HAL open science

Artificial Moral Advisors: enhancing human ethical decision-making

Marco Tassella, Rémy Chaput, Mathieu Guillermin

To cite this version:

Marco Tassella, Rémy Chaput, Mathieu Guillermin. Artificial Moral Advisors: enhancing human ethical decision-making. 2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS), IEEE, May 2023, West Lafayette, United States. pp.1-5, 10.1109/ETHICS57328.2023.10155026 . hal-04127849

HAL Id: hal-04127849

https://hal.science/hal-04127849v1

Submitted on 14 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Artificial Moral Advisors: enhancing human ethical decision-making

Marco Tassella
LUMSA University
Lyon Catholic University
Rome, Italy
m.tassella@lumsa.it

Rémy Chaput
Lyon, France
0000-0002-2233-7566

Mathieu Guillermin
CONFLUENCE: Sciences et Humanités research unit (EA 1598)
Lyon Catholic University
Lyon, France
mguillermin@univ-catholyon.fr

***Abstract*—This short paper focuses on understanding moral dilemmas, Artificial Moral Advisors, and their possible roles in ethical decision-making. After a brief analysis of the philosophical debate around dilemmas, we propose three different classes of dilemmas. We then discuss how AI-based advisors could be used to enhance human ethical decision-making, with a particular focus on three possible AI skills (identifying, presenting and settling dilemmas), as well as on their role as ethical experts. The resulting proposal opens up to new possible uses of AI moral advisors, and to the help they might offer in difficult decisions.**

***Index Terms*—Moral Dilemmas, Artificial Moral Advisors, AI Moral Enhancement, Reinforcement Learning**

I. Introduction

The world we live in is increasingly complex, and making *informed* decisions in everyday life is becoming progressively more difficult. Ethical considerations can be found in several, apparently innocent decisions, such as shopping or consuming energy [1]. With the emergence of Artificial Intelligence (AI), many started to believe that the automation of some of these decisions could help us lighten the burden of choosing. AI systems must take into account the ethical considerations that imbue these numerous choices. To do so, several approaches have been proposed: see, e.g., Dignum's Ethics By Design [2], or the Machine Ethics field [3]. However, we are currently unable to create an AI system that autonomously finds "the best decision" in every case: for example, in some particular instances — commonly called "dilemmas" — the *solution* seems to require a different approach; for this reasons, pure automation is not a viable option.

As we know, "settling" dilemmas is a difficult task even for us humans. In this paper, we suggest some ways in which AI systems could be used as tools to enhance our moral decision-making ability, at least by providing us with additional information and computing power. Although these systems will not be able to remove completely the problems arising with dilemmas, they can alleviate them, much like wearing a pair of glasses can improve our eyesight. All of this could be achieved by making humans and AI systems cooperate, in a virtuous cycle of mutual feedback.

This work was carried out within the framework of the NHNAI (nhnai.org) project, supported by the CONFLUENCE: Sciences et Humanités research unit (EA 1598) of the Catholic University of Lyon.

978-1-6654-5713-2/23/$31.00 ©2023 IEEE

II. Moral Dilemmas

Generally speaking, when we say that someone is "facing a dilemma", we put emphasis on how difficult it is for them to make a choice that they *know* they will not regret. From a more rigorous standpoint, we could understand dilemmas as difficult choices between two or more, equally desirable (or undesirable), mutually exclusive alternatives. In particular, we could say that an agent encounters a dilemma when all the possible alternatives of a choice are considered to be equally (comparably) satisfying, or equally (comparably) problematic.

Besides, most dilemmas "emerge" because of their moral implications: more often than not, dilemmas involve "a conflict between moral requirements" [4]. The emergence of a dilemma has many possible causes and, correspondingly, many kinds of dilemmas can arise. In moral philosophy, the amount of literature and debate on the subject is massive. The discussion goes so far as to challenge, or even entirely deny, the very possibility of genuine moral dilemmas. In this section, however, we suggest we could identify at least three fundamental causes for the emergence of dilemmas. Consequently, we could sketch out three corresponding kinds of "dilemmas" — that we call "Ontological", "Ethical", and "Epistemic" — based on their formal structure, and on the kind of challenges they offer. In the next subsections, we detail their scope, and the causes for their emergence.

A. (Onto)logical Dilemmas

In general, the core of dilemmas is their "insolubility". In particular, there is a kind of dilemma whose "strength" is precisely grounded in its logical structure: "(Onto)logical Dilemmas", which we could also call "Symmetrical Dilemmas". The very possibility of this kind of dilemma is among the most debated topics in analytical moral philosophy.

The debate is broadly structured as follows: on the one hand, some say, if real "dilemmas" do exist, then ontological dilemmas are their strongest instance. On the other hand, others suggest that the formal conditions of symmetrical dilemmas are not only counter-intuitive, but also straight-up invalid. Let us briefly explain why.

In moral dilemmas ($OA \oplus OB$, where O means "ought to do"), agents are morally required (ought) to do A and morally required (ought) to do B; however, they cannot successfully

achieve A without giving up on B, and vice versa. Regardless of the choice the agents will make, it seems that they will be doomed to do something that is morally wrong, or fail to do something that they were morally required to do. Symmetrical dilemmas bring this issue even further: here, the agent *ought* to do A and *ought* to do B, and both for the same reasons.

Here is a famous instance of a symmetrical dilemma, paraphrased from William Styron's novel *Sophie's Choice* [5].

Example 1. *Sophie is a Polish woman, living in the Auschwitz concentration camp with her two beloved children. One day, a Nazi guard informs her that one of her two children is to be executed, and that Sophie must choose which one. If she does not choose, warns the guard, both children will be killed.*

This choice is symmetrical, because Sophie's options are not actual alternatives. First and foremost, she *knows* she wants to ground her decision on a particular value: her unconditioned, motherly love for her children. Rationally speaking, simply making *any* choice is going to reduce the damage to its minimum, causing at most one of her children to die. Yet, ethically speaking, she has no way to solve (or "settle") the issue without doing the wrong thing. Regardless of the choice she will make, she will fail to do the right thing *for the same reasons*. In such instances, where no ethical choice can be made, there might be no real solution to the dilemma [6].

For this reason, some have proposed that, in cases like Sophie's, there is a substantial inconsistency in the dilemma's formal conditions: such dilemmas are, in reality, nonsensical. In more rigorous terms, logical inconsistency is a situation in which a statement (or set of statements) contradicts itself. For example, the statement "I always lie" is logically inconsistent, because *if it is true* then *it is false*, and vice versa. Moral dilemmas are logically inconsistent if the formal conditions of their structure cause a rational short circuit.

Intuitively, in fact, if one ought to do A *and* ought to do B, but cannot do A without forgoing B (and vice versa), it seems that A and B are both required *and* forbidden. Generally speaking, an action A can either be obligatory (desirability value of $A = 1$), permissible ($0 < A \leq 1$), or forbidden ($A = 0$), with the peculiar, yet disputed possibility of supererogation ($A > 1*$). In a "genuine" moral dilemma ($OA \oplus OB$), say the adversaries of dilemmas, the choices A and B would both have desirability values 0 *and* 1 at the same time. Nevertheless, if you cannot physically or logically do both (as one option forbids the other), then this kind of dilemma is to be considered "logically inconsistent": it is essentially incompatible with the *principle of deontic consistency* (PC) [4], whose logic is paramount.

B. Ethical Dilemmas

Let us now proceed with the second kind: sometimes, we are compelled to take a stance about irreconcilable ethical frameworks; in many cases, be they ordinary or extraordinary, this challenge can prove to be bigger than expected.

Since the dawn of moral theories, many approaches to the good and the right have been proposed. When facing a moral choice, there is often a clash between alternative values and alternative ethical frameworks. That is precisely when Ethical Dilemmas arise. Let us consider a famous example:

Example 2. *Anne is a businesswoman on her way to an important meeting. Along the way, she observes an assault taking place in an alley. An inner struggle ensues between her conscience, to stop and call for help, and her career ambitions, which tell her she cannot miss this meeting. She has to make an effort of will to overcome the temptation to go on.*

Cases such as "the Businesswoman" [7] show how crucial the choice of an ethical framework might be: is human dignity a value *per se*? Are values to be ordered hierarchically? If so, which priority order should we follow? Many answers are possible, depending on the set of values (or ethical frameworks) we wish to be guided by. Moreover, as for Anne's case, not all dilemmas are bounded to the violation of a moral rule, and yet they might subjectively look like torn choices. When we strive to "do the right thing", we are simply *forcing* one set of competing reasons to prevail over the others.

Sure enough, at least from an ethical standpoint, not every set of values is to be considered equally grounded in rational morality (if such a thing even *exists*), but at least some of them are. As an example, we could bring forth some traditional frameworks, like Kantian deontic ethics [8] and consequentialism – in its classical Aristotelian form, or in the utilitarian variant suggested by philosopher Jeremy Bentham [9]. In some fields, such as bioethics, the choice between such ethical frameworks is often considered a matter of harsh debate. However, when agents face a daily decision such as Anne's, the choice of a hierarchy is ultimately *up to them*.

C. Epistemic Dilemmas

The quandary of choosing sufficient reasons to act (i.e., hierarchizing irreconcilable values), however, is not the only obstacle in settling a moral dilemma. The third reason for the emergence of dilemmas is somehow even more disturbing: there is an essential dimension of human decision-making that further hinders our ability to settle dilemmas *for good*. It is what we call "epistemic failure": an essential misinterpretation of the broader context and the consequences of our choices, due to our brain's inability to continually present all relevant information to our conscious mind.

Human consciousness is an amazing and mysterious phenomenon. Yet, the human ability to fully understand broad contexts is rather limited. Processing information is a demanding task, and our brains seem to be only equipped to handle a limited amount of data at once. Studies in neuroscience, psychology, and education have shown the finite capacity of our working memory [10], and the effects of multitasking on cognitive processing [11]. Now, if moral judgment (e.g., facing dilemmas) is a kind of information-processing activity [12], then such cognitive limits must also apply to our moral life.

There are two main ways in which our psychological "boundedness" might hinder our decision-making: first, our

limited ability to consider the whole context and all its pertinent subtleties. Second, our inability to meaningfully predict possible consequences of our actions. Both of these issues give rise to what philosophers call "epistemic failures": in such a short period of time, our brain fails to manifest *and* connect all relevant thoughts. This is the reason we often need "more time to think", or to ask trusted people for advice.

Generally speaking, human information-processing is a slow and difficult task. Our inability to control the emergence of a certain consideration is just one of the many hindrances to a clear and rational process of reflection. The more our values are challenged – the "harder" the dilemma – and the more time we will require to make a choice. Our knowledge and comprehension of the context is usually incomplete, often lacking some crucial information, including what otherwise might have seemed adequate, perfectly reasonable predictions (cf. "moral slips" [13]). Here is the unromantic fact: a dilemma is not just a "moral debate", or a "philosophical problem". When discussing dilemmas and their solution, our psychological and physical limitations are an integrating part of the issue.

III. Artificial Moral Advisors

Such extensive premises about dilemmas are crucial, if we wish to understand how we could formalize them for our multidisciplinary proposal. Let us now cover the core questions: is there any way machines could assist in our moral endeavours? How could such machines work, and how could they improve our moral lives? Specifically, in our case: could dilemmas (even symmetrical ones) be "solved" by mere additional information and a stronger computing power? Could a machine help us to identify and solve dilemmas? These, and more philosophical and technical issues, have opened a broad debate in ethics, bioethics, and computer science.

In the last decade or so, moral philosophers and moral psychologists have deepened their work on artificial moral enhancement. Some have proposed that, in our ever-globalizing world, moral enhancement is progressively becoming a crucial and urgent matter. However, societal re-education, as effective as it may be, might not be a viable practice, at least not in such a short period of time. The alternative – moral medicine and moral bioenhancement – looks like a dangerous perspective, and a debatable solution [14].

All of this has led moral philosophers to inquire about the possibility of a totally reversible, non-invasive way of helping people in the task of becoming better citizens and more responsible agents. One option has been found in Moral AI and Artificial Moral Agents. Although the projects are still in an early development stage, most authors [15]–[17] agree on the possibility of integrating education with the use of a particular kind of AI-based software capable of helping humans to make better, more informed choices, while at the same time developing their ability to reason morally. The idea is to implement informing moral agents ("Artificial Moral Advisors") capable of broadening our comprehension of the context and the possible consequences of our decisions, including, but not limited to, "hard cases" such as dilemmas.

In the next subsections, we detail 3 different roles that AI advisors could take, in light of the dilemmas analysis proposed in Section II, and 4 skills they could offer, in order to enhance our ethical decision-making. These abilities will be more relevant to one or another type of dilemma; yet, an AI advisor may combine several roles, and nothing prevents a human user from taking advantage from a specific ability for a different type of dilemma. In order to illustrate how such AI advisors can be built, we exemplify these abilities on a use-case of energy distribution within a Smart Grid, and present early developments using Multi-Objective Reinforcement Learning (MORL).

Although RL and MORL are certainly not the only way to implement moral advisors, we find them a suitable technique, as RL deals with comparing actions in situations. MORL is a branch of RL that considers several objectives – as we naturally do when dealing with dilemmas, most of which are by definition "a conflict between values" (here assimilated to objectives to comply with).

A. Moral Observer

One of the Moral Observer's main abilities is to *identify dilemmas*, helping us in the task of recognizing that "there is a dilemma" in the first place. For example, when we want to turn on the washing machine, we (usually) do not know the current power grid's status. If the grid is, in fact, strained and under a high load, this might require turning on other power plants, such as coal-based ones, in order to prevent a power failure. This lack of knowledge hides a conflict between our comfort (powering up our household appliances when we want to) and ecology (avoiding using pollutant energy sources). An advisor might be linked to the energy provider to access the power grid's current status, and thus warn us of the current tension. More generally, in any situation, the system could check for unforeseen consequences. Then, if any of them go against the human users' (preset) values, give a warning. Some situations will then reveal to be dilemmas between moral values underlying the users' intended consequences, and the actual unexpected consequence(s).

Formally, a MORL algorithm could learn the *interest*, i.e., the gained advantage of taking an action, w.r.t. each of the moral values, in each situation, through a *Q-function*. It is defined as $Q : \mathcal{S} \times \mathcal{A} \times \mathcal{M} \rightarrow \mathbb{R}$, where $\mathcal{S} = \{s_1, s_2, \cdots\}$ is the set of all states, $\mathcal{A} = \{a_1, a_2, \cdots\}$ is the set of all actions, and $\mathcal{M} = \{m_1, m_2, \cdots\}$ is the set of moral values. In other words, $Q(s, a, m) \in \mathbb{R}$ is the interest of taking action a in state s w.r.t. moral value m. The higher the interest, the better it is to take this action, according to this moral value. This Q function is akin to the well-known Q-Learning [18], but extended to multiple moral values.

To determine whether a situation s is a dilemma, we propose to leverage the actions' interests, and to introduce the human's eye into the definition of a dilemma. First, we define *theoretical interests* Q^{th}, which correspond to the interests an action would have, if this action had always received the best (maximum) possible reward, i.e., if its impact was perfectly

good. In other words, they represent the maximal attainable interests for an action in a state. Theoretical interests provide a reference or *comparison* point for actions' interests, and allow human users to understand to which degree does the action satisfy the moral values.

The Qth are iteratively learned by adapting the traditional Bellman equation [19] to the multi-objective setting, and by using the (theoretically) maximum possible reward $\hat{r}$ instead of the actually received reward r_t. Next, we introduce the *ethical thresholds*, which represent the human user's expectations.

Definition 1 (Ethical thresholds)**.** *An ethical threshold is a vector $\zeta \in \mathbb{Z} = [0,1]^{|\mathcal{M}|}$, where $|\mathcal{M}|$ is the number of moral values. Each component ζ_m, $\forall m \in \mathcal{M}$, can be seen as a threshold between 0% and 100%, where 0% means that the user accepts any action, and 100% means the user accepts only an action with a perfect interest w.r.t. the moral value m.*

Comparing ethical thresholds and actions' interests allow us to determine whether an action is *acceptable*. For example, $\zeta = [0.8, 0.75]$ means that the user accepts an action that satisfies at least 80% of the first moral value, and 75% of the second moral value.

Definition 2 (Acceptable action)**.** *An action a in a state s is deemed acceptable with respect to ethical thresholds ζ, if the action's interests compared to its theoretical interests are higher than the thresholds, on each dimension. Formally, $acceptable(s, a, \zeta) \iff \forall m \in \mathcal{M}\ \frac{Q(s,a,m)}{Q^{th}(s,a,m)} \geq \zeta_m$.*

For example, an action a with $Q(s, a) = [8.5, 8]$ and $Q^{th}(s, a) = [10, 10]$ would be acceptable w.r.t. ζ, as $\frac{8.5}{10} = 0.85 > 0.8$ and $\frac{8}{10} = 0.80 > 0.75$. We can now define what should be recognized as a dilemma by our advisor system.

Definition 3 (Dilemma)**.** *A situation s is said to be a dilemma, with respect to ethical thresholds ζ, if, for all possible actions $a \in \mathcal{A}$, none of them is acceptable with respect to ζ. Formally, $dilemma(s, \zeta) \iff \nexists a \in \mathcal{A}$ s.t. $acceptable(s, a, \zeta)$.*

Remark. *As ethical thresholds are defined by human users, the same situation s can be deemed a dilemma by a first user, using thresholds ζ, and not a dilemma by another user, using thresholds ζ'. This definition of a dilemma effectively places the user back in the loop, which is crucial for ethics.*

B. Moral Organizer

The goal of the Moral Organizer is to *present a dilemma*, by first collecting available information, in 3 categories: those that we know and have in mind; those that we know but do not currently have in mind; and those that we do not or could not know, such as alternative or unclear options. All information should then be presented in an easily readable manner, e.g., through interfaces that display details about the dilemma, explaining why the dilemma occurred. Here, the User Experience (UX) is crucial, since users need to understand the dilemma, and what they can do to settle it. Additionally, to improve this UX, we can leverage the Explainable AI (XAI) [20] and Explainable RL (XRL) [21] fields.

In a practical setting, once a dilemma has been identified by a MORL algorithm, we propose that the user interface describes the current state s. This state can be either discrete, e.g., cells of a maze, in which case it should be clearly identified, or continuous, i.e., represented by a set of observations, which can be used to describe the state. For example, in the Smart Grid use-case, some notable parts of the state are: the current hour; the grid's current power load; the average consumption by other participants; etc. These data can be explicitly presented to the user, to highlight that there is not enough energy – that is, what triggered the dilemma.

Additionally, the interface should present the available actions. Actions can be described in terms of their interests, but also in terms of their impact on the world: what is the action really doing, what are its (potential) consequences? Thus, the interface depends on the actual application domain and use-case, particularly on the definition of states and actions. As such, it would also benefit from (domain) expert knowledge.

C. Moral Expert

Finally, the Moral Expert is associated to 2 abilities.

1) Settling a dilemma: Even once the actual dilemma is correctly presented to the user, its resolution might still be troublesome. As we highlighted in section II, the very possibility of "solving" a *genuine* moral dilemma is controversial, because it either grounds upon a (debatable) hierarchy of values, or ends in the admission of an impossible ethical choice. Generally speaking, however, both philosophy and the common sense treat dilemmas as a fact of life.

Artificial Moral Experts may assist human users by suggesting morally-charged options in a given situation. This requires the machine to be imbued with "ethical knowledge", and will particularly rely on the *Machine Ethics* field. Using knowledge about moral values and possible alternatives, AI advisors may compute the expected consequences of each decision.

When users are facing a dilemma, they must make a choice; their goal is to make a decision with the consequences that best align with their value preferences. Advisor systems could propose a range of tools to help users in their decision-making process. These tools offer different compromises between the technical difficulty of implementation, and the burden that is placed upon the human users' shoulders.

The first and simplest one, at least from a technical perspective, could simply be to ask the users to choose, among a set of proposed alternatives, the action they prefer. Proposed alternatives are derived from the action set $\mathcal{A}$, optionally filtered to remove uninteresting actions, e.g., actions whose interests are completely dominated by another action, so as to simplify the choice for users. From the user perspective, this might be difficult, as the burden mostly lies on the human's (ethical) decision-making skills. This solution relies heavily on the system's ability to *present dilemmas*, so that a user can make an informed choice – and possibly learn from it.

A second method could be to preemptively ask the users for their preferences (e.g., a lexicographic order, or a logic formula), and use them to automatically identify the most

suitable action. This method is slightly more difficult to implement, for it assumes that the users are able to express such preferences clearly (which is not guaranteed), and that the engine is able to evaluate them. In contrast, this method reduces the user's burden: they will only need to specify their preferences once, instead of selecting an action each time. However, if one considers preferences to be situational, then defining them *a priori* for most situations might be already a daunting task.

Finally, the advisor systems could go as far as to automatically derive user preferences from their data, ranging from *moral questionnaires* up to constantly observing the users' everyday life, as in Etzioni's *ethics bots* [22]. However, such systems may not be achievable yet, as they would require to accurately "understand" the users' everyday choices and their underlying preferences. One might also object that such systems would be (too) intrusive, as they require collecting extensive corpuses of data. The obvious advantage of this system would be its greater degree of autonomy, and its low-maintenance need for feedback.

2) Reflecting on our ethical preferences: Additionally, advisor systems can help users to reflect on their preferences *before* and *after* said users have settled a dilemma. Surely, not all preferences can be considered equal, as some of them are deemed unacceptable by society: for example, we might frown upon users considering that they are satisfied with a threshold of 0% with reference to the *ecology* value.

It might also happen that, because of an *epistemic failure*, users do not understand the consequences of their choices. For example, users might say that, in order to support the *ecology* moral value, they prefer taking the train rather than the plane, as long as the travel distance is less than 300 km. However, long plane trips might have a bigger impact than shorter ones: computing the real impact of trips, and comparing the consequences of replacing long and short trips (e.g. by boat) might help the users to *improve* their preferences.

Advisor systems could also work retroactively, after a dilemma is settled, to show counterfactuals, i.e., what would have happened if another action had been selected. At the very least, actions' interests $Q(s, a, m)$ can be used as counterfactuals, e.g., by comparing the chosen action's interests with those of a different action. Going further, model-based reinforcement learning [23] could be leveraged as some sort of simulator, to allow users to visualize and experiment the consequences of their preferences. This way, users could witness the impact of changing their priorities, tweaking the system's thresholds, or simply selecting different actions. This could lead them to discover new, better suited preferences for their ethical values.

IV. Conclusion and Remaining Questions

In this paper, we have first introduced moral dilemmas, and proposed 3 distinct classes based on their causes: *Ontological dilemmas*, *Ethical dilemmas*, and *Epistemic dilemmas*. We have then proposed several ways in which AI-based moral advisors could be used to enhance human ethical decision-making skills, through cooperation between the advisor and the human user. In particular, we have defined 3 roles (*Moral Observer*, *Moral Expert*, *Moral Organizer*) and 4 abilities (identifying, presenting, helping to settle, and helping to reflect on preferences) that AI advisors could offer to enhance human decision-making. In order to implement each one of these abilities, several options are possible; it is not yet clear which one would be "preferable", in terms of usability and ethical adequacy. Indeed, many technical and philosophical challenges are still open, and further testing will be needed.

To this end, moral philosophy might be of great help in designing our future interactions with AI systems. We believe that AI advisors might become a fundamental tool, if only to a certain degree, towards humanity's forthcoming moral growth.

References

[1] J. C. Dernbach and D. A. Brown, "The ethical responsibility to reduce energy consumption," *Hofstra L. Rev.*, vol. 37, p. 985, 2008.

[2] V. Dignum, *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature, 2019.

[3] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein, "Implementations in machine ethics: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–38, 2020.

[4] T. McConnell, "Moral Dilemmas," in *The Stanford Encyclopedia of Philosophy*, Fall 2022 ed., E. N. Zalta and U. Nodelman, Eds. Metaphysics Research Lab, Stanford University, 2022.

[5] W. Styron, "Sophie's choice.[1979]," *New York: Vintage*, 1992.

[6] M. J. Zimmerman, *The concept of moral obligation*. Cambridge University Press, 1996.

[7] R. Kane, "Responsibility, luck, and chance: Reflections on free will and indeterminism," *The Journal of Philosophy*, vol. 96, no. 5, pp. 217–240, 1999.

[8] I. Kant and J. B. Schneewind, *Groundwork for the Metaphysics of Morals*. Yale University Press, 2002.

[9] J. Bentham, *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press, 1996.

[10] A. Baddeley, "Working memory," *Science*, vol. 255, no. 5044, pp. 556–559, 1992.

[11] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[12] S. Guglielmo, "Moral judgment as information processing: an integrative review," *Frontiers in psychology*, vol. 6, p. 1637, 2015.

[13] F. Rudy-Hiller, "Give People a Break: Slips and Moral Responsibility," *The Philosophical Quarterly*, vol. 69, no. 277, pp. 721–740, 05 2019.

[14] I. Persson and J. Savulescu, *Unfit for the future: The need for moral enhancement*. OUP Oxford, 2012.

[15] M. Klincewicz, "Artificial intelligence as a means to moral enhancement," *Studies in Logic, Grammar and Rhetoric*, vol. 48, no. 1 (61), 2016.

[16] F. Lara, "Why a virtual assistant for moral enhancement when we could have a socrates?" *Science and engineering ethics*, vol. 27, no. 4, pp. 1–27, 2021.

[17] J. Savulescu and H. Maslen, "Moral enhancement and artificial intelligence: moral ai?" in *Beyond artificial intelligence*. Springer, 2015, pp. 79–95.

[18] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[19] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[20] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.

[21] S. Milani, N. Topin, M. Veloso, and F. Fang, "A survey of explainable reinforcement learning," *arXiv preprint arXiv:2202.08434*, 2022.

[22] A. Etzioni and O. Etzioni, "Incorporating ethics into artificial intelligence," *The Journal of Ethics*, vol. 21, no. 4, pp. 403–418, 2017.

[23] T. M. Moerland, J. Broekens, A. Plaat, C. M. Jonker *et al.*, "Model-based reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 16, no. 1, pp. 1–118, 2023.