# Why Algoenhancement Won’t Save the World: Raising three Challenges for Moral Enhancement in the Context of Major Moral Problems

Buczek, Felix

# Why *Algo*enhancement Won't Save the World

## Raising three Challenges for Moral Enhancement in the Context of Major Moral Problems

FELIX BUCZEK[*]

MASTER THESIS

'Philosophical Perspectives on Politics and the Economy'

Leiden University
s3025675@vuw.leidenuniv.nl

8th September 2022

### Abstract

*A central tenet of the standard account of moral enhancement qua algorithmic technology is that it has the potential to solve the mega-problems of our time, such as global poverty or the climate crisis. Thereby, it is simply assumed that the enhanced moral competence of individual agents will directly translate into solutions to our major moral problems. This paper sheds light on this key assumption and argues for a more sophisticated outlook on the potential effects of algorithmic moral enhancement. In particular, it is shown that our major moral problems are essentially political problems which are characterised by various kinds of dilemmas. The author shows that due to this peculiar nature of these problems, three distinct challenges arise when it comes to translating moral competence into political solutions. These challenges will have to be met by future proposals of algorithmic moral enhancement.*

[*] Supervised by Thomas Wells.

# Contents

## Introduction

In this paper, I raise three challenges for the standard account of moral enhancement *qua* algorithmic technology.[1] Advocates share the prevalent contention that the world's biggest problems, like the climate crisis or world poverty, could be solved if only people were more moral. Hence, they propose applications of algorithmic technology in order to create better moral agents. Thereby, it is usually taken for granted that better moral agents will produce less big-scale moral problems and thus 'save the world'. However, this inference is too quick, as I will demonstrate in this paper. Specifically, I draw attention to three distinct challenges that emerge when translating individual moral competence into reforms at the large-scale socio-political level.

In *Unfit for the Future: The Need for Moral Enhancement*, an influential contribution to the debate, Persson and Savulescu assert the dramatic claim that some kind of moral enhancement intervention will be necessary in order to prevent the extinction of humanity. Humanity suffers from an inherent limitation of moral capacities, preventing ethical behaviour and favouring destructive outcomes, so the argument goes. Indeed, the belief that the worlds worst problems could be solved by more competent moral individuals is a staple of the debate on moral enhancement.

Until recently, proposals were centred around the highly controversial genetical or neurobiological manipulation of the human constitution. However, recent advances in algorithmic technology bear the potential of pulling the debate into more mainstream territory. After all, most of the contentious byproducts of conventional moral enhancement can be bypassed by *algo*enhancement, securing full autonomy of the enhanced subject. Hence, it is time to reflect anew on the potential of moral enhancement.

As I will show, proponents of *algo*enhancement focus on targeting the moral capacities of individual human agents. But what is the upshot of making individuals more moral? It seems that advocates implicitly share the belief that moral competence leads directly to actual reform. That by improving the moral capacities of individual agents, major moral problems can be solved. In this paper I will show why this is a problematic premise and requires in-depth analysis. That is, I will present reasons for why *algo*enhancement won't 'save the world'.

[1] Henceforth '*algo*enhancement', abbreviated '*AE*'.

The structure of this paper is as follows: The aim of the first section is to present and contextualise the main claim of this paper. To ensure an informed framework for the subsequent discussion, I will first specify the general topic of moral enhancement and introduce key issues and figures of the debate. Subsequently, I will formulate and motivate my claim. Moreover, I will report who has held similar or opposing views and point out the original contribution of my thesis. Finally, I will disclose preliminary assumptions.

In the second section, I will reconstruct and analyse the standard account of *algo*enhancement. Drawing on textual evidence derived from the most prominent proposals, I will point out their shared motivational rational and derive the 'blueprint' of moral enhancement *qua* algorithmic technology. In order to provide a formal analysis, I will explicate the underlying argument. Finally, I will identify the argument's weakness and demonstrate that it suffers from an informal fallacy.

In section three, I will raise three challenges for the standard account of *algo*enhancement. Building on the previous section, I will explore the consequences of the informal fallacy for the project of *algo*enhancement. I will demonstrate that it draws attention to the fact that translating individual moral competence into solutions in the context of major moral problems is not as straight-forward as advocates assume and can run into problems. Specifically, I will show that major moral problems are 'wicked' problems, embossed by 'each-we' dilemmas and the problem of 'dirty hands'.

# 1. Claim & Context

Artificial intelligence[2], widely held to be a disruptive technology, is increasingly being hailed as both a universal panacea and cause of the most pressing challenges of our time. On the one hand, there is a lively discourse about the potential to improve almost every area of organised life, such as business, health, security, communication, *et cetera*. On the other hand, potential pitfalls are critically discussed—especially with regard to ethical concerns towards, *e.g.*, privacy, autonomy, bias, and transparency (see Powers and Ganascia 2020).

Consequently, there is a lot attention on how to regulate algorithmic technology, resulting in an overwhelming amount of principles and guidelines.[3] Thereby, AI is framed as a problem, rather than a solution. This paper, however, follows a different approach and sheds light on a concern that has been underrepresented so far: the potential of artificial intelligence technology for the pursuit of a very specific kind of improvement, namely moral enhancement. Rather than investigating how we can make AI more ethical, it looks into proposals on how AI can make ourselves more ethical.

## 1.1. Moral Enhancement

One might think that the discourse on moral enhancement is very old. Cultural practices like social norms, religion, and philosophy, aimed to shape humanity for the morally better since the dawn of men. However, such a broad understanding is ahistorical. In a narrow sense, the generic concept of moral enhancement was coined by bioethical debates of the 1990s (see Fenner 2019, p. 19). In contrast to traditional methods of moral education, moral enhancement generally builds on scientifically informed and technologically mediated interventions. Unlike medical treatment, these interventions aim programmatically for establishing moral virtues beyond 'ordinary' capacities.

Breakthroughs in neuroscience, pharmacology, and genetics inspired a debate on potential improvements of morally relevant human properties. Proposals include diverse approaches of fostering pro-social behaviour and elevating dispositions such as trust, empathy, cooperation, fairness, or self-control. Research suggests that relevant neurohormones such as

[2] Henceforth abbreviated 'AI'.

[3] See Turner 2019, pp. 207–62 for an comparative account on AI regulations.

serotonin, oxytocin, or testosterone can be effectively steered by the means of both invasive or noninvasive brain stimulation, exogenous administration of drugs, or genetic intervention (see Earp, Douglas and Savulescu 2017, pp. 166–7). Today, *bio*enhancement has become a major issue in applied ethics.

However, what constitutes moral enhancement beyond this descriptive definition is essentially contested (see Raus et al. 2014, p. 263). This is because moral enhancement is a normative concept. It inherently presumes a positive evaluation of the proposed intervention. Who would not want to be more moral, after all? Of course, what counts as moral is probably the most prominent open question in philosophy. Hence, any substantial proposal of *bio*enhancement, *i.e.*, any proposal that promotes a specific moral theory, is bound to be controversial. Is a heightened capacity for empathy unconditionally good? Arguably, it might do more harm than good in certain situations.

But this is not the only reason why *bio*enhancement is disputed. Critics argue that *bio*enhancement disregards core moral values like privacy, autonomy, or the conservation of humanity against profound unnatural alterations (see Douglas 2014). This led to debates about the moral justification, permissibility, and desirability of *bio*enhancement—whether it should even be compulsory or rather prohibited (see Lavazza and Reichlin 2019, p. 3). Subjects of *bio*enhancement are consequently suspected to shortcut the intellectual demands of morality. It is argued that, being imposed to doing the right thing, agents lack the freedom to do wrong. Authentic moral behaviour is thereby prevented.

On the other hand, proponents of *bio*enhancement emphasise the urgent need for moral intervention. In fact, it is commonly held that 'some of the world's most important problems' (Douglas 2014, p. 469) can be attributed to moral deficits. And it is a central tenet of *bio*enhancement, that traditional methods of moral education are not sufficient to mend these critical deficits (see Powell and Buchanan 2016, p. 239), thus further stressing the urgent need for moral enhancement. Prominent scholars even argue that in the face of a dramatic mismatch of technological power and moral capacities—potentially giving rise to mass destruction and environmental collapse—moral enhancement is a necessary tool to prevent the ultimate downfall of mankind (see Persson and Savulescu 2012, *passim*).

In the mean time, the debate has fallen into a stalemate between proponents and critics, confused by second-order disagreements and dominated by metaethical disputes (see Paulo and Bublitz 2019a, p. 96). Recently, however, a technological breakthrough once again inspired a new project of moral enhancement. While algorithmic technologies in general have been the subject of various philosophical studies for half a century (see Fetzer 2004, p. 119), they have only recently been discussed in the context of moral enhancement. Meanwhile, technological advances have unlocked the evocative power of machine learning and ambient computing, prompting afresh reflection on potential applications.

The philosophical debate on moral enhancement with the means of algorithmic technology ties in seamlessly with the preceding discussion of *bio*enhancement. Typically, it is presented as a solution to the above mentioned shortcomings of *bio*enhancement (see *e.g.* Lara and Deckers 2020, p. 276 or Savulescu and Maslen 2015, p. 80). It is generally argued that artificial intelligence (AI) is providing the tool for crafting better moral agents without entailing contentious first-order theoretical commitments or causing undesirable side-effects for the enhanced subject. In particular, it is argued that this can be achieved by the means of an AI-based ethical decision assistant.

The current outlook on *algo*enhancement is dominated by Savulescu and Maslen. They suggest phenotypical features of an AI advisor. Thereby, the general idea is this: Human beings are essentially imperfect moral agents. Our genetic heritage endows us with psychological traits that are not ideally suited to our contemporary lifeforms in global social communities, leading to biases and fallacies in decision-making and ultimately preventing moral action. To overcome these inherent moral limitations, AI could be utilised to prompt the user to actively engage in moral decision-making, to provide empirical insight and normative feedback to morally relevant judgements as well as organising and fostering motivation for moral action (see ibid.).

Although the exact design of *algo*enhancement is a matter of controversy, I show that all relevant proposals strikingly commit to the same underlying theoretical rationale. According to this implicit framework, humanity suffers from psychological deficits that prevent moral action and encourage wrongdoing. As a result, severe suffering is caused both actively—*e.g.* through terrorism, genocide, or violent oppression—and passively through omission to assist those in dire need. Consequently, the central targets of *algo*enhancement are the limited moral capacities of the individual agent with the ultimate goal of solving the major moral problems of our time. It is this implicit rationale—rather than a specific proposal of design—that I challenge in this paper.

## 1.2. Main Claim

I draw attention to a hitherto overlooked problem. The standard account of *algo*enhancement simply presumes without independent argument that the effects of *algo*enhancement translate from the level of individual moral competence to the level of large-scale socio-political organisation. Indeed, this problematic inference is a staple in the debate on moral enhancement in general. Crutchfield provides a book-length argument for compulsory *bio*enhancement in *Moral Enhancement and the Public Good*—without explaining how moral enhancement will practically serve the public good. The author simply takes for granted without independent argument that *bio*enhancement will prevent ultimate harm and extinction, once deployed.

I claim that this widely shared presumption is problematic and propose that the current debate would profit from a more detailed analysis of major moral problems. In its current formulation, the standard of *algo*enhancement suffers from an informal fallacy which obscures a pressing problem: major moral problems do not consist exclusively of problems that can be traced back to the moral incompetence of individual agents. Moreover, morally competent individuals are not necessarily more capable of solving major moral problems than regular individuals.

Hence, my point is carefully aimed. I do not set out to evaluate the potential of algorithmic technology to enhance moral capacities. Instead, my argument is targeted specifically at the motivational narrative which underlies the current debate. In line with other scholars, I show that the standard account of *algo*enhancement misrepresents the nature of major moral problems by suggesting that an aggregation of superior moral individuals will solve them. Hence, my main contention is that the current project of *algo*enhancement does not convincingly argue that it can achieve an important part of what it claims: the solution of large-scale moral problems.

## 1.3. Existing Critical Accounts

Notably, there are a few scholars who hold a similar view. In general, my argument is reminiscent of Harris' critical perspective on moral *bio*-enhancement. In response to key figures like Douglas, Persson and Savulescu he draws attention to their programmatic individualism. Consider the following quote of his 2016 monograph *How to be Good: The Possibility of Moral Enhancement*:

> What we need in order to solve, or even help mitigate, global poverty is a global solution and this must be attempted at a minimum at state level, and probably at an international or global level. [...] Let's [...] think about addressing these important problems at the level of policy and indeed of government or better, at a combined governmental, truly international, level. (Harris 2016, p. 144)

Here, Harris questions the need for moral enhancement and highlights a simple and striking fact: big-scale moral problems require big-scale solutions. He sees this fact in tension with the traditional perspective of *bio*enhancement. Contra its programmatic 'excessively (one might say "obsessively") individualist view' (ibid.), he favours a politically coordinated strategy. This intuitive scepticism towards an individualistic approach to solving large-scale problems is one of the main motivations for this paper.

Paulo and Bublitz criticise the project of *bio*enhancement in a similar spirit. In their recent article 'How (not) to Argue For Moral Enhancement: Reflections on a Decade of Debate', they claim that it is 'likely insufficient to [...] solve global problems.' (Paulo and Bublitz 2019a, p. 101) They argue that proponents 'fail to diagnose the often complex causes of contemporary moral maladies' (ibid., p. 95) and are committed to an erroneous impression of *methodological individualism*, 'the view that all higher level social processes can ultimately be exhaustively explained (and, hence, remedied) at the level of the individual.' (ibid., p. 100)

Ultimately, they conclude that '[s]olving the mega-problems of today very likely requires more than transforming individual brains, it requires structural and higher-level changes. By itself, moral bioenhancement is thus insufficient for solving these problems.' (ibid., p. 95) While proponents of *bio*enhancement argue that the necessary higher-level changes 'have not been undertaken because people today are not enough concerned about harmful effects in the remote future.' (Persson and Savulescu 2015, p. 55), Paulo and Bublitz debunk this claim with empirical support and conclude that

> [. . . ] it seems much more likely that something else along the way from citizens' minds to the conference tables in Copenhagen or Paris has gone astray. It might be that political representation is not working; it might have to do with powerful stakeholders, national interests, the protection of specific industries and global power structures. (Paulo and Bublitz 2019a, p. 99)

I belief Paulo and Bublitz expose a problem that is equally pressing for the standard account of *algo*enhancement, as it is for the project of *bio*enhancement, for the algorithmic moral advisory of individual agents might be prone to the same limitations as the biomedical approach of 'transforming individual brains'. After all, both strategies aim to enhance the capacities of individual agents. However, as noted above, it is questionable whether major moral problems that are the product of structural failure can be solved on the individual level.

While Paulo and Bublitz do not further look into the issue, there are two original strands of argument that pinpoint specific structural failures that are the root of big moral problems. For one, De Araujo draws attention to the competitive political arrangement of the international system of conflicting states, rather than to the limited moral capacities of individual agents (see De Araujo 2014, p. 30). In his 2014 article 'Moral Enhancement and Political Realism', he is concerned with the levels of social organisation in the context of moral enhancement.

Informed by the international relations theory stance of political structural realism, he claims that *bio*enhancement ultimately cannot help us solving major moral problems by making individuals morally better (see ibid., p. 31). According to *political structural realism*, political conflict is rooted in the 'anarchical structure' of the international system of states, in which every nation state is concerned first and foremost with its own survival and security (see ibid., p. 35).

Consequently, political leaders would 'often feel compelled to favour security over morality, even if, all other things being considered, they would *naturally* be more inclined to trust and to cooperate with political leaders of other states.' (ibid.; his italics) He concludes that even under the influence of effective and safe moral enhancement, there won't be any sufficient change in moral reality, if it is not for substantial change at the level of social structure (see ibid., p. 38).

A different perspective is provided by Bublitz. In contrast to De Araujo, he identifies the source of major moral problems with the socio-economical system rather than the political system. In his article 'Saving the World through Sacrificing Liberties? A Critique of some Normative Arguments in Unfit for the Future', he critically observes that the moral evaluations of individuals with notable political power are often overruled by economic motives (see Bublitz 2019, pp. 32–3). He concludes that '[r]ather than individual minds, the current economic system may be unfit for the future and in need of drastic reforms.' (ibid., p. 33)

## 1.4. Original Contribution

The above introduced existing critical accounts suffer from a blind spot. Firstly, Harris does not provide a convincing case for his position. His claim is backed by a scepticism towards the human capacity for altruism in general. He argues that major moral problems cannot be solved because 'of a combination of human weakness in the form not least of weakness of will, but also because of the human weakness of not being able to drum up much sympathy for the ugly and unsavoury or for those out of sight and out of mind' (see Harris 2016, p. 144). He thereby ignores the very fact that *bio*enhancement aims to overcome precisely these inherent human psychological limitations in order to enable altruistic behaviour.

Secondly, while I think that Bublitz and Paulo do have an important point as well, they attribute proponents of *bio*enhancement too strong a claim. The project of *bio*enhancement—just as the project of *algo*enhancement for that matter—does indeed recognise the need for structural change. In fact, moral enhancement is programmatically proclaimed as a requirement for structural reform, that is, proponents argue that reform is not achievable without substantial enhancement of individuals. Indeed, Savulescu himself notes that he has 'never thought that political action is unnecessary, but [. . . ] that moral enhancement is necessary for accomplishing requisite political actions' (Persson and Savulescu 2015, p. 53).

The design of structural and higher-level organisation—be it the political system of nation states or the socio-economical system—is typically a matter of politics. And at least in liberal democratic states, individuals do have notable influence on political action by the means of political participation and representation. One might hold that enhanced individuals would make better, more far-sighted political decisions and even revolutionise their organisation when necessary—and eventually solve big moral problems. De Araujo does recognise this and grants that moral enhancement might reinforce the political will to overcome the 'anarchical' global system of nation states and to implement an overseeing world state, which ensures national security (see De Araujo 2014, p. 36).

However, as I show below, the idea that individuals are able to realise the necessary transformations provided their moral capacities are enhanced by *algo*enhancement is deceptively simplistic. Against the backdrop of the existing critical accounts, the original contribution of this paper is that I explore this implicit assumption and show that it conceals difficulties for the standard account of *algo*enhancement. Indeed, these difficulties occur beyond the context of moral enhancement. In his programmatic book-length study about the *The World's Worst Problems*, Dodds develops a familiar rationale:

> Solving [the world's worst] problems will require a fundamental change in core values [...] By this I mean that people of the world will need to mostly adopt and act on the moral assumption that [...] death or suffering of any person, once they are born, now or in the future, is weighted equally across the world. [...] Successful adoption of this moral view would mean developing a 'global identity.' (Dodds 2019, pp. 127–8)

If people would only commit to the right moral values—in this case the equally shared intrinsic value of human life as the axiological basis of an emphatic 'global identity'—major moral problems would be solved. Hence, although Dodds provides an in-depth analysis of the biggest contemporary challenges, he beliefs that a change in moral values of individuals is required to solve these issues. With this paper, I hope to resolve the common misconception that the enhancement of moral dispositions does unconditionally warrant for the solution of major moral problems.

To conclude, this paper sets out to critically evaluate the potential of *algo*enhancement and to correct the outlook on the role of moral capacities in the context of big moral problems. However, I do not see why my argument would only hold for the philosophical project of *algo*enhancement, for it does not rely on any premises committed to the context of algorithmic technology. I aim to revise a narrative that is a common misconception, not only shared by both proponents and critics of *bio*enhancement, but of moral enhancement in general.

Hence, my findings will inform future debates of moral enhancement *en gros* and prevent naïve and overly optimistic outlooks on technological solutionism of the big moral problems that define our time. No doubt, artificial intelligence is indeed a powerful technology which very well might help us overcoming our biggest moral problems. For this very reason, it is important that we resolve any conceptual confusion regarding the possible aims and targets of moral enhancement technology. Otherwise we are prone to miss out on the groundbreaking potential of this technology.

## 1.5. Preliminaries

For the sake of the argument, I take for granted the feasibility of *algo*enhancement in principle without independent argument. However, it is noteworthy that this is a presumptuous premise. Many scholars are sceptical about the project's feasibility, both on technological and conceptual grounds (see Beck 2015). Considering conceptual feasibility, the strand of critique lead by Klincewicz is particularly pressing. He draws attention to the antecedent metaethical commitments and investigates resulting limitations for the project of *algo*enhancement (see Klincewicz 2016).

In this context, it is noteworthy that proponents usually do not elaborate on what exactly they mean when they refer to algorithmic technology. Instead, the technological specifics are black-boxed by rather vague concepts like '"[p]ervasive" or "ubiquitous" computing and the more recent concept of "ambient intelligence"' (Savulescu and Maslen 2015, p. 84). Thereby, the actual technical features are only roughly outlined as, for example, 'a system that gathers information form multiple sensors and processes the functional significance of this information in "awareness" of environmental and user context.' (ibid.)

Of course, conceptual vagueness bears the risk of causing critical delusion about the actual faculties and requirements of algorithmic technology.[4] This holds especially for 'artificial intelligence', since it is an umbrella term that subsumes a variety of algorithmic technologies and is therefore notoriously confusing for the uninformed.[5] Hence, special alertness is in order considering the hazardous 'buzzword jungle' (Thamm, Gramlich and Borek 2020, p. xxii) that surrounds algorithmic technology.

[4] See Hagendorff and Wezel 2020 for a clarifying overview about common misconceptions and challenges with regard to AI.

[5] See Russell et al. *Artificial Intelligence: A Modern Approach* for an elaborate introduction and Ertel's *Introduction to Artificial Intelligence* for a more accessible overview.

However, mind that the debate about *algo*enhancement is a philosophical debate, first and foremost. As such, it is primarily concerned with philosophical questions. Consequently, for this paper I assume *algo*enhancement is both technologically and conceptually feasible. Instead, I focus on the philosophical question whether it would solve our biggest moral problems. A lot has been written about the feasibility of moral enhancement. Here, I rather aim to evaluate if the proposed interventions actually are likely to succeed.

This brings me to the last preliminary remark: To a considerable degree, the debate on moral enhancement *en gros* is theoretical speculation. Considering the current state of technological progress, this holds especially for the case of *algo*enhancement. However, the fact that this paper is based on a thought experiment rather than actual feasible technology does not deny its relevance *per se*. Indeed, a lot of philosophy is organised around thought experiments, highlighting the value and importance of this essential methodological instrument.

# 2. The Standard Account

The aim of this section is to provide a reconstruction and critical analysis of the standard account of *algo*enhancement. In order to secure an informed framework for the subsequent evaluation, I sketch a 'blueprint' of the envisioned functionality and derive the underlying motivational argument. The standard account of *algo*enhancement is especially prone to misunderstandings since it is not set out by a single proposal, but a composite of various sources from different authors. Hence, I aim cautiously for a charitable interpretation. Although I provide textual evidence for my interpretation, mind that my main concern is systematic rather than exegetic.

## 2.1. The Implicit Argument

Of course, the project of *algo*enhancement in general entails various arguments. However, this paper highlights a rationale that is usually taken for granted in the debate. Drawing on the prominent proposals of Savulescu and Maslen, Lara and Deckers, Klincewicz, Giubilini and Savulescu, and Seville and Field, I propose the following outline:

**The Standard Account of *AE* (Formalised)**

*P1* In all cases where human agents tend to behave morally defective, they are subject to their inherently limited moral capacities.

*P2* In all cases where human agents cause and perpetuate major moral problems, they tend to behave morally defective.

> *C1* *Ergo*, in all cases where human agents cause and perpetuate major moral problems, they are subject to their inherently limited moral capacities.

*P3* In all cases where human agents are guided by *algo*enhancement, they are less subject to their inherently limited moral capacities.

> *C2* *Ergo*, in all cases where human agents are less subject to their inherently limited moral capacities, they tend to cause and perpetuate less major moral problems.
>
> *C3* *Ergo*, in all cases where human agents are guided by *algo*enhancement, they tend to cause and perpetuate less major moral problems.

In the following, I will explain each step of the argument and provide textual evidence for the proposed formulation.

## 2.2. The Motivational Rationale

The fact that humanity is confronted with a variety of man-made major moral problems is a widely shared background assumption of prominent proponents of *algo*enhancement. In fact, it already served as the main motivation for earlier proposals of *bio*enhancement (see for example DeGrazia 2014, p. 362 or Persson and Savulescu 2012, p. 1). Consider the following quote for a particularly illustrative example:

> Human beings in the twenty-first century are confronted with a daunting array of moral problems, from climate change, poverty, and genocide to the prospects of nuclear war and terrorism—ethical challenges that human moral psychology, which evolved to function under very different social and technological circumstances, is arguably ill-equipped to address. (Powell and Buchanan 2016, p. 239)

Powell and Buchanan declare the first two premises as a 'key framing assumption' of proponents of *bio*enhancement (see ibid.). Indeed, they are staples of the debate on moral enhancement in general. In his 2014 article 'Moral Enhancement', Douglas discusses the reasons for and against moral enhancement, citing 'the world's most important problems' as main motivation for any moral enhancement project:

> There is clearly scope for most people to morally enhance themselves. According to every plausible moral theory, people often have bad or suboptimally good motives. Moreover, according to many plausible theories, some of the world's most important problems – such as developing-world poverty, climate change, and war – can be attributed to these moral deficits. (Douglas 2014, p. 469)

It is notable that the more recent proposals of *algo*enhancement often do not introduce these notions as explicitly as it is the case in the above quotes. Instead, they are usually assumed or merely implied by phrases such as 'the pressing challenges inherent in a globalised world' (Savulescu and Maslen 2015, p. 79) or 'the threats of a new, globalised world' (Lara and Deckers 2020, p. 275). Nevertheless, it is evident that contemporary major moral problems, such as climate change, world poverty, and genocide, play a crucial role in motivating the standard account of *algo*enhancement.

Interestingly, proponents identify the source of these major moral problems programmatically with the morally defective behaviour of individual human agents. According to this rationale, these problems exist because of active wrongdoing or omitted assistance of individual agents. Both are widespread, resulting in 'lying, corruption, racism, murder, and paedophilia' (Lara and Deckers 2020, p. 275), causing 'environmental degradation and [. . . ] harmful climate change' (Savulescu and Maslen 2015, p. 82), perpetuating world hunger—even raising the prospect of nuclear annihilation of humanity (see ibid.).

Proponents usually explain these widespread moral failures with recourse to evolutionary biology: Human psychology would be evolutionarily oriented towards living in close-knit communities and would therefore systematically favour short-term thinking and partial in-group action over long-term considerations and altruistic motives (see ibid., pp. 79–80, *locus classicus*). Once effective, this hard-wired psychological programming would develop a grave momentum when confronted with our contemporary global mass societies, resulting in collective-action problems and free-rider problems (see ibid., p. 81).

Informed by findings of cognitive science, it is furthermore held that 'we are suboptimal information processors, moral judges, and moral agents' (Giubilini and Savulescu 2018, p. 170). Human agents would be 'biologically predisposed to have limited cognition and to have a limited level of altruism' (Lara and Deckers 2020, p. 275). Individuals would be 'naturally disinclined to act to avert' (Savulescu and Maslen 2015, p. 82) major moral problems that involve collective action problems. Klincewicz simply calls this condition the 'The Moral Lag Problem' which is 'a shorthand name for all the things that cause us to be not as moral as we could or should be.' (Klincewicz 2016, p. 172)

The general idea is this: Human agents usually act on the basis of a decision-making process. Unfortunately, this decision-making process is often either compromised or completely overruled by irrational psychological dispositions, genetically inherited and grounded in our neurobiological constitution. However, it is held that moral capacities—be it impartial rational deliberation, (subconscious) moral motivation, or actualised personal virtue—provide the power to effectively correct such compromised decisions (see Savulescu and Maslen 2015, p. 81).

### 2.3. The 'Blueprint' of Algoenhancement

How exactly algorithmic technology can be utilised to neutralise these irrational psychological dispositions (*P3*) usually resembles the heart of the proposals of *algo*enhancement. It is the original claim that proponents spend the most effort on. Following the above analysis of the genesis of major moral problems, the aim of the project of *algo*enhancement is 'to improve moral cognition, motivation and behaviour' (Savulescu and Maslen 2015, p. 82) in order 'to help us to reach better decisions ourselves and, consequently, to act better' (Lara and Deckers 2020, p. 280), so that ultimately 'our judgments and actions [are] more consistent with our explicit moral goals' (Giubilini and Savulescu 2018, p. 171).

Thereby, it is argued that algorithmic technology provides the means to achieve moral enhancement 'more rapidly and successfully than traditional methods, and with fewer risks and controversies than bio-enhancement.' (Lara and Deckers 2020, p. 276) The decisive advantage of *algo*enhancement over *bio*enhancement addressed here is that it optimally grants the enhanced subject full autonomy and prevents any kind of moral paternalism. This is enabled by instances of 'weak' AI.[6]

Rather than chasing the creation of autonomous artificial moral agents with superior moral capacities, proponents are united in the belief that it is more promising to utilise the unique strengths of both technology and humanity, in order 'not [. . . ] to change our behaviour, but the ways in which we make moral decisions.' (ibid., p. 280) In the case of Giubilini and Savulescu, the ideal is a fully rational moral agent, denouncing emotions and intuitions as unreliable epistemic instruments (see Giubilini and Savulescu 2018, p. 171). To get a grasp on how this might be achieved, consider the following proposal:

> We therefore argue that the contender for serious consideration is a type of 'weak' moral AI that [. . . ] gathers, computes and updates data to assist human agents with their moral decision-making. This data will comprise information about the individual agent and his

[6] The distinction between 'weak' and 'strong' AI is a staple of the literature on artificial intelligence. Systems of 'strong' or 'general' AI are usually characterised 'by their ability to develop creatively and produce behaviours the developers could not program, design, or even imagine. These are systems that can think and reflect about their own state, so they know who they are and what they are for.' (Henning 2021, p. 35) Furthermore, it is even speculated about the potential creation of 'superintelligence', that is 'any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest' (Bostrom 2014, p. 26, quoted in Sutrop 2020, p. 56). Following this technical distinction, there are potentially two distinct approaches to *algo*enhancement. On the one hand, it is speculated that strong AI may provide the means of creating autonomous artificial moral agents (see Savulescu and Maslen 2015, p. 84). Far superior to their human counterpart, they would dictate their ideal (moral) judgements, leaving humans to passively obey (see Lara and Deckers 2020, p. 277). Yet, considering the mere fact that it is widely accepted that 'strong' AI is technologically unfeasible, proponents of *algo*enhancement agree on exploring the potential of 'weak' AI instead.

> environment, about his moral principles and values and about the common cognitive biases that affect moral decision-making. The moral AI will use this data to alert the agent to potential influences and biases, will suggest strategies for ameliorating these influences and biases, and will advise the agent of particular courses of action at his request. (Savulescu and Maslen 2015, p. 84)

This is a quote from the most dominant proposal of the debate, where Savulescu and Maslen provide the 'blueprint' of *algo*enhancement. Here, the advantage of *algo*enhancement becomes apparent: Rather than overriding faulty moral judgements on the neurobiological level, the enhanced subject is prompted to correct them autonomously by the means of active deliberation. Their proposal features four distinct ideas on how to utilise algorithmic technology for the means of moral enhancement. Since all later proposals more or less build on these ideas, it is worth giving a brief overview.

The first envisioned function is the *continuous moral environment monitor*, which acts as a 'bio-feedback facility, where the physiological, psychological and environmental data is analysed from the perspective of optimal moral functioning.' (ibid., p. 85) Thereby, the AI 'monitor[s] the agent's physiology, mental states and his environment [. . . ] to alert the agent to particular factors that tend to affect moral decision-making and behaviour.' (ibid.) Relevant factors are, *e.g.*, the amount of sleep, the time between meals, environmental effects like temperature, crowdedness, or tidiness, and physiological patterns of arousal. In addition, levels of hormones and neurotransmitters associated with judgement and behaviour are monitored. Whenever the AI registers a notable effect, the enhanced subject is alerted to reflect on its decision-making and behaviour.

The second presented function is the *continuous moral organiser*. It aims to realise the pre-existing moral motivation of the enhanced subject, assuming that individual moral goals are often failed due to a lack of information, organisation and consistency of behaviour. Examples of moral goals include donations to charity, volunteering for altruistic causes, reducing carbon footprint, or keeping promises (see ibid., p. 86). In this case, the AI would be 'aware of opportunities for the agent to meet his goals (for example new charitable organisations or events; alternative travel options), make suggestions about how best to achieve his goals, and alert him when he misses his targets.' (ibid.)

Thirdly, Savulescu and Maslen present the *situation-specific moral prompter*. This feature seeks to provide neutral guidance for agents already engaged in moral deliberation. Whenever the agent is faced with a moral dilemma, the AI raises relevant questions to challenge the agent's judgement and thus achieve a more refined decision-making process, ultimately resulting in deliberate behaviour. Thereby, the AI is informed by 'a variety of ethical considerations drawn from different accounts of what constitutes right action.' (Savulescu and Maslen 2015, p. 87) Typical outputs would cover concerns like these:

> 'what would be the consequences of your act for your self and others?', [. . . ] 'do you think you will feel shame or remorse if you go ahead with the act?', [. . . ] 'would one course of action result in more overall benefit than the other?', 'are you being influenced by any irrelevant characteristics of the two parties, such as race or gender?', 'do you think that if you have the time and capacity to help the person in need you should?' (ibid.)

The fourth and last proposed feature is the *situation-specific moral advisor*. Whenever the subject is confronted with a morally challenging situation, this system provides action-guiding output in the form of: taking into account the input of your weighted values x, in your situation y, action z is most advisable. Of course, this output depends crucially on the computation of the given moral input. The enhanced subject is requested to 'indicate which of a long list of morally significant values or principles he holds and is guided by' and to 'assign a weight (between 0 and 1) to each value.' (ibid., p. 88) Thereby, suggested values include *e.g.* benevolence, justice/fairness, legality, maximising net utility, and environmental protection. Based on this input, the AI calculates the most consistent action. In the words of the authors:

> For any given scenario, the AI would compute the extent to which the courses of action open to the agent would uphold or compromise these values (fully uphold value = 1; fully compromise value = -1), amplifying or diminishing based on the weight indicated by the agent. The AI would then use these weighed values to suggest the best course of action. (ibid.)

All four subsystems are designed to synergise. The idea is to neutralise the impact of the neurobiological constitution on the agent's decision-making through increased sensitivity to external influences, strengthened consistency of value-driven behaviour, and refined deliberation, thus ultimately improving the agent's moral capacity.[7] Provided that the source of

[7] Although all proponents agree in that regard, they do entertain different foci: Seville and Field 2011 provide an early discussion of the themes of *algo*enhancement, Giubilini

major moral problems are to a big part attributable to the limited moral capacities of individual agents (*C1*), it only seems reasonable to develop tools to overcome these limitations in order to address these problems. Furthermore, given that algorithmic technology bears the means to develop these tools (*P3*), *algo*enhancement effectively helps to overcome major moral problems (*C3*). This is the standard account of *algo*enhancement.

## 2.4. Formal Analysis

Now that I reconstructed and contextualised the standard account of *algo*enhancement, I point out why it is problematic. For this purpose, I propose a two-step analysis. Since it is an inductive argument, I examine whether it is strong and cogent. Firstly, I analyse whether its conclusions follow from its premises. Secondly, I examine whether its premises are convincing. I repeat here the previously given formalisation so that you can easily follow the subsequent analysis:

**The Standard Account of *AE* (Formalised)**

*P1* In all cases where human agents tend to behave morally defective, they are subject to their inherently limited moral capacities.

*P2* In all cases where human agents cause and perpetuate major moral problems, they tend to behave morally defective.

> *C1* *Ergo*, in all cases where human agents cause and perpetuate major moral problems, they are subject to their inherently limited moral capacities.

*P3* In all cases where human agents are guided by *algo*enhancement, they are less subject to their inherently limited moral capacities.

> *C2* *Ergo*, in all cases where human agents are less subject to their inherently limited moral capacities, they tend to cause and perpetuate less major moral problems.
>
> *C3* *Ergo*, in all cases where human agents are guided by *algo*enhancement, they tend to cause and perpetuate less major moral problems.

For the purpose of a formal analysis, it is helpful to provide a symbolised form of this argument. Drawing on basic propositional logic, I propose the following formulation, in which the logical relationships between the implicit propositions are indicated and the respective types of syllogisms are explicated in parentheses:

and Savulescu 2018 and Klincewicz 2016 focus on the elaboration of the *situation-specific moral advisor*, while Lara and Deckers 2020 and Lara 2021 promote and develop the *situation-specific moral prompter*.

**The Standard Account of *AE* (Symbolised)**

*P1* $B \implies A$

*P2* $C \implies B$

*C1* $C \implies A$ [*hypothetical syllogism*, *P1*, *P2*]

*P3* $D \implies \neg A$

*C2* $\neg A \implies \neg C$ [*transposition*, *C1*]

*C3* $D \implies \neg C$ [*hypothetical syllogism*, *P3*, *C2*]

The first two premises (*P1*, *P2*) translate the suggested causal link between the existence of major moral problems and the inherently limited moral capacities of individual agents into separate conditional claims. Note that the assumption of causality in the context of social behaviour in general and global politics in particular is clumsy, if not misleading. Social outcomes are inherently multi-causal. However, since it is an inductive argument, the described effect is gradual—rather than categorical—and abstracts *ceteris paribus* from any potentially interfering factors.

On the one hand, *P1* expresses the supposed connection between the inherently limited moral capacities of human agents and the occurrence of morally defective human behaviour. On the other hand, *P2* posits that morally defective behaviour is prone to create and perpetuate major moral problems. Based on *P1* and *P2*, the conclusion is then drawn *qua hypothetical syllogism* that human agents cause and perpetuate major moral problems mainly because their moral capacities are inherently limited (*C1*).

*Qua transposition* of this conclusion, it can be derived that if human agents were less subject to their inherently limited moral capacities, they would probably tend to cause and perpetuate less major moral problems (*C2*). While *P3* represents the central claim of the current proposals of *algo*enhancement, namely that algorithmic technology bears the means to overcome these inherent moral limitations, it is ultimately inferred *qua hypothetical syllogism* that in all cases where human agents are guided by *algo*enhancement, they tend to cause and perpetuate less major moral problems.

The above symbolisation demonstrates that the standard account of *algo*enhancement is backed by a formally unproblematic inductive argument: All three conclusions are derived by valid inferences. Hence, *prima facie*, this argument might be convincing. However, considering content rather than form, it comes apparent that some of its premises are dubious and its conclusions indeed do not follow from its premises. In the following, I show that the argument suffers from the informal *fallacy of composition*. Consequently, *C3* is a much stronger conclusion than its premises allow.

## 2.5. Fallacy of Composition

To make the informal fallacy apparent, reconsider the aforementioned existing critical accounts on moral *bio*enhancement. In particular Paulo and Bublitz, who argue that it is not only individual moral failure that is the source of major moral problems, but that it 'might be that political representation is not working; it might have to do with powerful stakeholders, national interests, the protection of specific industries and global power structures.' (Paulo and Bublitz 2019a, p. 99) Following this thought, it seems that a major moral problem typically entails at least two disjunct sets of problems: problems that are rooted in individual moral misconduct *qua* suboptimal moral capacity—and problems that are not.

To make this distinction clear, consider for example the major moral problem of ethnic and racial conflicts. *Prima facie*, wrongdoings motivated by racism might be attributable to individual moral deficits, for example genetically inhibited xenophobia. However, further reflection reveals that this rationale is prone to introduce a fishy biological reductionism of human behaviour: By reducing moral failure to neurobiological dispositions, it abstracts a single causal factor out of a much more complex reality. Thereby, it disregards any cultural, political, and economic factors.

In fact, the problem of neurobiological reductionism points at an even more extensive problem: The reductionism of major moral problems to the failed moral decision-making process of individual agents is a worrying tendency inherent to the standard account of *algo*enhancement. As explained above, the default picture is that *algo*enhancement is designed to correct moral decision-making which is compromised or overruled by irrational psychological dispositions. However, the belief that persistent ethnic conflicts can be resolved simply by enhancing the individual participant's moral capacity *qua algo*enhancement seems dubious.

Focussing on individual moral failure disregards the evidently much more complex nature of major moral problems. We have reason to expect that the source of ethnic conflict is not only rooted in a lack of moral motivation, compromised deliberation, or epistemic shortcomings, but also structural features. For example, post-cold war history suggests that ethnic conflict is rather strongly facilitated by perceptions of relative deprivation, ethnic territorial claims, state collapse, security dilemmas, or economic inequality (see Taras and Ganguly 2016, pp. 6–10).

I will come back to this thought below, but for now consider another quick example for the complex composition of major moral problems: In his 2016 monograph *Blood Oil: Tyrants, Violence, and the Rules that Run the World*, Wenar provides an illustrative example of one of today's most pressing major moral problems. His lucid ethical analysis of the global crude oil market exposes the complex causal chains between the individual private consumer and oppressive autocratic regimes, which effectively result in concealed complicity. This example highlights the widespread phenomenon of unintended wrongdoing that is not based on individual moral shortcomings.

The above examples suggest that there is a difference in quality rather than quantity between cases of major moral problems and cases of individual moral wrongdoing. It is indeed sensible to expect that *algo*enhancement will make a difference in particular cases of compromised individual moral decision-making. An instance of the *continuous moral environment monitor*, for example, might very well prevent an unjust judgement in court by informing the accountable judge of his or her biases and any physiological factors that tend to affect moral decision-making and behaviour (see Savulescu and Maslen 2015, p. 85).

However, a (moral) judgement in the context of jurisprudence is a rather special and artificial situation under 'laboratory conditions'. As the above examples show, real-world major moral problems are much more messy. In the context of major moral problems such as ethnic conflicts, an indication by the *situation-specific moral prompter* such as 'are you being influenced by any irrelevant characteristics of the two parties, such as race or gender?' (ibid., p. 87) would probably have a very restricted impact on the problem at large, even if prompted to all individual stakeholders.

This finding suggests that the problem of the standard account of *algo*enhancement is that it tends to mistakenly extrapolate from one part of the problem to the whole of the problem. Crucially, while *algo*enhancement may indeed help to overcome cases of individual moral wrongdoing, these cases of individual moral behaviour do not constitute the whole of major moral problems. The *terminus technicus* for this kind of faulty reasoning is the informal *fallacy of composition*. In *Bad Arguments: 100 of the Most Important Fallacies in Western Philosophy*, Waller provides the following generic example of this kind of fallacy (see Waller 2019, p. 251):

**Fallacy of Composition (Generic)**

1 Congressmen Jones, Mark, and Smith are all radicals.

2 Therefore, Congress is radical.

In this generic example, the problem is evident: The property of some parts is mistakenly extrapolated to the whole. Without further information, we cannot infer from the fact that some members of the Congress are radical that the whole Congress is radical. After all, it could be that other moderate members counter these radical tendencies (see Waller 2019, p. 251). Crucially, the standard account of *algo*enhancement is prone to committing the same fallacy. The corresponding formulation looks like this:

**Fallacy of Composition (Standard Account of *AE*)**

1 Major moral problems contain moral problems that are based on individual failure *qua* impaired moral capacities which can be solved by *algo*enhancement.

2 Therefore, major moral problems are based on individual failure *qua* impaired moral capacities and can be solved by *algo*enhancement.

This formulation pinpoints the problem that the standard account of *algo*enhancement suffers from. The first two conclusions *C1* and *C2* suggest a direct relation between moral capacities and major moral problems: Wherever their moral capacities are impaired, agents facilitate major moral problems—and *vice versa*, wherever moral capacities are optimal, major moral problems are prevented or solved. However, the assumption of this direct link is problematic because major moral problems involve wrongdoing that is not attributable to a lack of individual moral competence, as well as problems that cannot be overcome by means of optimal moral competence. This is what I will argue for in the next section.

# 3. Three Challenges

In the last section, I established that the standard account of *algo*enhancement suffers from the informal *fallacy of composition*. That is, it extrapolates inadmissibly from the fact that major moral problems entail wrongdoings that are grounded in the impaired moral competence of individual agents, that major moral problems at large are reducible to impaired moral competence of individual agents. Furthermore, since *algo*enhancement enhances individual moral competence, the argument goes, it can solve major moral problems. In the following, I will discuss what this fallacy means for the project of *algo*enhancement. Crucially, it generates a pressing problem for the standard account: adherents will have to demonstrate that individual moral competence translates into the solution of large-scale moral problems.

## 3.1. The Problem of Translation

Since this central inferences of the standard account of *algo*enhancement is not sufficiently warranted, it is *question begging* to assume without independent argument that the enhancement of moral capacities has the suggested effect on the solution of major moral problems. However, as Waller rightly notes in the context of the informal *fallacy of composition*, '[i]nferences from a part to a whole can be made if additional assumptions are added to guarantee that the whole will have the property if the parts do.' (see Waller 2019, p. 250)

In the context of this paper, the respective necessary 'additional assumption' would have to indicate that morally enhanced agents are usually better in solving even those problems that are not directly reducible to impaired moral capacities, such as structural problems of poor political representation or economic policy. Indeed, as noted above, proponents grant that higher level structural change is necessary to overcome major moral problems, but that this change requires the intervention of moral enhancement (see p. 9). However, that *algo*enhancement satisfies this function is simply assumed without independent argument. I call this the *problem of translation*:

> **The Problem of Translation**: Major moral problems are dauntingly complex issues with manifold stakeholders and high levels of uncertainty. How can individual moral competence translate into successful reforms under these circumstances?

The *problem of translation* draws attention to the fact that the assumption that morally enhanced agents are better in solving complex major moral problems—*qua* enhanced moral competence—is more complex than assumed and runs into problems. Specifically, I shed light on three challenges for the standard account of *algo*enhancement that arise in the context of translating individual moral proficiency to the level of coordinated global socio-political problem-solving.

In the following, I will discuss the *problem of translation* from three levels of analysis: the individual, the organisational, and the societal level. Firstly, I investigate whether there is a promising way to effectively bypass political organisation in order to shortcut the translation from individual moral competence to socio-political problem-solving. Secondly, I examine potential effects of *algo*enhancement on organisational leaders, individuals which inherit the power to bring significant structural change by themselves. Finally, I explore the level of societal organisation and whether higher voter morality translates into moral success in policymaking.

## 3.2. 'Effective Altruism' and 'Each-We' Dilemmas

In order to avoid *question begging*, proponents of *algo*enhancement need to show that morally competent individuals can solve major moral problems. Above, I indicated that this entails also solving problems that are not grounded in moral incompetence. In the following, I shed light on an influential philosophical school of thought that pursues this very goal on the individual level: the movement of *Effective Altruism*. Largely inspired by Singer's 1972 paper 'Famine, Affluence, and Morality', it is a comprehensive program that promotes individual altruism in order to solve major moral problems. MacAskill, a prominent expert on the field, recently provided the following definition:

> (i) the use of evidence and careful reasoning to work out how to maximize the good with a given unit of resources, tentatively understanding 'the good' in impartial welfarist terms, and (ii) the use of the findings from (i) to try to improve the world. (MacAskill 2019, p. 14)

Accordingly, the project of *Effective Altruism* is characterised by both an intellectual and a practical aspect (see MacAskill and Pummer 2020, p. 4). On the one hand, it is about the scientific search for the most cost-effective way to utilise the given resources in order to do good, drawing on empirical findings and ethical theory. On the other hand, these findings are to be implemented, usually through charitable engagement of individual agents. To illustrate the approach of *Effective Altruism*, consider the real-world major moral problem of persistent global poverty.

World poverty is a notoriously complex problem that bears for persistent widespread disagreement about its causal mechanisms, both in politics and academia (see Gauri and Sonderholm 2012, p. 193). Furthermore, although many consider world poverty to be 'the key moral problem of our time' (Caranti 2010, p. 36), what exactly makes it morally problematic—and what kind of responsibility arises from it—is disputed (see Mieth 2008). Yet, despite all controversy, it is evident that poverty causes a staggering amount of human suffering all over the world. Although the *World Bank* notes that the proportion of people enduring extreme poverty steadily declined over the past two decades (see *Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population) - World* 2018), the numbers are still alarming. According to recent studies, '85% of the world live on less than $30 per day, two-thirds live on less than $10 per day, and every tenth person lives on less than $1.90 per day' (see Roser and Ortiz-Ospina 2013).

In the face of such a daunting challenge, what could individuals possibly do to solve this major moral problem? *Effective Altruism* sets out to answer this question. So-called 'meta-charities' such as *GiveWell* and *Giving What We Can* provide a ranking of the most cost-effective and best performing charity organisations. Drawing on these recommendations, the *Effective Altruist* knows where charitable engagement produces the maximum amount of goodness. In the case of world poverty, this might entail providing mosquito-nets, medical supplies, or critical infrastructure.

At this point, the inherent affinity between the approaches of *Effective Altruism* and *algo*enhancement becomes apparent. Indeed, it does not surprise that Savulescu and Maslen themselves explicitly promote charity in their proposal of *algo*enhancement (see Savulescu and Maslen 2015, p. 86). After all, the *situation-specific moral advisor* (see p. 18) could raise attention to the positive effects of charitable engagement and thus foster moral motivation, while the *continuous moral organiser* (see p. 17) could counteract potential weakness of will by reminding agents to donating more money on a more regular basis in order to comply with their personal moral goals.

Hence, *Effective Altruism* offers a potential solution to the *problem of translation*: by shortcutting the causal chain from the morally competent individual to the actual humanitarian intervention, the individual's moral capacity is directly translated into the partial solution of the major moral problem of world poverty. Thereby, the overwhelmingly complexity of the problem is reduced to the simple act of donating money to the right place at the right time (see Gabriel 2017, p. 457). However, this approach may run into problems: the immediacy of *Effective Altruism* comes at the cost of a lack of coordination.

One of the most classic problems of coordination discussed in game-theory is the *Prisoner's Dilemma*. Imagine two inmates who have to decide whether to back their partner's alibi or not. The irritation of this example comes from the fact that there is a conflict between the individual rational choice and the collective rational choice: While it would be better for both inmates to keep quiet, it is actually more rational to blow the whistle on the partner—from the individual's point of view, that is. Disturbingly, this problem of cooperation exists not only for rational choices but also for moral choices.

Going back to Parfit's *On What Matters: Volume One* (see Parfit 2011, p. 303), it was Temkin who raised the challenge of so-called 'each-we' dilemmas in the specific context of *Effective Altruism* (see Temkin 2019, p. 11). Drawing on Parfit, he argues that in some cases it is true that if each of us—individually—does what he or she ought to do according to a given moral theory, we may—collectively—do more harm than good, so we ought to refrain from that action (see ibid.). Consider the following example:

Agent A seeks to meet the moral obligation to address the problem of world poverty. Hence, she donates to the charity that she considers the most effective. With this money, the charity is financing a vaccination program against deadly diseases. Although agent A knows that programs like these tend to undermine the legitimacy of the local government, she deems her obligation fulfilled. After all, her individual impact on this abstract negative effect is rather small, compared to the fact that her donation directly saved three lives.

Now, as Temkin emphasises, the problem with this case is that 'while [agent A], individually, may have virtually no impact on a government's responsiveness to its citizens; we, together, can have a substantial impact on its responsiveness.' (ibid., p. 15) Thus, there is a conflict between what agent A ought to do from the individual perspective and what agent A ought to do from the collective perspective. After all, together with all the other *Effective Altruists*, agent A's moral dedication might have devastating consequences.

Evidently, this conflict is particularly challenging for the standard account of *algo*enhancement, because it would make things only worse: For what it's worth, enhanced individuals would probably be 'better' *Effective Altruists* and engage even more in philanthropic charity. And chances are, that they run into *each-we dilemmas* and paradoxically create worth outcomes collectively—although being more moral agents individually. Hence, proponents have to meet the challenge of 'each-we' dilemmas in order to avoid that *algo*enhancement does more harm than good.

Ultimately, *Effective Altruism* offers a potentially problematic solution to the *problem of translation*. Importantly, this argument is not dependent on a general scepticism towards the private relief sector. It is but one example of a more general phenomenon. That is, you might agree or disagree with the proposition that relief aid undermines political legitimacy. Either way, the challenge of potential *each-we dilemmas* persist. Proponents of *algo*enhancement thus have to show how to cope with these cases. This is why the phenomenon of 'each-we' dilemmas poses the *problem of translation*.

### 3.3. Politics and the Problem of 'Dirty Hands'

Above, I showed that the potential of *algo*enhancement on the individual level can be challenged because bypassing coordination in the sense of *Effective Altruism* can run into *each-we dilemmas*. In this subsection, I would like to investigate what effect could *algo*enhancement have on the organisational level. Does the moral competence of high-level decision-makers such as prime ministers or CEOs translate successfully into solutions on the global socio-political scale?

In their study *Blind Spots: Why We Fail to Do What's Right and What to Do about It*, Bazerman and Tenbrunsel deny that ethics training of organisational leaders could have prevented the financial collapse of 2008. They point out that 'millions of dollars [were spent] on corporate codes of conduct, value-based mission statements, ethical ombudsmen, and ethical training, to name just a few types of ethics and compliance management strategies'—evidently with disappointing results (Bazerman and Tenbrunsel 2011, p. 4). Their worrying findings vividly underscore the appeal of *algo*enhancement. Drawing on behavioural ethics, Bazerman and Tenbrunsel uncover the psychological limitations in human morality and challenge the traditional emphasis of moral philosophy on rational deliberation. They show that unconscious systematic constraints on our morality have a big impact on human decision-making. One of the strengths of *algo*enhancement over traditional ethical education is that it can raise awareness towards these unconscious psychological limitations. But would it have prevented the financial collapse of 2008?

In order to assess the potential effect of *algo*enhancement on high-level decision-makers in the context of major moral problems, consider the issue of climate change. Since 1970 the amount of flooding, droughts, tropical storms, and forest fires have increased significantly—the symptoms of global warming are evident (see Archer and Rahmstorf 2010, p. 67). Although awareness increased for decades, the reasons for why this is a moral problem is less obvious. While scientific, economic, and geopolitical perspectives usually dominate the discussion, the climate crisis is in fact a major moral tragedy. Importantly, the negative effects of the climate change have the strongest impact on the world's most vulnerable groups, such as the global poor, future generations, and nature itself (see Williston 2019, p. 2). Thus, the climate crisis entails pressing moral problems of global justice and intergenerational equality.

Hence, while the climate crisis poses a scientific challenge, it is also a matter of ethical and political evaluation. What preventive policy measures will be adopted depends on the level of risk society is willing to take and the intrinsic value that it attributes to the survival of people, animals and nature itself (Archer and Rahmstorf 2010, p. 226). Thanks to enormous scientific efforts, we are now well informed about the causes of global warming and also about the measures that would prevent further escalation. So why don't our leaders realise the policies that scientists and activists are calling for so desperately? Is it because they evidently lack moral competence? Are they corrupt and bribed by big business, betraying their own moral values to stay in power? Is their decision-making process biased in favour for personal interests at the expense of the rights of political minorities, like future generations or the global poor? Is it the hypocrisy, pandering, or complacency of the mighty that prevents the solution of this major moral problem?

These are all reasonable hypotheses. Indeed, Gardiner notes that the potential for moral corruption is especially high in the context of the climate crisis, since 'the victims are not yet around to defend the discourse, the potential for moral corruption is especially high.' (Gardiner 2011, p. 46) And if these assumptions are correct, we have reason to believe that *algo*enhancement could help tackling the climate crisis. After all, improving individual moral competence would probably prevent or at least mitigate persistent moral corruption. However, individual vices are but only one side of the coin, as I will show below. Indeed, there is an insightful alternative explanation for the persistence of major moral problems.

A classic historical position of political philosophy can help to explain what is going on here. However, it does not only explain the persistence of major moral problems, but also poses a challenge for the standard account of *algo*enhancement. In his infamous 1513 treatise *The Prince*, Machiavelli developed an account of the successful politician, which is still controversially discussed. It remains to be disturbing, because it points to a substantial conflict between morality and politics: a good politician sometimes ought to do bad things—for the greater good of securing a stable polity.

Whether you agree with this paradoxical statement or not, Machiavelli draws attention to the important insight that public decisions differ substantially from private decisions: they usually have to take conflicting interests into account, imply far-reaching consequences for a large number of people, and are often enforced by coercion (see Bellamy 2010, p. 414). But does this difference in quality warrant for divergent moral standards? Building on Machiavelli's historic challenge, contemporary scholars discuss the phenomenon of so-called 'dirty-hands' dilemmas in the context of political decision-making. Primoratz gives the following example:

> [...] think of a national leader in whose capital a series of bombs will go off in the next 24 hours if they are not discovered and defused, and whose security service have captured a rebel leader who (probably) knows where they are, but refuses to tell. The only way to obtain the information and so prevent the disaster is by torturing him. The leader authorizes torture, although he believes, with the rest of us, that torture is always wrong. How should we judge the leader's decision, and how should the leader feel about it afterwards? (Primoratz 2007, p. xvi)

This example draws attention to the fact that politics is dominated by difficult decisions about how to tackle public problems for which there are no right or wrong solutions, but only better or worse attempts. In the case of the climate crisis, it might be necessary to cooperate with a bellicose state in order to achieve sustainability goals. However, whether a compromise like this is a morally justified decision in the traditional sense is essentially contestable. This is why organisational leaders can not be moral in the same sense that ordinary people can be. Hence, the concept of the 'dirty-hands' dilemmas provides a tool to better understand the nature of major moral problems and what to expect, when we try to overcome them. It debunks the belief that major moral problems would be solved, if only people would be more moral. The moral competence of enhanced organisational leaders will not translate directly into moral decisions in politics. Reality is a lot more tricky than this.

*Algo*enhancement of organisational leaders is a promising approach to overcome the inherent limitations of human moral psychology. However, the phenomenon of 'dirty-hands' dilemmas suggest that even morally enhanced leaders will struggle to meet their obligations when it comes to making difficult public decisions. In the context of major moral problems, sometimes there is no politically viable and morally justified decision, but only foul compromises. This suggests that even in a world exclusively organised by morally enhanced organisational leaders, some major moral problems will persist.[8]

## 3.4. Moral Proficiency and 'Wicked' Problems

The *problem of translation* states that major moral problems are dauntingly complex issues with manifold stakeholders and high levels of uncertainty and poses the question of how individual moral competence can be translated into successful reforms under these difficult circumstances. Above, I looked into how this problem might be solved on the individual level and the organisational level. However, it showed that both approaches run into problems. Hence, lastly, I would like to explore the effects that *algo*enhancement might have on the societal level.

In the beginning of this paper (section 1.3), I explained that existing critical accounts towards *algo*enhancement argue that it is faulty political organisation that is the root of major moral problems. Furthermore, it is argued that moral enhancement might reinforce the political will to realise drastic reforms to overcome our critically flawed socio-economic system.

[8] In this scenario, it comes apparent that *algo*enhancement would pose a collective action problem itself, because in order to have a significant impact, the majority of organisational leaders will have to participate in the moral enhancement programme. Otherwise, the positive effects are prone to fall flat in the global interplay of political agents (see Glannon 2018). However, note that a compulsory programme of moral enhancement would run into problems of political legitimacy (see Paulo and Bublitz 2019b).

Indeed, Persson and Savulescu explicitly propose to enhance the democratic citizenry to achieve morally competent voter's behaviour in order to solve major moral problems (Persson and Savulescu 2012, p. 100). In order to assess this belief and the potential effects of *algo*enhancement on the societal level, I would like to provoke your intuitions with the help of a thought experiment. Consider the following noteworthy quote of the proposal of Lara and Deckers:

> [T]he system would receive, through computers, virtual reality devices or brain interfaces, information from many databases on science, linguistics, logic, and on how people think and reason morally. Moreover, it would collect information from experts in argumentation theory and ethical theory. With the help of sensors, it would also monitor the actual biology and the environment of the agent. The system would then process all this information and, using the aforementioned criteria, engage in a conversation with the agent through a virtual voice assistant. In this conversation, the system would ask a number of questions. These may include the following: *Why? And why? What makes you think that? Is this your last reason? Why do you think this is the best reason? [...]* (Lara and Deckers 2020, p. 284; their italics)

Drawing on the ancient greek philosophical method of maieutics, Lara and Deckers modify Savulescu and Maslen's concept of a *situation-specific moral prompter* (see p. 17) and design an 'artificial Socratic advisor'. Thereby, this proposal comes as close as possible to what a 'pocket-philosopher' might be. After all, the targeted moral capacity is a key discipline of philosophy: deliberate moral reasoning. To pick up on this thought, imagine a world of people with access to their personal 'pocket-philosopher'—or, for the sake of the argument, simply a world of professional moral philosophers. Now ask yourself: Would this world be free from any major moral problem?

For the sake of the argument, let's assume that they are indeed superior moral agents.[9] Hence, shouldn't a polity of morally enhanced professional ethicists be more efficient in tackling major moral problems—by supporting the right parties, by openly demanding adequate domestic and foreign

[9] You might expect moral philosophers to behave differently—morally better—than regular agents. After all, professional ethicists are experts in their field, so that they 'reach correct moral judgments with high probability and for the right reasons.' (Gesang 2010, p. 155) Or are they? In fact, the claim that philosophers are superior moral agents is debatable. For example, Schwitzgebel and Rust provide empirical evidence for that professional ethicists do not vote significantly more often than the average citizen (see Schwitzgebel and Rust 2010). But also beyond political responsibility, experimental data warrants 'to reject the view that ethicists behave, on average, morally better than do non-ethicists.' (Schwitzgebel and Rust 2014, p. 320) However, mind that these are problems of moral motivation and consistency that are explicitly targeted by *algo*enhancement.

policies, or participating in grassroots non-governmental organisations? Shouldn't *algo*enhancement foster moral consensus and political cooperation in order to overcome major moral problems? Surely, a morally competent polity would refrain from supporting outright racist or sexist political parties and candidates. It would probably also be concerned with realising equitable and just conditions of life for everyone. However, at this point, it already becomes apparent that beyond cases of indisputable evil, the situation might be more complex.

History is witness to the fact that moral concepts like justice or equality are not only essentially contested abstract concepts, but require political interpretation and implementation. Hence, despite the challenge of reasonable disagreement about questions of value and their interpretation, it is usually a notoriously difficult task to translate these abstract values into practical political reform. Indeed, when it comes to major moral problems, it might even be an impossible task. After all, according to findings in contemporary studies of public policy, for these kind of problems 'there are no "solutions" in the sense of definitive and objective answers' (Rittel and Webber 1973, p. 155)—major moral problems are 'wicked' problems.

Originally going back to the seminal 1973 paper 'Dilemmas in a general theory of planning', Rittel and Webber introduced the concept of 'wicked' problems in the context of dilemmas in policy-making. Against the backdrop of the popularity of technocratic approaches of the 1960s, the authors argue that societal problems are fundamentally different from the problems of the natural sciences. Whereas the 'tame' problems of, say an engineer, are well-defined and potentially solvable, there are no definite formulations of the 'wicked' problems of political planning (see ibid., pp. 160–1).

To illustrate the 'wicked' nature of major moral problems, reconsider the above discussed examples of world poverty and the climate crisis. Both problems are essentially ill-defined: there is considerable disagreement about what constitutes the problem and even how to describe it properly. Especially in the case of climate change, key scientific findings were publicly contested (see Head 2022, p. 99). In fact, the description is already part of the attempted solution. For example, whether poverty is defined as a lack of income or insufficient social entitlements has a decisive influence on the measures to be taken (see Spicker 2016, p. 1). Furthermore, there are no in principle permissible moves in solving these problems, resulting in persistent disagreement and each attempted solution is a 'one-shot operation' which cannot be undone, has indeterminable consequences and can only evaluated under contrafactual terms (see Rittel and Webber 1973, pp. 161–67). And lastly, world poverty and global warming are interconnected: attempts to reduce poverty often increase global warming and *vice versa*.

When it comes to questions of morality, 'competent reasoners can reach different conclusions about the answer to a given question.' (McMahon 2009, p. 26) Consequently, there is no one single right thing to do in politics from the perspective of morality. It is not a question of moral truth whether a particular republican or democratic proposal constitutes the better policy. If this were the case, we should establish technocratic regimes led by professional moral philosophers. Instead, democratic policies are backed by various, sometimes contrasting values and moral reasons. In foreign affairs, for example, a different concept of justice might be realised compared to domestic affairs. A democratic citizenry is complex and heterogeneous, and so are the programs and policy proposals of democratic parties. And contrary to the natural sciences, 'there is no agreed epistemology or method for selecting between these views other than the process of politics itself.' (see Bellamy 2010, p. 415)

Thus, increased moral competence of a citizenry does not translate directly into better policies to solve 'wicked' major moral problems. The unavoidable uncertainty under which political measures must be developed and implemented guarantees a broad spectrum of reasonable political approaches. This is not about advocating moral relativism. In extreme cases, party programmes can be ruled out from a moral point of view. Nevertheless, various competing policy solutions remain. The 'wickedness' of major moral problems explain why even a polity of moral experts will struggle with solving major moral problems. In best case scenarios, *algo*enhancement can help identifying, expressing, and negotiating latent conflicts of values among stakeholder perspectives—and this is an essential, potentially life-sustaining skill in politics. Moral expertise can provide a guiding compass for navigating societies under circumstances of extreme uncertainty. However, it will not necessarily bring about better political solutions. This is why the 'wickedness' of major moral problems poses the *problem of translation*.

## 3.5. Evaluation

Lastly, I would like to put the above insights into perspective. Despite the potential occurrence of 'each-we' and 'dirty hands' dilemmas, *algo*enhancement may very well have the beneficial effect that more people engage in charitable commitment, participate in politics, or simply open themselves up to political discourse and endure different opinions. And these effects should not be undervalued. Mind that I do not aim to deny the potential positive effects of *algo*enhancement, but rather refine the existing account by drawing attention to critical open challenges. *Algo*enhancement might help to realise the liberal ideal of a tolerant and constructive discourse on political decision-making. Indeed, insofar 'understanding the aspirations and values of the people' (Head 2022, pp. 25–6) is a fundamental skill in tackling 'wicked' problems, *algo*enhancement is a potentially powerful approach and should not be dismissed light-heartedly. However, it is exactly the great potential of *algo*enhancement that should make us wary about illusionary technological solutionism. This is the aim of this paper: to debunk the oversimplified belief that major moral problems are solved, once *algo*enhancement is deployed.

# Conclusion

In this paper, I discussed issues that come apparent when engaging with political and economic problems from the perspective of philosophy. Contributions to the debate of *algo*enhancement simply presuppose without argument that the enhancement of the moral capacities of individual agents has the potential to solve major moral problems. It is assumed that if only people would be better, the problem of poverty or the climate crisis could be solved. In this paper, I shed light on this implicit assumption. The key finding is that this inference is too quick: moral proficiency does not simply translate from the individual level to the level of major moral problems.

In the first section of this paper, I provided an introduction to the field and presented my main claim. The current project of *algo*enhancement does not convincingly argue that it can achieve an important part of what it claims: the solution of large-scale moral problems. Against the backdrop of existing critical accounts, the original contribution of this paper is that it provides a closer investigation of the presupposed link between individual moral competence and the solution of major moral problems.

The second section focussed on the reconstruction and analysis of the standard account of *algo*enhancement. Drawing on dominant proposals, I derived the 'blueprint' of *algo*enhancement and made the underlying argument explicit. The formal analysis demonstrated that the argument suffers from the informal *fallacy of composition*. That is, it tends to mistakenly extrapolate from one part of the problem—individual failure *qua* moral incompetence—to the whole of the problem—a major moral problem, for example the climate crisis. Crucially, this fallacy draws attention to the assumption of a direct link of individual moral competence and the occurrence of major moral problems.

In the third section, I demonstrated that this link of individual moral competence and the occurrence of major moral problems is dubious. Specifically, I showed that it poses the *problem of translation*: since it is *question begging* to assume without argument that *algo*enhancement has the suggested effect on major moral problems, proponents will have to show how individual moral competence translates into successful reform under the peculiar circumstances of major moral problems. Ultimately, I argued that, in order to do so, proponents will have to meet three distinct challenges on the individual, the organisational, and the societal level.

# References

ARCHER, David and Stefan RAHMSTORF (2010). *The Climate Crisis: An Introductory Guide to Climate Change*. Cambridge; New York: Cambridge University Press. URL: www.cambridge.org/9780521407441.

ARP, Robert, Steven BARBONE and Michael BRUCE, eds. (2019). *Bad Arguments: 100 of the Most Important Fallacies in Western Philosophy*. Oxford: Wiley Blackwell.

BAZERMAN, Max H. and Ann E. TENBRUNSEL (2011). *Blind Spots: Why We Fail to Do What's Right and What to Do about It*. Princeton; Oxford: Princeton University Press.

BECK, Birgit (May 2015). 'Conceptual and Practical Problems of Moral Enhancement'. In: *Bioethics* 29.4, pp. 233–240. DOI: 10.1111/bioe.12090. URL: https://onlinelibrary.wiley.com/doi/10.1111/bioe.12090.

BELLAMY, Richard (Oct. 2010). 'Dirty hands and clean gloves: Liberal ideals and real politics'. In: *European Journal of Political Theory* 9.4, pp. 412–430. DOI: 10.1177/1474885110374002. URL: http://journals.sagepub.com/doi/10.1177/1474885110374002.

BOSTROM, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

BUBLITZ, Jan C. (Apr. 2019). 'Saving the World through Sacrificing Liberties? A Critique of some Normative Arguments in Unfit for the Future'. In: *Neuroethics* 12.1, pp. 23–34. DOI: 10.1007/s12152-016-9265-8. URL: http://link.springer.com/10.1007/s12152-016-9265-8.

CARANTI, Luigi (Jan. 2010). 'The Causes of World Poverty: Some Reflections on Thomas Pogge's Analysis'. In: *Theoria* 57.125. DOI: 10.3167/th.2010.5712503. URL: http://berghahnjournals.com/view/journals/theoria/57/125/th5712503.xml.

CRUTCHFIELD, Parker (May 2021). *Moral Enhancement and the Public Good*. Routledge, pp. 1–174. DOI: 10.4324/9781003181798. URL: https://www.taylorfrancis.com/books/9781000401806.

DE ARAUJO, Marcelo (2014). 'Moral Enhancement and Political Realism'. In: *Journal of Evolution and Technology* 24.2, pp. 29–43. URL: http://jetpress.org/v24/araujo.pdf.

DEGRAZIA, David (June 2014). 'Moral enhancement, freedom, and what we (should) value in moral behaviour'. In: *Journal of Medical Ethics* 40.6, pp. 361–368. DOI: 10.1136/medethics-2012-101157. URL: https://jme.bmj.com/lookup/doi/10.1136/medethics-2012-101157.

DODDS, Walter (2019). *The World's Worst Problems*. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-30410-2. URL: http://link.springer.com/10.1007/978-3-030-30410-2.

DOUGLAS, Thomas (Mar. 2014). 'Moral Enhancement'. In: *Enhancing Human Capacities*. Oxford: Blackwell Publishing Ltd, pp. 465–485. DOI: 10.1002/9781444393552.ch34. URL: https://onlinelibrary.wiley.com/doi/10.1002/9781444393552.ch34.

EARP, Brian D, Thomas DOUGLAS and Julian SAVULESCU (July 2017). 'Moral Neuroenhancement'. In: *The Routledge Handbook of Neuroethics*. Ed. by L. Syd M JOHNSON and Karen S. ROMMELFANGER. Vol. 5. New York : Routledge, Taylor & Francis Group, 2017. | Series: Routledge handbooks in applied ethics: Routledge. Chap. 11. DOI: 10.4324/9781315708652. URL: https://www.taylorfrancis.com/books/9781317483526.

ERTEL, Wolfgang (2017). *Introduction to Artificial Intelligence*. 2nd Ed. Undergraduate Topics in Computer Science. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-58487-4. URL: http://link.springer.com/10.1007/978-3-319-58487-4.

FENNER, Dagmar (2019). *Selbstoptimierung und Enhancement: ein ethischer Grundriss*. utb ; 5127 : Philosophie. Tübingen: Narr Francke Attempto Verlag. URL: http://deposit.dnb.de/cgi-bin/dokserv?id=f1d0d7735b374f868cc975e2d5622b9a&prov=M&dok_var=1&dok_ext=htm.

FETZER, James H. (2004). 'The Philosophy of AI and its Critique'. In: *The Blackwell Guide to the Philosophy of Computing and Information*. Ed. by Luciano FLORIDI. Malden; Oxford; Victoria: Blackwell Publishing Ltd. Chap. 9, pp. 119–134.

GABRIEL, Iason (Aug. 2017). 'Effective Altruism and its Critics'. In: *Journal of Applied Philosophy* 34.4, pp. 457–473. DOI: 10.1111/japp.12176. URL: https://onlinelibrary.wiley.com/doi/10.1111/japp.12176.

GARDINER, Stephen M. (May 2011). *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. Oxford; New York: Oxford University Press, pp. 1–512. DOI: 10.1093/acprof:oso/9780195379440.001.0001. URL: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780195379440.001.0001/acprof-9780195379440.

GAURI, Varun and Jorn SONDERHOLM (Dec. 2012). 'Global poverty: four normative positions'. In: *Journal of Global Ethics* 8.2-3, pp. 193–213. DOI:

10.1080/17449626.2012.705787. URL: http://www.tandfonline.com/doi/abs/10.1080/17449626.2012.705787.

GESANG, Bernward (2010). 'Are Moral Philosophers Moral Experts'. In: *Bioethics* 24.4, pp. 153–159. DOI: 10.1111/j.1467-8519.2008.00691.x.

GIUBILINI, Alberto and Julian SAVULESCU (2018). 'The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence'. In: *Philosophy and Technology* 31.2, pp. 169–188. DOI: 10.1007/s13347-017-0285-z.

GLANNON, Walter (Oct. 2018). 'Moral Enhancement as a Collective Action Problem'. In: *Royal Institute of Philosophy Supplement* 83, pp. 59–85. DOI: 10.1017/S1358246118000292. URL: https://www.cambridge.org/core/product/identifier/S1358246118000292/type/journal_article.

HAGENDORFF, Thilo and Katharina WEZEL (June 2020). '15 challenges for AI: or what AI (currently) can't do'. In: *AI & SOCIETY* 35.2, pp. 355–365. DOI: 10.1007/s00146-019-00886-y. URL: http://link.springer.com/10.1007/s00146-019-00886-y.

HARRIS, John (2016). *How to be Good: The Possibility of Moral Enhancement*. New York: Oxford University Press.

HEAD, Brian W. (2022). *Wicked Problems in Public Policy: Understanding and Responding to Complex Challenges*. Cham: Palgrave Macmillan. DOI: 10.1007/978-3-030-94580-0. URL: https://link.springer.com/10.1007/978-3-030-94580-0.

HENNING, Klaus (2021). *Gamechanger AI: How Artificial Intelligence is Transforming our World*. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-52897-3. URL: http://link.springer.com/10.1007/978-3-030-52897-3.

KLINCEWICZ, Michał (2016). 'Artificial Intelligence as a Means to Moral Enhancement'. In: *Studies in Logic, Grammar and Rhetoric* 48.1, pp. 171–187. DOI: 10.1515/slgr-2016-0061.

LARA, Francisco (Aug. 2021). 'Why a Virtual Assistant for Moral Enhancement When We Could have a Socrates?' In: *Science and Engineering Ethics* 27.4, p. 42. DOI: 10.1007/s11948-021-00318-5. URL: https://doi.org/10.1007/s11948-021-00318-5%20https://link.springer.com/10.1007/s11948-021-00318-5.

LARA, Francisco and Jan DECKERS (2020). 'Artificial Intelligence as a Socratic Assistant for Moral Enhancement'. In: *Neuroethics* 13.3, pp. 275–287. DOI: 10.1007/s12152-019-09401-y.

LAVAZZA, Andrea and Massimo REICHLIN (2019). 'Introduction: Moral Enhancement'. In: *Topoi* 38.1, pp. 1–5. DOI: 10.1007/s11245-019-09638-5. URL: http://dx.doi.org/10.1007/s11245-019-09638-5.

MACASKILL, William (2019). 'The Definition of Effective Altruism'. In: *Effective Altruism: Philosophical Issues*. Ed. by Hilary GREAVES and Theron PUMMER. Oxford: Oxford University Press. Chap. 1, pp. 10–28.

MACASKILL, William and Theron PUMMER (June 2020). 'Effective Altruism'. In: *International Encyclopedia of Ethics*. Wiley, pp. 1–9. DOI: 10.1002/9781444367072.wbiee883. URL: https://onlinelibrary.wiley.com/doi/10.1002/9781444367072.wbiee883.

MACHIAVELLI, Niccolo (Jan. 2019). *Machiavelli: The Prince*. Ed. by Quentin SKINNER and Russell PRICE. Cambridge: Cambridge University Press. DOI: 10.1017/9781316536223. URL: https://www.cambridge.org/highereducation/books/machiavelli-the-prince/ACCEE83504D6C76F2D8D93064F20BEF5#contents.

MCMAHON, Christopher (July 2009). *Reasonable Disagreement: A Theory of Political Morality*. Cambridge; New York: Cambridge University Press. DOI: 10.1017/CBO9780511596742. URL: https://www.cambridge.org/core/product/identifier/9780511596742/type/book.

MIETH, Corinna (Feb. 2008). 'World Poverty as a Problem of Justice? A Critical Comparison of Three Approaches'. In: *Ethical Theory and Moral Practice* 11.1, pp. 15–36. DOI: 10.1007/s10677-007-9088-0. URL: http://link.springer.com/10.1007/s10677-007-9088-0.

PARFIT, Derek (2011). *On What Matters: Volume One*. Ed. by Samuel SCHEFFLER. Oxford; New York: Oxford University Press.

PAULO, Norbert and Jan C. BUBLITZ (2019a). 'How (not) to Argue For Moral Enhancement: Reflections on a Decade of Debate'. In: *Topoi* 38.1, pp. 95–109. DOI: 10.1007/s11245-017-9492-6.

— (Apr. 2019b). 'Pow(d)er to the People? Voter Manipulation, Legitimacy, and the Relevance of Moral Psychology for Democratic Theory'. In: *Neuroethics* 12.1, pp. 55–71. DOI: 10.1007/s12152-016-9266-7. URL: http://link.springer.com/10.1007/s12152-016-9266-7.

PERSSON, Ingmar and Julian SAVULESCU (July 2012). *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199653645.001.0001. URL: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199653645.001.0001/acprof-9780199653645.

PERSSON, Ingmar and Julian SAVULESCU (Jan. 2015). 'The Art of Misunderstanding Moral Bioenhancement'. In: *Cambridge Quarterly of Healthcare Ethics* 24.1, pp. 48–57. DOI: 10.1017/S0963180114000292. URL: https://www.cambridge.org/core/product/identifier/S0963180114000292/type/journal_article.

*Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population) - World* (2018). URL: https://data.worldbank.org/indicator/SI.POV.DDAY?end=2018&locations=1W&name_desc=true&start=1981&view=chart.

POWELL, Russell and Allen BUCHANAN (2016). 'The Evolution of Moral Enhancement'. In: *The Ethics of Human Enhancement: Understanding the Debate*. Ed. by Steve CLARKE et al. New York: Oxford University Press. Chap. 17. Pp. 239–160.

POWERS, Thomas M. and Jean-Gabriel GANASCIA (July 2020). 'The Ethics of the Ethics of AI'. In: *The Oxford Handbook of Ethics of AI*. Ed. by Markus D. DUBBER, Frank PASQUALE and Sunit DAS. New York: Oxford University Press. Chap. Chapter 2, pp. 27–52. DOI: 10.1093/oxfordhb/9780190067397.001.0001. URL: https://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190067397.001.0001/oxfordhb-9780190067397.

PRIMORATZ, Igor (2007). *Politics and Morality*. Ed. by Igor PRIMORATZ. London: Palgrave Macmillan UK. DOI: 10.1057/9780230625341. URL: http://link.springer.com/10.1057/9780230625341.

RAUS, Kasper et al. (Dec. 2014). 'On Defining Moral Enhancement: A Clarificatory Taxonomy'. In: *Neuroethics* 7.3, pp. 263–273. DOI: 10.1007/s12152-014-9205-4. URL: http://link.springer.com/10.1007/s12152-014-9205-4.

RITTEL, Horst W. J. and Melvin M. WEBBER (June 1973). 'Dilemmas in a general theory of planning'. In: *Policy Sciences* 4.2, pp. 155–169. DOI: 10.1007/BF01405730. URL: http://link.springer.com/10.1007/BF01405730.

ROSER, Max and Esteban ORTIZ-OSPINA (2013). *Global Extreme Poverty*. URL: https://ourworldindata.org/extreme-poverty.

RUSSELL, Stuart Jonathan et al. (2016). *Artificial Intelligence: A Modern Approach*. Ed. by Stuart Jonathan RUSSELL and Peter NORVIG. 3rd Ed. Vol. 48. Hoboken: Pearson Education.

SAVULESCU, Julian and Hannah MASLEN (2015). 'Moral Enhancement and Artificial Intelligence: Moral AI?' In: *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*. Ed. by Jan ROMPORTL, Eva ZACKOVA and Jozef KELEMEN. Vol. 9. Topics in Intelligent Engineering and Informatics. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-09668-1. URL: http://link.springer.com/10.1007/978-3-319-09668-1.

SCHWITZGEBEL, Eric and Joshua RUST (June 2010). 'Do Ethicists and Political Philosophers Vote More Often Than Other Professors?' In: *Review of Philosophy and Psychology* 1.2, pp. 189–199. DOI: 10.1007/s13164-009-0011-6. URL: http://link.springer.com/10.1007/s13164-009-0011-6.

— (June 2014). 'The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude, and directly observed behavior'. In: *Philosophical Psychology* 27.3, pp. 293–327. DOI: 10.1080/09515089.2012.727135. URL: http://www.tandfonline.com/doi/abs/10.1080/09515089.2012.727135.

SEVILLE, Helen and Debora G. FIELD (2011). 'What Can AI Do for Ethics?' In: *Machine Ethics*. Ed. by Michael ANDERSON and Susan Leigh ANDERSON. New York: Cambridge University Press. Chap. 28. Pp. 499–511.

SINGER, Peter (1972). 'Famine, Affluence, and Morality'. In: *Philosophy & Public Affairs* 1.3, pp. 229–243.

SPICKER, Paul (2016). 'Poverty as a Wicked Problem'. In: *CROP Poverty Brief* 35, pp. 1–4. URL: http://www.crop.org/CROPNewsEvents/Poverty-as-a-wicked-problem.aspx.

SUTROP, Margit (2020). 'Challenges of aligning artificial intelligence with human values'. In: *Acta Baltica Historiae et Philosophiae Scientiarum* 8.2, pp. 54–72. DOI: 10.11590/ABHPS.2020.2.04.

TARAS, Raymond C. and Rajat GANGULY (2016). *Understanding Ethnic Conflict*. London; New York: Routledge.

TEMKIN, Larry S (2019). 'Being Good in a World of Need : Some Empirical Worries and an Uncomfortable Philosophical Possibility'. In: *Journal of Practical Ethics* 7.1, pp. 1–23. URL: http://www.jpe.ox.ac.uk/papers/being-good-in-a-world-of-need-some-empirical-worries-and-an-uncomfortable-philosophical-possibility/.

THAMM, Alexander, Michael GRAMLICH and Alexander BOREK (2020). *The Ultimate Data and AI Guide: 150 FAQs About Artificial Intelligence, Machine Learning and Data*. Munich: Data AI Press.

TURNER, Jacob (2019). *Robot Rules: Regulating Artificial Intelligence*. London: Palgrave Macmillan. DOI: 10.1007/978-3-319-96235-1.

WALLER, Jason (2019). 'Composition'. In: *Bad Arguments: 100 of the Most Important Fallacies in Western Philosophy*. Ed. by Robert ARP, Steven BARBONE and Michael BRUCE. Oxford: Wiley Blackwell. Chap. 53. Pp. 250–251.

WENAR, Leif (2016). *Blood Oil: Tyrants, Violence, and the Rules that Run the World*. Oxford; New York: Oxford University Press.

WILLISTON, Byron (2019). *The Ethics of Climate Change: An Introduction*. Oxon; New York: Routledge.