**AI-Aided Moral Enhancement – Exploring Opportunities and Challenges**

**Abstract**

In this chapter, I introduce three different types of AI-based moral enhancement proposals discussed in the literature – substitutive enhancement, value-driven enhancement, and value-open moral enhancement. I analyse them based on the following criteria: effectiveness, examining whether they bring about tangible moral changes; autonomy, assessing whether they infringe on human autonomy and agency; and developmental impact, considering whether they hinder the development of natural moral skills. This analysis demonstrates that no single approach to AI enhancement can satisfy all proposed criteria, suggesting a need for pluralism in devising AI-based enhancement tools. The most advisable approach to AI-based enhancement is to explore and develop a variety of tools that cater to the diverse needs of users and address different weaknesses in moral decision-making.

**Keywords**: moral enhancement; AI-based moral enhancement; moral autonomy; moral agency; effectiveness of moral enhancement; development of moral skills; moral decision-making; moral mistakes; substitutive moral enhancement; auxiliary moral enhancement.

# 1 Introduction

Moral enhancement is broadly best understood as any kind of intervention that aims to improve a person's moral capacities, ultimately leading to moral improvement, such as better motives and behaviour (DeGrazia, 2014).[[1]](#endnote-1) The necessity of morally enhancing human beings is widely accepted, although there is significant disagreement about the appropriate means to achieve this. We inhabit a world plagued by morally troubling issues such as wars, poverty, and social and racial segregation. These pervasive problems, at least partially, can be viewed as manifestations of the inherent limitations of human nature, which lead to suboptimal moral reasoning. As Giubilini and Savulescu (2018, p. 170) put it, human beings are “suboptimal information processors, moral judges, and moral agents.” This means that we often fail to consider important information due to adverse physical conditions such as stress, hunger, or sleep deprivation. Additionally, we are susceptible to various reasoning fallacies, such as inconsistency and a range of implicit biases. Frequently, we experience cognitive dissonance (Festinger, 1957), where we hold incompatible beliefs or fail to align our behaviour with our values. We also exhibit in-group favouritism, the tendency to favour members of our own group over outsiders (see, for example, Everett et al., 2015). Persson and Savulescu (2012) argue that our evolved moral psychology, designed for life in small, close-knit groups, is inadequate for the complexities of the modern world. Given the challenges of global warming and weapons of mass destruction, they contend that moral enhancement is urgently needed for the survival of the human race. However, even if we agree on the necessity and urgency of moral enhancement, the most appropriate methods for achieving it remain a matter of debate.

The first and most obvious way of moral enhancement is through traditional means – education, socialisation, and the establishment of social and legal norms. In contrast, bioenhancement uses biological and technological means, such as pharmacological interventions, genetic engineering, neurotechnology, and hormonal treatments, to directly influence the cognitive and emotional capacities related to morality. The ethical debate about moral enhancement (Douglas, 2008, 2013; Harris, 2016; Persson and Savulescu, 2008, 2012, 2013; Rakić, 2014, 2017) usually revolves around moral bioenhancement and its advantages over traditional forms of moral enhancement, which are deemed slow and inefficient. Recently, there have been proposals to use Artificial Intelligence (AI) for moral enhancement (Savulescu and Maslen, 2015; Giubilini and Savulescu, 2018; Lara and Deckers, 2020; Lara, 2021; Volkman and Gabriels, 2023). Although this type of enhancement, since it is intended to use technology for moral enhancement, could be subsumed under the umbrella of bioenhancement, it is usually viewed as a separate type since it does not typically include any form of direct bodily intervention.[[2]](#endnote-2) For the purposes of this chapter, I will consider AI-based moral enhancement (AIME) as any form of moral improvement that uses AI-based technology. When discussing AI, I refer to systems capable of performing tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, and language translation.[[3]](#endnote-3) AI encompasses a broad range of technologies designed to simulate human-like intelligence in machines, enabling them to learn from data, adapt to new inputs, and perform tasks autonomously. For example, this includes machine learning techniques like deep neural networks and natural language processing tools such as ChatGPT. Since AI finds broad application in aiding decision-making (e.g., in healthcare, legal systems, banking), it is natural to consider whether it can be used to improve our moral decision-making. Additionally, many proposals for AI moral enhancement see this option as preferable to more invasive bioenhancement, proposing it as a superior alternative (e.g., Lara and Deckers, 2020).

The aim of this chapter is to evaluate some of the most prominent proposals for AI-based moral enhancement against different parameters for evaluating moral enhancement generally: effectiveness, effects on human autonomy and agency, and influence on human moral skills development. Other types of moral enhancement are not the subject of this chapter, except as background for comparison with AI moral enhancement. The debate on bioenhancement, which has been present on the philosophical scene for a longer time, will be utilized as a valuable source of general insights into the nature, implications and ways to assess moral enhancement proposals.

# 2 AI-Based Moral Enhancement: General Background

First of all, it is necessary to point out that all the proposals for AI moral enhancement are speculative. They do not analyse any existing tools but instead propose tools that could and should be developed in the future. To my knowledge, no tool specifically aimed at moral enhancement has been developed to date. Although multipurpose tools such as ChatGPT can be used for asking for moral advice and discussing moral topics, these tools are not designed with this purpose in mind. As already mentioned, at least part of the inspiration and justification for proposals for AI moral enhancement is the fact that AI-based systems are already being used as decision-making aids in many fields. The advantages that AI systems bring are numerous. They surpass humans in terms of efficiency and speed as data processors, can process huge amounts of data in short periods of time, and consider a larger portion of data and parameters when making a decision (Rodríguez-López and Rueda, 2023; Berber and Srećković, 2023). Given all this, these systems can extract patterns and information from data that humans would miss (Srećković et al., 2022). However, it should also be kept in mind that the use of AI systems for decision-making brings many challenges: these systems can be biased and error-prone, and there is still no consensus on how to ascribe moral responsibility for decisions made using these systems (Berber and Srećković, 2023; Berber and Mijić, 2024).

## 2.1 The General Classification of Proposals

When it comes to the general classification of AIME proposals, they can be classified as either exhaustive or auxiliary (Lara and Deckers, 2020). Exhaustive methods would imply completely outsourcing moral decision-making to the AI system. The system would make a moral recommendation, and humans would simply execute it without actively involving themselves in the moral decision-making process. Thus, exhaustive moral enhancement would mean giving up the human role as moral decision-makers and completely delegating it to AI. The human role would be reduced to accepting and executing the system’s recommendations. Of course, humans would need to build this system in the first place, but after that, they would be passive recipients of the system’s instructions. On the other hand, auxiliary enhancement involves using AI systems as support to make better decisions. Depending on the model of AI moral enhancement, the systems are envisioned to help humans with different aspects of moral decision-making. However, essentially, humans remain 'in the loop' and actively participate in the decision.

Another way to differentiate between various proposals for AIME is based on their intended improvement targets. Liu and colleagues (2022) distinguish between 'narrow' and 'broad' moral enhancement. Narrow moral enhancement involves systems that assist users in making moral decisions by pointing to the right decision outcome, whereas broad moral enhancement involves systems designed to improve users' moral character. Yet another distinction introduced in the context of AIME is between the operational types of AI-based moral enhancement systems: outcome-oriented and process-oriented navigation (Poszler and Lange, 2024). The first type, outcome-oriented systems, points the user towards a specific moral decision outcome, which is conditioned by previously selected and embedded principles. Since it targets the decision output directly, this type can be said to offer narrow enhancement. On the other hand, process-oriented navigation systems do not strive for a predetermined decision outcome but support and guide users through their own moral decision-making process without aiming to indoctrinate particular ethical principles or values (Poszler and Lange, 2024). Process-oriented systems usually aim to enhance a person's moral reasoning capacities, thereby bringing about broad moral enhancement. Finally, we can think of AIME as targeting either moral behaviour or moral capacities (Raus et al., 2014). Typically, the former would be narrow and outcome-oriented, while the latter would be broad and process-oriented.

## 2.2 The Parameters for Assessing AIME Proposals

The criteria I will use to assess the proposed methods of AIME are those that have emerged from the broader debate on moral enhancement, where different proposals are often criticized for failing to meet or infringing upon certain standards. Based on my research of the enhancement literature, I have identified the three most important criteria for successful moral enhancement. Regarding the conceptual origin of these parameters, I would say that they are, in a sense, intuitive and follow naturally from the very idea of moral enhancement. However, it is possible that someone might contest the soundness or importance of a certain parameter relative to their own conception of moral enhancement. My aim is to provide a more or less unproblematic (if this is even possible in philosophy) overview of the AIME proposal, although I do not claim that this kind of evaluation will be universally acceptable to everyone.

The first thing we should take into account is whether AIME proposals are effective, i.e., whether they actually bring about moral improvements. Some methods may intend to produce particular moral improvements but fail to do so (Shook, 2012; Raus et al., 2014). The obvious way to demonstrate effectiveness is to manifest tangible effects on the moral behaviour outputs, i.e., better moral decisions. However, if we choose to understand effectiveness in this manner, it seems to favour methods of enhancement that are oriented towards changing outputs, or in terms of the introduced distinctions, narrow, outcome-oriented, and behaviour-targeting methods. But if we look closer at the motivation behind the moral enhancement project, which is to make a positive moral change in the world, it makes sense to expect that all methods of enhancement, regardless of whether they directly target the outcomes of moral decision-making, should ultimately result in enhancing these outcomes. Thus, methods aimed at enhancing the moral character and capacities of users should lead to better moral decisions and manifest in users' moral behaviour. This way of understanding effectiveness has some implications for the distinctions introduced in the previous paragraph. It implies that ultimately, all methods that bring about broad enhancement should also result in narrow enhancement. All process-oriented methods should ultimately lead to changed outputs, and all capacity-targeted methods should be behaviourally manifested, but not necessarily *vice versa*.

Another important parameter is how AIME affects the autonomy of persons using the proposed system. When I speak of autonomy, I refer to several aspects of a person's self-directedness that have all emerged as important topics in the debate on moral enhancement. The first aspect is freedom from coercion (Rakić, 2014, 2017), a concept tracing back to Aristotle and deemed essential for ascribing moral responsibility according to the control or choice model of moral responsibility introduced by Fischer and Ravizza (1998). According to this model, an agent is not morally responsible for an act if it was not fully under their control or if they did not act voluntarily. Thus, if actions are performed under coercion, whether physical or chemical (such as the effects of drugs), the agent is not responsible, and we cannot attribute blame or praiseworthiness. Even if coercion leads to better moral outcomes, it does not count as moral enhancement of the agent because these acts cannot be considered the agent's own in the strict sense.

The second aspect of autonomy is the freedom to fall and choose morally wrong actions. Persson and Savulescu (2012b) introduced a thought experiment featuring a "God machine" that monitors everyone's intentions and intervenes only if someone forms an intention to commit a great evil, such as murder or rape, by altering that intention and preventing the crime before it happens. Although freedom in all other cases would be preserved, many authors argue that this type of intervention would infringe on a person's autonomy. It would prevent them from committing morally wrong acts, meaning that refraining from wrongdoing couldn't be credited to the individual and, thus, wouldn't count as moral enhancement (Harris, 2011, 2014).

The third aspect of autonomy is authenticity or self-directedness, viewed as a positive aspect of freedom that goes beyond mere protection against coercion or choice limitation. It entails acting according to one's own beliefs and values, which also implies freedom from manipulation. The crucial idea is that the autonomous agent may indeed rely on others and delegate various tasks to them, whether technological tools or other humans, but only to the extent that doing so does not compromise their own capacity for self-direction (Carter, 2017). Thus, autonomy involves relying on others to an appropriate extent without relinquishing autonomy, which would entail becoming others-directed rather than self-directed (Roberts and Wood, 2007).

To summarise, preserving autonomy in all its aspects is important in this context because if a method prevents persons from acting autonomously, it is hard to see how it enhances them. To be morally enhanced, it is necessary to maintain one's status as a moral agent, and this status requires autonomy.

The last parameter I will consider is how using an AIME system affects our natural moral skills. By "natural," I mean the skills we would have without the AI system. Unlike the previous parameters, this question did not feature prominently in the moral bioenhancement debate but has emerged in the context of AI-based enhancement. Generally, the question of how different technologies affect our natural abilities has been considered in various contexts, dating back to Plato’s concerns about how the invention of writing would impact our natural memory abilities to more contemporary worries about whether internet usage negatively affects our memory and reasoning abilities (Sparrow et al., 2011; Greenfield, 2014). Taking this parameter into account may seem to favour broad, process-oriented, and capacity-oriented approaches to moral enhancement, as they directly aim to enhance a person's moral skills. To avoid this bias, this parameter is best understood negatively—the method for enhancement should not lead to the loss of natural moral skills or prevent their development. Erosion of natural abilities can render people in a state of learned helplessness, where they are unable to direct their cognitive affairs without the enhancement tool in question (Carter, 2017; Lynch, 2016). This would, consequently, infringe on the third aspect of autonomy, which is self-directedness. Of course, it would be desirable for AIME to have a positive impact on our natural abilities, but this should not be seen as a disqualifying condition.

In the following section, I will introduce three different types of AIME proposals that are discussed in the literature – substitutive enhancement, value-driven enhancement, and value-open moral enhancement – and analyse them based on the parameters mentioned above. Based on this analysis, I will offer some general insights on the prospects for moral enhancement using AI-based technologies in the last section of the chapter.

# 3 Types of AIME

# 3.1 Substitutive Moral Enhancement

The main idea behind these proposals is that, given humans are suboptimal moral decision-makers, AI-based decision systems—if developed to be superior to humans—could eventually substitute human decision-making in moral matters. Some discussions of these possibilities are presented as proposals for moral enhancement (Gips, 1995; Dietrich, 2001), while others are treated more as thought experiments designed to explore important topics concerning human moral agency and autonomy (e.g., Beavers, 2011; Berber, 2023). Dietrich (2001) advocated a utopian idea that human beings should be replaced by ethical robots. He refers to these robots as "Homo sapiens 2.0," as they are envisioned to be enhanced versions of human beings, which will overcome human biological conditioning and realize that they are not the centre of the universe and act accordingly. Gips (1995) discusses the development of ethical robots capable of determining actions that result in the best possible outcomes. These robots could be used to override or correct human decisions. Beavers ponders the consequences of having and using an all-knowing moral machine that would have the answer to what is the right thing to do in every morally contentious situation. This type of moral enhancement would be exhaustive since it is aimed at substituting human decision-making. Additionally, it is narrow, outcome-oriented, and behaviour-focused, as humans would simply execute the prescribed system's outputs without involvement in the decision-making process.

These types of AIME face numerous challenges and associated risks. The first issue is attainability – whether we will be able to develop AI systems that are superior to humans in moral decision-making and what the best strategy to achieve this would be. This problem is extensively discussed in relation to the machine ethics project, which aims to create autonomous artificial moral agents (Anderson and Anderson, 2007; Sison and Redín, 2023; Sparrow, 2021). Although the machine ethics project is not explicitly aimed at moral enhancement, if successful, it would open up the possibility for substitutive enhancement. Nonetheless, the dangers and risks of using AI-based systems for substitutive moral enhancement are manifold, as analysed in depth by Lara and Deckers (2020). There is a possibility that the AI system’s designers or the system itself might promote a hidden agenda and that there could be flaws in the system that are not easily detectable. Since the stakes are high – practically the fate of all humankind – placing faith in an artificial system, no matter how superior it may seem, is very dangerous. In fact, the very superiority of the system would make humans vulnerable.

For argumentative purposes, I will assume that we have successfully developed AI-based systems that are superior to humans and can be used as all-knowing moral oracles. As previously mentioned, this type of enhancement does not have many proponents and is often treated more as a thought experiment or a backdrop for discussing less utopian proposals. I will briefly comment on it with this perspective in mind, aiming to gain a broader understanding of AI enhancement in general. When it comes to effectiveness, this approach to moral enhancement would bring about swift and tangible change. If humans followed the recommendations of the morally superior system, it would lead to better moral decisions and reduce the amount of human-induced evil (Berber, 2023), aligning with the core idea of moral enhancement. On the other hand, concerning the requirement not to infringe on human autonomy, this approach does not score well. Of the three aspects of autonomy, only the first one—the absence of coercion—is satisfied, assuming that the use of these AI systems would not be enforced on humans but rather chosen freely. (This issue is not specified in scenarios of substitutive enhancement, but if the use of these systems were mandated by law, this aspect of autonomy would be nullified as well.) Thus, some minimal sense of autonomy would be retained. This situation is similar to the model proposed by Rakić (2014, 2017) in the context of bioenhancement, which suggests that techniques undermining freedom can still be considered moral enhancements if chosen freely and voluntarily by the person subjected to them. However, the freedom to fall would be cancelled since, by following the system, humans would not be able to make moral mistakes. Even if the system is fallible and makes a mistake, the human user would not be responsible. Additionally, self-directedness is completely undermined as humans are guided by the system. Finally, the development of human moral skills would be undermined by the use of this kind of system, as humans would never have to decide for themselves and, therefore, wouldn't have the opportunity to develop moral reasoning skills. To be fair, the system could be designed to provide explanations or rationales for its outputs, but without the active involvement of humans in the decision-making process, this would not suffice to develop moral reasoning capacities.

In conclusion, the substitutive systems excel primarily in terms of effectiveness, while two other parameters pose significant challenges. Moreover, even this favourable effectiveness score is contingent upon the perhaps unrealistic prospect of developing systems superior to humans in moral decision-making, and accepting all the associated risks of using them.

# 3.2 Value-Driven Moral Enhancement

The value-driven approach is presented in two proposals by Savulescu and Maslen (2015) and Giubilini and Savulescu (2018). These proposals advocate for the development of AI-based moral personal assistants that can be customized by users to align with their moral values and beliefs. In the first proposal, users select and prioritize basic values from a list provided by the system according to their personal preferences. The system then processes the available information based on this hierarchy of values and offers recommendations for moral behaviour. In the second proposal, Giubilini and Savulescu suggest that users choose an advisor system from the available preset options that align with their personal values. The system would then recommend decisions that a hypothetical ideal observer—omniscient, omnipercipient, impartial, dispassionate, and consistent—who shares the user's value perspective would make in specific situations. These systems are not intended to change or affect users' moral values and principles but to help users implement them more efficiently. Thus, the system itself is value-neutral—the users can set up values as they wish. But once the user sets the values, the system is driven by this choice (hence, I decided to term this approach 'value-driven' since the moral enhancement of the users is driven by their own values).

The proposed systems would help users gather all relevant information for particular decisions. They would help users overcome their cognitive limitations, reach decisions more quickly and efficiently, and remain more consistent with their own moral standards. Additionally, they would protect users from the distorting effects of being too personally and emotionally involved in the situation or from adverse physical conditions such as tiredness and sleep deprivation at the time of making a decision, which could negatively affect moral reasoning. Giubilini and Savulescu illustrate the type of advice the artificial moral assistant would give users in the following manner: "For example, it would tell us where exactly to throw our empty cup once we have told it that we want to make an environmentally friendly choice, or which eggs to buy once we have told it we care about animal welfare" (2018: p. 171). These systems would essentially guarantee more thoughtfulness, awareness, and consistency, safeguarding the user from failing their own values. In terms of the classificatory distinctions introduced in the previous section, this approach is auxiliary, meaning that it helps the user but doesn’t supplant them. It is narrow, outcome-oriented, and behaviour-targeted. The system primarily aims to point the user to a specific decision outcome and behaviour, without intending to change the user’s character or improve their reasoning abilities.

Concerning effectiveness, this approach would certainly bring about changes in users' behaviour, as the users would receive precise guidelines on what actions to take in order to follow their own values and principles. This would be helpful in terms of enhancing executive skills. However, there is a danger that this approach, despite its promise to yield real changes in users' behaviour, may not result in genuine moral enhancement. If users input the wrong values from the start, making them more consistent and thoughtful would not be sufficient for genuine moral enhancement (Klincewicz, 2016). Thus, if the system does not provide any normative guidance and serves only to ensure that users' normative values are respected and implemented, it may easily fail to morally enhance the users. On the other hand, if we limit or forbid certain values from the start as a system's limitation, we infringe on users' autonomy in terms of their freedom to fall and their self-directedness. As proposed, driven by users' values, this approach fully respects the users' autonomy in all of its aspects, and this is seen as a very important strength by its proponents.

The objection against the value-driven approach is that since it doesn't require users "to understand the rational connections between their values and the decisions that are made by the system, their moral skills may not be enhanced a great deal" (Lara and Deckers, 2020). However, as I already said, not enriching our moral skills cannot be taken as a decisive reason against some approach. But we can push this objection a step further and suppose that constantly using AI assistance for decision-making could cause our natural reasoning abilities to erode, similar to some other technologies (e.g., using satellite navigation causes our natural orientation abilities to atrophy as shown by McKinlay, 2016). Lara and Deckers made yet another objection that using a value-driven moral personal assistant would prevent users' personal moral growth in the sense that users would be denied the opportunity to change and evolve their values and principles; they wouldn’t develop but get stuck with the value set they initially chose in the system.

Although the value-driven personal moral assistant is an output-oriented enhancement system primarily targeting decisions rather than moral reasoning, it is not obvious that it would negatively affect users' moral skills. Under closer scrutiny, it may even enhance them. For example, an important part of understanding a value or principle is comprehending its practical implications and how this proclaimed value translates into practice. Thus, by allowing the user to be exposed to impartial and consistent implications of their proclaimed values, the system can help the user gain a better understanding of those values. Additionally, even when not using the system, it is not implausible to suppose that the user would apply this implicit knowledge and become better at moral decision-making. Furthermore, when seeing how some of their values translate into practice and what their real-world implications are, the user may be prompted to question these values. And since the moral personal assistant certainly wouldn't be set to prevent the user from updating their values, the user could change the initial settings, which would constitute personal moral development. Ultimately, maybe the best way to understand and test our beliefs is to consistently and thoroughly put them into practice.

To summarize: This approach preserves autonomy. It may even foster the development of moral skills; however, this is uncertain as it might also impair them in the long run, with arguments available for both positive and negative perspectives. Unfortunately, its effectiveness in guaranteeing better moral outcomes is not assured, as it is dependent on the users choosing the right values to begin with.

# 3.3 Value-Open Moral Enhancement

Lara and Deckers (2020) proposed the development of a Socratic moral adviser or interlocutor, a concept further elaborated by Lara in 2021. Their approach emphasizes continuous interaction between the system and the user. Rather than relying on pre-established lists or sets of values to enhance the user's morality, their method fosters ongoing dialogue, increasing the likelihood of the user's values evolving through this interaction. The key is that the system aids users in arriving at better decisions independently, without binding them to any predetermined ethical frameworks. Users' decisions would emerge from engaging with a system that does not come with a built-in set of ethical guidelines. This proposal draws inspiration from the deliberative exchange in Socratic philosophy as a means to develop better moral judgments (Lara and Deckers, 2020). The system would provide a moral assessment achieved by applying criteria of empirical, logical, and ethical rigour to the user's convictions. The user would occupy a position of privilege, initiating the initial resolution of a significant conflict. This resolution would then undergo systematic examination by the system, posing pertinent inquiries to uncover potential weaknesses in the user's reasoning. The system's role is to enhance the user's moral reasoning process, remaining value-neutral or value-open: it doesn't advocate any predetermined value agenda, although it may assist the user in reshaping their values through critical exchange with the system.

Volkman and Gabriels (2023) expand on the concept of an AI-based Socratic interlocutor by proposing a modular system comprising multiple AI interlocutors, each with unique perspectives shaped by their training in various wisdom traditions. The idea is to ensure moral progress through engagement with an AI system that incorporates modules based on diverse traditions such as Stoicism, Rawlsian theory of justice, Socratic philosophy, Confucianism, and other moral traditions. Each AI module or mentor would offer a unique perspective, engaging in continuous dialogue not only with the user but potentially with one another as well. Such an AI would provoke meaningful engagement like books and other information technologies, but its interactivity promises a richer and more compelling experience. The authors highlight that this approach reduces the risk of moral disengagement, and the presence of multiple modules from diverse traditions guarantees the preservation of pluralism. Although each module has its own value perspective, none of these perspectives is imposed on the users, as they are free to compare and confront them. This ensures that value neutrality is maintained.

Both proposals for a Socratic interlocutor and a multimodal pluralistic interlocutor aim at broad enhancement, intending to improve users' moral skills. They are process-oriented and capacity-oriented, as they do not impose any particular decision choice but focus on enhancing the user's moral reasoning abilities. Concerning the parameters for assessing AIME proposals, I will start with the favourable ones. Value-open types of enhancement preserve all aspects of autonomy. The users are fully involved, engaged, and in charge of the whole process, with no values being imposed on them. When it comes to enhancing users' moral skills, particularly their moral reasoning abilities, the proposed systems should excel. These systems challenge users' choices, prompting them to explain and reconsider their initial stance. They offer alternative viewpoints and provide a background understanding of the normative theories behind the issues they address. This approach should not only improve users' moral reasoning abilities concerning specific matters but also foster a broader habit of moral reasoning and self-questioning. In this way, value-open tools for moral enhancement could serve as significant aids in moral education, acting as personal moral instructors or guides for users.

Nevertheless, this raises the question of the advantages these AI systems offer over traditional forms of moral enhancement such as moral education, including taking intensive courses in normative ethical theory, seeking advice from friends or counsellors, or reading books on morality. One immediate advantage is accessibility. These AI systems can consolidate various resources and perspectives into a single platform or application on our phones, making them more readily available. Additionally, Lara and Deckers (2020) highlight that users may be more inclined to accept advice from machines than from other humans, particularly if the machines demonstrate high effectiveness rates. Although these interlocutor systems may bring the discussed advantages, it is hard to see them as anything else but advanced moral education tools. Many disadvantages of traditional moral education would emerge here as well: it would be time-consuming, slow, and ineffective, and many users would just lack persistence and motivation to engage with these systems. Additionally, the abundance of diverse viewpoints presented by the system proposed by Volkman and Gabriels could overwhelm users, leading to indecisiveness, moral relativism, or even nihilism. As users engage with various ethical frameworks and conflicting values, they may struggle to discern a clear path forward, ultimately questioning the validity of moral principles altogether. Similarly, as Lara and Deckers' Socratic interlocutor offers questioning and counterperspectives, it could fail to provide users with guidance and clarity while navigating complex moral dilemmas, and fail to lead to conclusive moral decisions.

 In conclusion, although this proposal scores well on the preservation of autonomy and development of moral skills, its effectiveness is questionable.

# Table 1: The Summary of Section 3 Analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type of AI-Based Moral Enhancement** | **Effectiveness** | **Impact on Autonomy** | **Impact on Moral Skills Development** | **Challenges** | **Potential Benefits** |
| **Substitutive** | Highly effective if AI surpasses human decision-making, leading to quick and tangible moral improvements. | Significant infringement on autonomy: Users lose the freedom to fall and self-directedness. Autonomy is limited to freedom from coercion. | Negative impact: Users do not engage in decision-making, leading to a loss of moral reasoning and impeding skills development. | Dependence on AI's superiority over humans in moral reasoning; risk of hidden agendas or system flaws; loss of human moral agency. | Reduces the likelihood of human error in moral decisions; ensures consistency and objectivity. Provides fast and tangible results. |
| **Value-Driven** | Effective in aligning behaviour with pre-set values; it may not result in genuine moral enhancement if the input values are flawed. | Preserves autonomy: Users retain the freedom to set their values, including the freedom to fall and self-directedness. | Can be both positive and negative: It can enhance implicit understanding of values, but it also can cause atrophy of moral reasoning skills due to overreliance. | Risk of reinforcing flawed or harmful values; potential for stagnation in personal moral growth. | Allows users to act consistently with their values, even under stress or cognitive limitations; can support users in upholding personal moral standards. |
| **Value-Open** | Questionable effectiveness: Encourages moral reasoning and self-questioning, it may be slow and less conclusive. | Fully preserves autonomy: Users engage in dialogue, retain control, and are free from imposed values. | Positive impact: Focuses on enhancing moral reasoning abilities and promoting moral development through dialogue and reflection. | Can be time-consuming; and risks overwhelming users with diverse viewpoints, leading to indecisiveness or moral relativism. | Enhances moral reasoning and self-reflection; fosters moral growth and adaptability. |

# 4 General Prospects for AIME

Based on the analysis of the proposed AIME systems, we can observe an inherent tension in the very idea of moral enhancement through non-traditional means such as bioenhancement or AI-based enhancement. On one hand, the urgency and drive to seek these new techniques for enhancement stem from the slowness and ineffectiveness of traditional methods, creating an imperative for swift, effective, and tangible moral change in the rapidly changing world we live in. On the other hand, there is also an imperative to preserve human moral agency and autonomy and to develop human moral skills, which significantly limits the project. Thus, there is a noticeable trade-off between effectiveness and protecting human autonomy and agency.

As we saw, in the case of substitutive enhancement, autonomy would be infringed upon in two significant aspects (freedom to fall and self-directedness), but the change would be swift and highly effective (provided of course that we manage to design AI systems superior to humans in moral decision-making). In the value-driven approach, allowing users to choose their own values guarantees respect for all the aspects of autonomy, including self-directedness, but could result in individuals improving their ability to act on wrong values, leading to ineffectiveness in making the world better. Ultimately, in the value-open case, where all aspects of autonomy and human skill development are safeguarded, we find ourselves circling back to traditional moral education, albeit now supplemented by AI-based tools. The very ineffectiveness of traditional moral education which initially spurred the quest for moral bio- and AI- enhancement reappears. This means that the pursuit of AI-based moral enhancement must carefully navigate the tension between effectiveness, on the one hand, and autonomy, agency, and human moral development, on the other hand. Highly efficient systems that provide swift and tangible changes in behaviour risk undermining user autonomy, agency, and moral development, while systems respecting all of these may lead to suboptimal moral outcomes and prove ineffective.

Maybe the very idea that we should propose an AI moral enhancement model that excels in all the aspects required for successful moral enhancement is flawed. Building upon Volkman and Gabriels' pluralistic approach of integrating multiple modules into a single AI system for enhancement, we might consider adopting a similar perspective when envisioning and developing AIME tools. This would involve developing diverse tools with varying aims. Thus, some of these tools may be more suitable to target particular problems and errors in moral decision-making, while others can be suitable for others. For instance, if we do not have a firm and elaborated stance on particular moral topics, such as the question of animal rights, we may want to use the system proposed by Volkman and Gabriels to gain perspective. If we have a stance on these topics but are not quite sure that we are right, the Socratic interlocutor tool proposed by Lara and Deckers may be more suitable for us. Finally, if we are quite convinced in our stance concerning animal rights, the personal assistant proposed by Savulescu and Giubilini may be the right choice for us. It is important to recognize the diverse needs of users and that diverse tools may be adequate for satisfying these needs.

Moral decision-making is inherently complex and can fail for various reasons. It might be unrealistic to expect that a single AI tool, no matter how advanced, could address all potential causes of failure. Additionally, the trade-off between effectiveness and human agency and autonomy teaches us that we cannot have it all, at least not in one tool. The presence of trade-offs between different values poses a significant problem for the development of the AI sector in general. Different principles and values we cherish and want to protect may collide with each other; for example, data protection and privacy might endanger safety and efficiency, just as safety and efficiency might come at the cost of human autonomy and agency (Berber, 2023; Whittlestone et al., 2019).

While authors proposing AI enhancement methods often present their own models as superior to previously proposed ones, it is flawed to think about this topic in terms of 'winner takes it all.' There is certainly room for different tools and applications to be developed and utilized in this field. In some cases, it may even be necessary to combine AI tools with certain methods of bioenhancement. The main takeaway of this chapter is that we should encourage exploring a diversity of potential methods for AIME, as this may be the most methodologically sound approach to making real progress in morally enhancing human beings and creating a better world, which are the ultimate goals of the moral enhancement debate.

# Notes

1. Later in the paper, we will see that some approaches target moral behavior directly, bypassing the improvement of moral capacities. [↑](#endnote-ref-1)
2. To be fair, we can imagine an AI-based microchip designed to influence moral reasoning being implanted in a person's head. This case could blur the distinction between bioenhancement and AI enhancement; however, having a clear-cut distinction between these types of enhancement is not a matter of great importance for the debate. [↑](#endnote-ref-2)
3. Whether these systems should be considered truly intelligent and what kind of intelligence should be attributed to them is a matter of philosophical debate. Some authors believe it is misleading to call them intelligent (e.g. Floridi 2020). However, we will adhere to the colloquial and widely accepted use of the term artificial intelligence, without delving into the underlying philosophical controversies.

Acknowledgements

I would like to thank my colleagues Jelena Mijić and Nenad Filipović for their kindness and patience in reading and commenting on early versions of this chapter.

# References

Anderson, M. and Anderson, S.L. (2007). Machine ethics: creating an ethical intelligent agent. *AI Magazine*, 28(4), pp. 15–26.

Beavers, A. (2011). Moral machines and the threat of ethical nihilism. In: P. Lin, G. Bekey and K. Abney, eds. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: MIT Press, pp. 333–344.

Berber, A. and Mijić, J. (2024). Understanding moral responsibility in automated decision-making: responsibility gaps and strategies to address them. *Theoria*, 67(3): 177-192

Berber, A. (2023). Automated decision-making and the problem of evil. *AI & Society*. <https://doi.org/10.1007/s00146-023-01814-x>.

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

Carter, J.A. (2020). Intellectual autonomy, epistemic dependence and cognitive enhancement. *Synthese*, 197, pp. 2937–2961.

DeGrazia, D. (2014). Moral enhancement, freedom, and what we (should) value in moral behaviour. *Journal of Medical Ethics*, 40(6), pp. 361–368.

Dietrich, E. (2001). Homo sapiens 2.0: Why we should build the better robots of our nature. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(4), pp. 323–328.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy*, 25(3), pp. 228–245.

Douglas, T. (2013). Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics*, 27(3), pp. 160–168.

Everett, J.A., Faber, N.S. and Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in Behavioral Neuroscience*, 9, p. 15. <https://doi.org/10.3389/fnbeh.2015.00015>.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Evanston, IL: Row, Peterson.

Fischer, J.M. and Ravizza, M.S.J. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36, p. 15.

Gips, J. (1995). Towards the ethical robot. In: K.M. Ford, C. Glymour and P. Hayes, eds. *Android Epistemology*. Cambridge: MIT Press, pp. 243–252.

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

Greenfield, S. (2014). *Mind Change: How Digital Technologies are Leaving Their Mark on Our Brains*. London: Rider Books.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics*, 25(3), pp. 102–111.

Harris, J. (2014). Taking liberties with free fall. *Journal of Medical Ethics*, 40(6), pp. 371–374.

Harris, J. (2016). *How to be Good. The Possibility of Moral Enhancement*. Oxford: Oxford University Press.

Lara, F. (2021). Why a virtual assistant for moral enhancement when we could have a Socrates? *Science and Engineering Ethics*, 27(4), p.42.

Lara, F. and Deckers, J. (2020). Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics*, 13, pp. 275–287.

Liu, Y., Moore, A., Webb, J. and Vallor, S. (2022). Artificial moral advisors: a new perspective from moral psychology. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 436–445.

Lynch, M.P. (2016). *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. London: W.W. Norton.

McKinlay, R. (2016). Technology: Use or lose our navigation skills. *Nature*, 531(7596), p. 573–575.

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), pp. 162–177.

Persson, I. and Savulescu, J. (2012). *Unfit for the Future*. Oxford: Oxford University Press.

Persson, I. and Savulescu, J. (2013). Getting moral enhancement right: the desirability of moral bioenhancement. *Bioethics*, 27(3), pp. 124–131.

Plato (1925). *Phaedrus*. Translated by H.N. Fowler. Cambridge, MA: Harvard University Press.

Poszler, F. and Lange, B. (2024). The impact of intelligent decision-support systems on humans' ethical decision-making: a systematic literature review and an integrated framework. *Technological Forecasting and Social Change*, 204. <https://doi.org/10.1016/j.techfore.2024.123403>.

Rakić, V. (2014). Voluntary moral enhancement and the survival-at-any-cost bias. *Journal of Medical Ethics*, 40, pp. 246–250.

Rakić, V. (2017). Compulsory administration of oxytocin does not result in genuine moral enhancement. *Medicine, Health Care and Philosophy*, 20, pp. 291–297.

Raus, K., Focquaert, F., Schermer, M., Specker, J. and Sterckx, S. (2014). On defining moral enhancement: a clarificatory taxonomy. *Neuroethics*, 7, pp. 263–273.

Roberts, R.C. and Wood, W.J. (2007). *Intellectual Virtues: An Essay in Regulative Epistemology*. Oxford: Oxford University Press.

Rodríguez-López, B. and Rueda, J. (2023). Artificial moral experts: asking for ethical advice to artificial intelligent assistants. *AI Ethics*, pp. 1–9.

Savulescu, J. and Maslen, H. (2015). Moral enhancement and artificial intelligence: moral AI? In: J. Romportl, E. Zackova and J. Kelemen, eds. *Beyond Artificial Intelligence*. Springer Publishing, pp. 79–96.

Savulescu, J. and Persson, I. (2012). Moral enhancement, freedom and the God Machine. *The Monist*, 95(3), pp. 399–421.

Shook, J.R. (2012). Neuroethics and the possible types of moral enhancement. *AJOB Neuroscience*, 3(4), pp. 3–14.

Sison, A.J.G. and Redín, D.M. (2023). A neo-aristotelian perspective on the need for artificial moral agents (AMAs). *AI & Society*, 38, pp. 47–65.

Sparrow, B., Liu, J. and Wegner, D.M. (2011). Google effects on memory: cognitive consequences of having information at our fingertips. *Science*, 333(6043), pp. 776–778.

Sparrow, R. (2021). Why machines cannot be moral. *AI & Society*, 36, pp. 685–693.

Srećković, S., Berber, A. and Filipović, N. (2022). The automated Laplacean demon: how ML challenges our views on prediction and explanation. *Minds & Machines*, 32, pp. 159–183.

Volkman, R. and Gabriels, K. (2023). AI moral enhancement: upgrading the socio-technical system of moral engagement. *Science and Engineering Ethics*, 29(2), p. 11. <https://doi.org/10.1007/s11948-023-00428-2>.

Whittlestone, J., Nyrup, R., Alexandrova, A. and Cave, S. (2019). The role and limits of principles in AI ethics: towards a focus on tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. pp. 195–200. [↑](#endnote-ref-3)