal Values on Music Taste: Towards Value-Based Music Recommendations', *RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems*, 10 September, 501–505. doi: https://doi.org/10.1145/3298689.3347021.

Martin, Kristen (2019), 'Ethical Implications and Accountability of Algorithms', *Journal of Business Ethics* 160(4), 835–850. doi: https://doi.org/10.1007/s10551-018-3921-3.
Google Scholar WorldCat

Milano, Silvia, Taddeo, Mariarosaria, and Floridi, Luciano (2020), 'Recommender Systems and Their Ethical Challenges', *AI and Society* 35(4), 957–967. doi: https://doi.org/10.1007/s00146-020-00950-y.
Google Scholar WorldCat

Mittelstadt, Brent Daniel, Allo, Patrick, Taddeo, Mariarosaria, Wachter, Sandra, and Floridi, Luciano (2016), 'The Ethics of Algorithms: Mapping the Debate', *Big Data and Society* 3(2), 1–21. doi: https://doi.org/10.1177/2053951716679679.
Google Scholar WorldCat

Mogensen, Andreas L. (2017), 'Moral Testimony Pessimism and the Uncertain Value of Authenticity', *Philosophy and Phenomenological Research* 95(2), 261–284. doi: https://doi.org/10.1111/phpr.12255.
Google Scholar WorldCat

Ng, Andrew Y., and Russell, Stuart J. (2000), 'Algorithms for Inverse Reinforcement Learning', *Proceedings of the Seventeenth International Conference on Machine Learning*, June, 663–670.

Nickel, Philip (2001), 'Moral Testimony and Its Authority', *Ethical Theory of Moral Practice* 4, 253–266. doi: https://doi.org/10.1023/A:1011843723057.
Google Scholar WorldCat

Nissenbaum, Helen (2001), 'How *Computer* Systems Embody Values', *Computer* 34(3), 120–119. doi: https://doi.org/10.1109/2.910905.

Nyholm, Sven (2018), 'Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci', *Science and Engineering Ethics* 24, 1201–1219. doi: https://doi.org/10.1007/s11948-017-9943-x.
Google Scholar WorldCat

O'Rourke, Dara, and Ringer, Abraham (2016), 'The Impact of Sustainability Information on Consumer Decision Making', *Journal of Industrial Ecology* 20(4), 882–892. doi: https://doi.org/10.1111/jiec.12310.
Google Scholar WorldCat

Olteanu, Alexandra, Castillo, Carlos, Diaz, Fernando, and Kıcıman, Emre (2019), 'Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries', *Frontiers in Big Data*, 11 July. doi: https://doi.org/10.3389/fdata.2019.00013.

Paul, Laurie A. (2014), *Transformative Experience* (Oxford: Oxford University Press).
Google Scholar Google Preview WorldCat COPAC

Paulo, Norbert (2018), 'Moral-Epistemic Enhancement', *Royal Institute of Philosophy Supplements* 83, 165–188.
Google Scholar WorldCat

Paulo, Norbert, and Bublitz, Christoph (2019), 'Introduction: Political Implications of Moral Enhancement', *Neuroethics* 12(1), 1–3. doi: https://doi.org/10.1007/s12152-018-9352-0.
Google Scholar WorldCat

Persson, Ingmar, and Savulescu, Julian (2012), *Unfit for the Future: The Need for Moral Enhancement* (Oxford: Oxford University Press).
Google Scholar Google Preview WorldCat COPAC

Peterson, Martin (2017), *The Ethics of Technology: A Geometric Analysis of Five Moral Principles* (Oxford: Oxford University Press).
Google Scholar Google Preview WorldCat COPAC

Phillips-Wren, Gloria (2012), 'AI Tools in Decision Making Support Systems: A Review', *International Journal on Artificial Intelligence Tools* 21(2), 1240005. doi: https://doi.org/10.1142/S0218213012400052.
Google Scholar WorldCat

Pugh, Jonathan (2019), 'Moral Bio-Enhancement, Freedom, Value and the Parity Principle', *Topoi* 38, 73–86. doi: https://doi.org/10.1007/s11245-017-9482-8.
Google Scholar WorldCat

Raus, Kasper, Focquaert, Farah, Schermer, Maartje, Specker, Jona, and Sterckx, Sigrid (2014), 'On Defining Moral Enhancement: A Clarificatory Taxonomy', *Neuroethics* 7(3), 263–273. doi: https://doi.org/10.1007/s12152-014-9205-4.
Google Scholar WorldCat

Rudin, Cynthia (2019), 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead', *Nature Machine Intelligence* 1(5), 206–215.
Google Scholar WorldCat

Sandel, Michael J. (2004), 'The Case against Perfection: What's Wrong with Designer Children, Bionic Athletes, and Genetic Engineering', *Atlantic Monthly* 293(3), 50–54, 56–60, 62.
Google Scholar WorldCat

Savulescu, J., and Maslen, H. (2015), 'Moral Enhancement and Artificial Intelligence: Moral AI?', in Jan Romportl, Eva Zackova, and Jozef Kelemen, eds, *Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics*, Vol 9 (Heidelberg: Springer), 79–95.
Google Scholar Google Preview WorldCat COPAC

Schaefer, G. Owen (2015), 'Direct vs. Indirect Moral Enhancement', *Kennedy Institute of Ethics Journal* 25(3), 261–289. doi: https://doi.org/10.1353/ken.2015.0016.
Google Scholar WorldCat

Schaefer, G. Owen, and Savulescu, Julian (2017), 'Better Minds, Better Morals: A Procedural Guide to Better Judgment', *Journal of Posthuman Studies: Philosophy, Technology, Media* 1(1), 26–43.
Google Scholar WorldCat

Schaefer, G. Owen, and Savulescu, Julian (2019), 'Procedural Moral Enhancement', *Neuroethics* 12, 73–84. doi: https://doi.org/10.1007/s12152-016-9258-7.
Google Scholar WorldCat

Serafimova, Silviya (2020), 'Whose Morality? Which Rationality? Challenging Artificial Intelligence as a Remedy for the Lack of Moral Enhancement', *Nature: Humanities and Social Sciences Communications* 7(1), 1–10.
Google Scholar WorldCat

Seville, Helen, and Field, Debora (2000), 'What Can AI Do for Ethics?', *AISB Quarterly* 104, 31–34.
Google Scholar WorldCat

Shook, John R. (2012), 'Neuroethics and the Possible Types of Moral Enhancement', *AJOB Neuroscience* 3(4), 3–14. doi: https://doi.org/10.1080/21507740.2012.712602.
Google Scholar WorldCat

Sinnott-Armstrong, Walter and Skorburg, Joshua A. (2021), 'How AI Can Aid Bioethics', *Journal of Practical Ethics* 9(1). doi: https://doi.org/10.3998/jpe.1175.

Google Scholar WorldCat

Skitka, Linda J., Bauman, Christopher W., and Sargis, Edward G. (2005), 'Moral Conviction: Another Contributor to Attitude Strength or Something More?', *Personality and Social Psychology Bulletin* 88(6), 895–917. doi: https://doi.org/10.1037/0022-3514.88.6.895.
Google Scholar WorldCat

Sparrow, Robert (2014), 'Egalitarianism and Moral Bioenhancement', *American Journal of Bioethics* 14(4), 20–28. doi: https://doi.org/10.1080/15265161.2014.889241.
Google Scholar WorldCat

Specker, Jona, Focquaert, Farah, Raus, Kasper, Sterckx, Sigrid, and Schermer, Maartje (2014), 'The Ethical Desirability of Moral Bioenhancement: A Review of Reasons', *BMC Medical Ethics* 15. doi: https://doi.org/10.1186/1472-6939-15-67.
Google Scholar WorldCat

Strohminger, Nina, and Nichols, Shaun (2014), 'The Essential Moral Self', *Cognition* 131(1), 159–171. doi: https://doi.org/10.1016/j.cognition.2013.12.005.
Google Scholar WorldCat

Susser, Daniel (2019), 'Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures', *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, January, 403–408. doi: https://doi.org/10.1145/3306618.3314286.

Tang, Tiffany Y., and Winoto, Pineta (2016), 'I Should Not Recommend It to You Even If You Will Like It: The Ethics of Recommender Systems', *New Review of Hypermedia and Multimedia* 22(1–2), 111–138. doi: https://doi.org/10.1080/13614568.2015.1052099.
Google Scholar WorldCat

Taylor, Linnet (2016), 'Safety in Numbers? Group Privacy and Big Data Analytics in the Developing World', in Linnet Taylor, Luciano Floridi, and Bart van der Sloot, eds, *Group Privacy: New Challenges of Data Technologies*, Vol. 126 (Cham: Springer). doi: https://doi.org/10.1007/978-3-319-46608-8.
Google Scholar Google Preview WorldCat COPAC

Taylor, Linnet, Floridi, Luciano, and van der Sloot, Bart, eds (2016), *Group Privacy: New Challenges of Data Technologies*, Vol. 126 (Cham: Springer).
Google Scholar Google Preview WorldCat COPAC

Vallor, Shannon (2015), 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character', *Philosophy and Technology* 28, 107–124. doi: https://doi.org/10.1007/s13347-014-0156-9.
Google Scholar WorldCat

Veliz, Carissa (2020), *Privacy is Power* (London: Penguin; Bantam Press).
Google Scholar Google Preview WorldCat COPAC

Whitby, Blay (2011), 'On Computable Morality: An Examination of Machines', in Michael Anderson and Susan Leigh Anderson, eds, *Machine Ethics* (Cambridge: Cambridge University Press), 138–150.
Google Scholar Google Preview WorldCat COPAC

Wright, Jennifer Cole, Cullum, Jerry, and Schwab, Nicholas (2008), 'The Cognitive and Affective Dimensions of Moral Conviction: Implications for Attitudinal and Behavioral Measures of Interpersonal Tolerance', *Personality and Psychology Bulletin* 34(11). doi: https://doi.org/10.1177/0146167208322557.
Google Scholar WorldCat

Yang, Qiang, Liu, Yang, Chen, Tianjian, and Tong, Yongxin (2019), 'Federated Machine Learning: Concept and Applications', *ACM Transactions on Intelligent Systems and Technology* 10(2), 1–19. doi: https://doi.org/10.1145/3298981.
Google Scholar WorldCat

Young, Garry (2018), 'How Would We Know If Moral Enhancement Had Occurred?', *Journal of Speculative Philosophy* 32(4), 587–606. doi: https://doi.org/10.5325/jspecphil.32.4.0587.
Google Scholar WorldCat

Notes

1 Much of this discussion has been presented in terms of enhancement. In contrast with past debates about human enhancement, though, which are generally concerned with improving a trait beyond what is normal or typical for some population, the question of moral enhancement is at least sometimes treated as including any sort of moral improvement, even if it does not surpass the normal or typical (e.g. DeGrazia 2014). We are concerned with improvement broadly, not just enhancement in the narrow sense, so to make that explicit we will use the term 'improvement' rather than 'enhancement' (see also Klincewicz 2019).
For additional discussion on the question of whether machines can aid human moral reasoning and decision-making, see Cave et al. (2019: 568); Boddington (2021).

2 We will not rely on a principled distinction between moral and non-moral values; we pragmatically count an individual's value as a moral value if that individual considers it to be a moral value.

3 We use this term in place of Giubilini and Savulescu (2018)'s 'artificial moral advisor' first, to avoid ambiguity about whether the AI system *itself* is moral/ethical and, second, because we want to consider not only AI systems that give advice but also those that provide assistance in other ways. (See Danaher (2018) for discussion on the ethics of AI assistants generally.) Throughout the chapter, we use the terms 'ethical' and 'moral' interchangeably.

4 A few examples of other types of uses that we will not discuss are: moral decision-making as a group (e.g. at the family, company, or city level) and moral decision-making by individuals who are in special positions, such as politicians, doctors, or designers.

5 Many people have highlighted the importance of this question within debates on moral enhancement; see, e.g. Shook (2012), DeGrazia (2014); Beck (2015), Young (2018).

6 For some discussion on this problem for moral enhancement see, e.g. Paulo (2018); de Sio and van den Hoven (2018). Indeed, one of the biggest hazards associated with ostensible moral improvement via technology is the prospect of some people using such technologies to impose their values on others with different values. Numerous people have addressed various aspects of this important topic, including, e.g. Sparrow (2014) and Paulo and Bublitz (2019).

7 There are also other approaches one could take to carve up the types of ways one might seek to morally improve. For instance, Shook (2012) distinguishes five ways in which one might morally improve; DeGrazia (2014) distinguishes three ways—motivational improvement, improved insight, and behavioural improvement.

8 For instance, one's intentions, motivations, and emotional reactions.

9 Some might dispute whether an agent can coherently or rationally believe that some unspecified subset of their own core moral beliefs and values might be wrong; this may raise an issue like the preface paradox. Nonetheless, (descriptively) at least some individuals have this belief.

10 Manolios et al. (2019) investigate the possibility that music recommendations might be based on personal values (see also Tang and Winoto 2016).

11 In reality, some internet services that use recommendation algorithms have attempted to remove content containing, for instance, hate speech (Hern 2018).

12 Giubilini and Savulescu (2018: 174) discuss an app, the Humane Eating Project, that promises to provide guidance on where to eat if one wants to avoid supporting animal cruelty.

13 For a number of years, the internet service GoodGuide supplied information about whether a product meets various standards, such as 'Leaping Bunny Certified', indicating lack of animal testing (O'Rourke and Ringer, 2016).

14 See also Conitzer et al. (2017); Freedman et al. (2020).

15 The distinction we are drawing between preparatory and on-the-spot assistance is a simplification: of course, people do not deliberate on a single moral question or case without interruption until they reach a conclusion. As a rough-and-ready distinction, however, it is useful.

16 Whether a given act of assistance is preparatory or on-the-spot is defined with reference to a particular moral question. In general, as one is considering a particular question, one cannot be trained to better deliberate about that particular question. (One might make an exception for a very lengthy period of deliberation over the course of which one's deliberation capacities could improve.) Any given instance of deliberation can be part of one's training in advance of future consideration of further cases.

17 See Schaefer and Savulescu (2017, 2019); see Paulo (2018) for discussion of procedural moral enhancement. Paulo (2018) argues that procedural moral enhancement amounts to *moral epistemic* enhancement.

18 For instance, Peterson (2017) suggests an approach to ethical reasoning that one could imagine a computer implementing

to help an individual evaluate their own moral consistency. The method involves identifying principles that best explain paradigm moral cases, assessing moral similarity between cases, and ascertaining what principle to apply to a new case by assessing which paradigm case the new case most closely resembles, morally. Suppose that the AEA possesses a model of the human's moral views, which represents the human's moral views geometrically in a 'moral space' of principles and cases as Peterson's describes. Suppose also that the system has learned to predict with some degree of reliability the human's pairwise moral similarity judgements between new cases and paradigm cases. The AEA could then indicate to the human the presence of a possible inconsistency in the human's view, in cases where the human's judgement about a new case diverges from what the AEA predicted the human would judge, on the basis of the AEA's prediction of the human's pairwise moral similarity judgements between the new case and paradigm cases and the AEA's knowledge of what principle is associated with each paradigm moral case.

19 For this intervention to be pertinent in the context of self-improvement, the user would have to consider these to be questions that are conducive to moral improvement or would need to believe more generally that the AEA's prompting method is conducive to moral improvement.

20 They emphasize, though, that the ultimate aim of the system would be to help the user get better at doing this reasoning on their own. This system, then, would be providing on-the-spot assistance as well as preparatory assistance for future cases.

21 Another possibility is that the system guides the user through other (non-reasoning) kinds of processes that can change one's ends or one's understanding of one's ends, such as meditation or life-transition traditions that involve hallucination-inducing drugs. (See Earp (2018) on the possible relevance of psychedelic interventions for moral enhancement).

22 Many potential ethical problems have already been identified for efforts to use technologies, particularly biomedical and genetic technology, for moral enhancement. For useful review, see Raus et al. (2014) and Specker et al. (2014).

23 Also see Milano et al. (2020) on privacy risks associated with recommender systems, and Kreitmair (2019: 158–159) on privacy worries related to direct-to-consumer neurotechnologies.

24 Federated learning (see, e.g. Yang et al. 2019) would be one strategy for reducing the transmission of specific information about individuals; it remains to be seen what form this might take and how much it would help in addressing the risks associated with sensitive moral data.

25 Already, digital technologies are changing this—for instance, some indication of values can be inferred using social media data, though such efforts are both patchy and noisy (Kosinski et al. 2013; Olteanu et al. 2019).

26 On this type of phenomenon, see Alfano et al. (2021)'s analysis of YouTube recommender systems as causing transformative experiences that may lead to radicalization.