Probing the risks of moral enhancement

Joao Lourenco de Araujo Fabiano

St Cross College, University of Oxford

October 2018

Thesis for the degree of Doctor of Philosophy

Word-count: 58,269

Probing the risks of moral enhancement

Joao Lourenco de Araujo Fabiano

St Cross College, University of Oxford

October 2018

Thesis for the degree of Doctor of Philosophy

Word-count: 58,269

Abstract

Attempting to improve fundamental moral dispositions with technology is prone to unexpected consequences. These dispositions are complex, and fragile when faced with technological interventions. For instance, a drug increasing the disposition to co-operate between individuals can unexpectedly lead to group conflict because it can increase parochialism. Furthermore, these interventions might be detrimental to a person's psychological unity over time and thus might undermine interests whose realisation is sensitive to the continuation of this unity over time. Likewise, they might be detrimental to interests sensitive to the continued chain of human generations. Notwithstanding these risks, there are strong reasons in favour of attempting to improve moral dispositions; it could significantly decrease a wide range of extreme risks that arise from our moral failings (for example, nuclear war from lack of global co-operation). Appealing to a form of virtue theory as a guiding framework is likely to avoid these risks and to address other concerns expressed in the literature. If we apply such a framework, technological interventions aimed at improving fundamental moral dispositions are extremely desirable.

Acknowledgements

I am grateful for receiving the support of more people than I can name or remember. I am indebted to Julian Savulescu, who was my supervisor throughout my doctorate. He offered encouragement and support when applying to doctoral programmes in philosophy, helping me to pursue a doctorate at an elite university. He has done seminal work in the subject of this thesis, making it possible for me even to think about moral enhancement in the first place. Together with Joshua Shepherd and Roger Crisp, who co-supervised me at different periods, Julian has provided me with countless comments and meetings. Joshua gave me feedback on the initial rough drafts of this thesis and helped sharpen my philosophical reasoning. I had many insightful meetings with Anders Sandberg, who collaborated with me on related research. Guy Kahane gave me extensive feedback both times he acted as an examiner of my early doctoral research – once with Roger and a second time with Jeff McMahan. I am thankful for the many constructive comments from my defence examiners, Allen Buchanan and Tom Douglas. Moreover, without the works of Jeff and Roger on personal identity and virtue ethics, respectively, it would have been harder (if not impossible) to build a systematic understanding of these questions. I am thankful they took the time to write accessible and foundational materials on these matters. Jeff met with me twice to discuss earlier drafts of a chapter of this thesis when he was not under any expectation to do so. Nick Bostrom provided me with guidance when I was still an undergraduate student in Brazil with little idea of how to build a career in academic philosophy. His work motivated me to investigate the long-term risks of moral enhancement.

From the writing of my doctoral research proposal to submitting this thesis, I appreciate the various levels and types of feedback I received from (in no particular order): Daniel Kokotajlo, Emma Bates, Tena Thau, Nick Beckstead, Ricardo Bindi, Neil Levy, Diego Caleiro, Regina Reni, Owen Schaefer, Lucas Nascimento Machado, C'zar Bernstein,

Brian Earp, Jonathan Pugh, Dominic Wilkinson, Nadira Faber, Osvaldo Pessoa Jr., and the many supporters whose names have gone unacknowledged here.

For as long as I can remember, my father – Luiz Hermenegildo Fabiano – has instigated my philosophical curiosity and influenced me to spend an unusual amount of my pre-adult life reading philosophy. I thank him and my mother, Maria Regina de Araujo, for their material, emotional and intellectual support throughout my life. Through the most stressful moments of this thesis, I am thankful for having had the emotional and grammatical support of my girlfriend, Jane Turula.

Table of contents

Introduction

The aim of this thesis is to investigate the possible negative consequences of using technology to alter human morality. I will focus on the possibility that attempting to improve human morality with the use of technology – *moral enhancement*[1] – could in fact yield a future with diminished moral value. My analysis will be novel in that I will focus on risks that could arise even if the claims of moral enhancement advocates are true and some arguments against it unsound. Such an investigation will be broadly relevant as any sensible view of morality ought to take significant negative outcomes seriously when analysing the use of a potential new intervention.

Consideration of the history of technology reveals past cases where most people would agree we failed to use advances correctly and ethically, such as nuclear and chemical weapons. Persson & Savulescu have argued that many of these ethical mistakes arose in part due to flaws in our moral character.[2] Alongside others, Persson & Savulescu have examined the possibility of improving us morally as a response to our moral unfitness to solve co-operation problems arising from the development of powerful technologies. We may only be able to survive the introduction of even more powerful and complex technologies if we enhance our moral dispositions and ability to co-operate (Persson & Savulescu, 2012, pp. 100- 134).

My thesis, while conceding the desirability of moral enhancement and considering some critiques of moral enhancement to be unsuccessful, aims to investigate as yet unexplored long-term risks of moral enhancement. Solving humanity's moral unfitness is deeply needed. If we assume that moral enhancement is feasible, while denying that there is

[1] See second paragraph of Chapter 1 and associated footnote for a discussion of definitional issues.

[2] Their main work on the topic is Persson & Savulescu, 2012. In this thesis, when their names are mentioned without an year, it refers to their whole body of work on moral enhancement used in this thesis, that is Persson & Savulescu (2008, 2010, 2011, 2012, 2013a, 2013b, 2014a, 2014b, 2017) and Savulescu & Persson (2012).

anything necessarily wrong about changing our moral character, then could we still have reasons to be wary of moral enhancement?

Moral enhancement carried out to the necessary extent to solve our moral unfitness will cause significant changes to fundamental traits; it will be what I call *deep moral enhancement*. Deep moral enhancement is relatively prone to unexpected consequences and is directed at traits with an unusually high degree of complexity when compared to most forms of human enhancement. What's more, deep moral enhancement might be detrimental to personal identity, thus undermining interests whose realisation depends on the existence of the same individuals, or type of individuals, who originally held those interests. However, deep moral enhancement could significantly decreases a wide range of extreme risks that arise from our moral failings (e.g. nuclear war from lack of global co-operation). I argue that using some form of virtue theory as a guiding framework for deep moral enhancement is likely to avoid these risks and to address other concerns present in the literature. I conclude that if we were to apply such a framework, deep moral enhancement would be extremely desirable.

To introduce the discussion, I will review the literature on moral enhancement, with a specific focus on its critiques. I examine more than 100 academic articles and classify their criticisms into five general categories as relating to unfeasibility, failure to improve morality, threats to freedom, flawed reasons, and undesirable social effects. I conclude many of these critiques fail to offer any strong reason against moral enhancement, while some point to real potential risks on which I will focus the next two Chapters.

In the second Chapter, I will argue that (1) a proper account of human moral traits faces major epistemic difficulties, and (2) deep moral enhancement is relatively prone to unexpected consequences. I draw support for the first claim from the aetiology of moral traits, its unusually high susceptibility to contingencies, and epistemic difficulties arising

from the first-person view of human morality. To support the second claim, I first argue that even an apparently straightforward example of moral enhancement such as increasing human co-operation could plausibly lead to unexpected harmful effects. Secondly, I generalise the example and argue that technological intervention on individual moral traits will often lead to paradoxical effects on the group level. Thirdly, I contend that in so far as deep moral enhancement targets motivation, it is prone to be self-reinforcing and irreversible. Finally, I conclude from those two claims that attempts at deep moral enhancement pose greater risks than other enhancement technologies. For example, one of the major problems that moral enhancement is hoped to address is lack of co-operation between groups. If humanity developed and distributed a drug that dramatically increased co-operation between individuals, we would be likely to see a decrease in co-operation between groups and an increase in the disposition to engage in further modifications, which are both potential problems.

I will explore the consequences of deep moral enhancement to moral status in the third chapter. If we accept both that human enhancement could produce beings with higher moral status than our own (i.e. supra-persons), and that this scenario will be detrimental to persons, then there are still several ways that the creation of supra-persons could be defended. I will argue that some of these defences fail because they do not account for interests whose realisation depends on the existence of the same individuals, or type of individuals, who held those interests. As such, I will concentrate on scenarios where moral status enhancement produces significant changes to psychological continuity. I will consider three scenarios: the creation of supra-persons within one generation, the enhancement of persons into supra-persons, and the generational replacement of persons by supra-persons. For each of these, I will briefly mention the concerns that they most often elicit in the literature, explain some of the strongest arguments against these concerns, and then argue

that a thorough consideration of individual interests undermines these arguments. In all three cases I conclude that although the creation of drastically morally better persons may be desirable and likely to help realise most of what we currently value, this realisation is undermined by the fact the beings enjoying this future might not be our continuants. Our interest in living in a more co-operative and efficient society can be wholly fulfilled by us living in such a society; it can only be somewhat fulfilled by radically enhanced versions of us living in such a society. It will be safer, not to mention more pragmatic and intuitive, to centre the question around what sort of people we want to become using our current moral traits as a starting point.

In the fourth Chapter, I will argue that deep moral enhancement is likely to address a wide range of extreme risks that threaten the long-term future of humanity, while also causing the extreme risks addressed in my preceding arguments. I investigate the various reasons why extreme risks should be among the most important considerations when evaluating the impact of a new powerful technology. I enumerate known sources of these risks and how they relate to large-scale co-operation. Deep moral enhancement may introduce new extreme risks, but can also be effective in alleviating known sources of extreme risks. Arguments in favour of traditional forms of moral improvement are considered, but I argue they do not justify a strict preference for traditional methods because that would require an unreasonable certainty regarding the efficiency of standard political processes to deal with dramatically new risks.

The last Chapter will outline some prescriptions to avoid the risks explored in the previous chapters. The extreme risks of deep moral enhancement require the use of a safety framework for the project of improving moral traits; such a framework should guarantee practical robustness to moral uncertainty, empirical adequacy, correct balance between traits, preservation of identity, and be sensitive to practical considerations such as emergent social

effects. I argue that some form of virtue theory is likely to fulfil those desiderata and to address other concerns present in the literature. I compare previous attempts at developing a virtue theory framework for deep moral enhancement with other possible suggestions and conclude some alternatives are superior but still incomplete. I propose research directions for those frameworks and suggest some practical applications. My final conclusion will be that if we properly develop and apply this safety framework, the project of moral enhancement can avoid the risks I explored while decreasing most known sources of extreme risks.

Chapter 1: Critical review of arguments against moral enhancement

1. Introduction

I will briefly introduce both Persson & Savulescu's moral obligation and Douglas' moral desirability arguments and proceed to classify and analyse some of the major criticisms such arguments have encountered. Since the publication of the first two articles on the subject in 2008, moral enhancement has provoked intense debate within applied ethics and related disciplines. In just ten years, more than 100 academic articles have been published on the issue. The main advocates of moral enhancement have contended that it is either a moral obligation, given it would prevent self-destruction due to our moral failings, or at least that it is morally permissible, given it would increase a person's moral desirability. Both statements were met with criticism, even by those who elsewhere endorse human enhancement.

I will define *moral enhancement* as technological interventions reasonably expected to improve moral behaviour or motives, similar to Douglas' (2008, p. 229) definition of biomedical moral enhancement. My definition will be looser than that used by Douglas because I intend to provide an overview of most of the literature criticising moral enhancement and some authors slightly depart from Douglas' definition. In the next Chapter, I will define more rigorously the type of moral enhancement this thesis will explore (i.e. deep moral enhancement).[3]

[3] From the meaning of the words alone, one would be inclined to classify traditional interventions as moral enhancement. Some authors agree with this classification and then proceed to define a more specific term such as biomedical moral enhancement or technological moral enhancement; other authors do not (for a review of definitions see Raus, Focquaert, Schermer, Specker, & Sterckx, 2014). As a matter of usage, a search for the term on academic databases will reveal that the overwhelming majority of academic publications that use the

A variety of connected counter-arguments have been put forward. Bearing in mind the further investigation they will receive in this thesis, they will be divided into five categories: (1) moral enhancement is not feasible, (2) moral enhancement will not result in an improved morality, (3) moral enhancement is a threat to freedom, (4) reasons to morally enhance are flawed, and (5) moral enhancement will not produce the desired social and political effects.[4]

2. Two defences of moral enhancement

2.1 Moral enhancement as socially mandatory: Persson & Savulescu

Over the course of the last century, humanity's inability to co-operate on an international scale has become a major concern, particularly when problems on the scale of global warming and nuclear proliferation have arisen. Persson & Savulescu (2008, 2012) have argued that we are not equipped with the right set of traits and morals to solve these problems. We can co-operate in small groups because being able to efficiently interact in a small gathering was a recurrent evolutionary pressure that selected traits adapted to do so. Generosity, altruism, a sense of fairness and the desire to punish cheaters are easily expressed by most people, enabling us to co-operate. However, we do not possess the ability to co-operate well in extremely large groups spread across countries and territories or from different ethnicities and backgrounds. In addition, most people now have the capacity to cause harm to a large group of individuals (e.g. through a terrorist attack), whereas few can

term moral enhancement are about technological methods. In fact, there seems to be no publication on traditional moral education or moral progress using the term moral enhancement. As observed by Raus et al. (2014), a lack of rigorous definition is a problem for anyone attempting to conduct a thorough investigation in the area. I fully agree and offer a detailed definition in my next Chapter. However, due to the usage of the term, adopting a loose definition of moral enhancement, for now, will enable a more extensive literature review of the subject, and it will not incorrectly include studies on traditional means. Furthermore, I reserve the use of the term *moral improvement* to mean traditional means of improving human morality, such as moral education.

[4] This literature review was mostly conducted in 2015, after which only a selected number of articles were included. For a more extensive and purpose-neutral literature review on moral enhancement carried out around the same year see Specker, Focquaert, Raus, Sterckx, & Schermer (2014).

save the same amount of lives. This situation is aggravated because we are being conferred an ever-increasing destructive power and technology is rapidly becoming globalised. This means that the probability of any particular individual having enough power to destroy the whole of humanity has increased. Guaranteeing none of them will not exercise such power requires high levels of large scale co-operation. Therefore, the two authors conclude, we have a moral imperative to pursue moral enhancement, for not doing so will expose humanity to extreme risks of catastrophes or extinction – what they call *ultimate harm*. Further scientific progress, although providing small increases to our high quality of life, would also increase our already alarming destructive powers and would, thus, be harmful in the absence of moral enhancement.

2.2 Moral enhancement as individually permissible: Douglas

Instead of focusing on the moral obligation that society has to promote the development of moral enhancement, Douglas (2008) has analysed the moral permissibility of a single individual voluntarily performing moral enhancement. According to Douglas, while many forms of human enhancement are commonly considered morally impermissible on the grounds that they produce an advantage for the individual at the cost of a disadvantage for society, these same grounds could not be applied to moral enhancement. If a person enhances herself so that she will have morally better motives, or so that her actions conform better to common moral expectations, then this has clear advantages to society as a whole. He has later argued that one straightforward way to perform moral enhancement would be through direct emotional modulation, without the need to target deliberative capacities directly. Douglas' (2013b) central example is that of an US judge in a multi-ethnic area under the influence of negative emotional biases against African-Americans. He decides to take a drug to attenuate this emotional reaction that he deems wrong. This drug would function

merely by reducing the emotional response, without any direct alteration of beliefs or deliberation.

3. Five areas of criticism

3.1 Moral enhancement is not feasible

One common belief of many opponents of moral enhancement is that it will simply not happen. Such a belief is at times the result of other scepticisms towards moral enhancement or a reason itself for opponents to be sceptical.

Agar (2013a) contends that improving human morality beyond normal levels would be extremely difficult. He maintains that human moral dispositions lie on a delicate balance that moral enhancement will be likely to disturb, shattering our morality into an undesirable configuration rather than actually enhancing it. According to him, our morality is based on both rational cognitions and emotional intuitions – an equilibrium he calls moral normalcy – and shifting our inner tendencies in either direction holds the risk of causing catastrophic imbalance. He thus states: "[It is] unlikely to be any pills or injections that directly produce in us morally superior judgments or motivations" (p. 1). In the next Chapter, a stronger version of this objection will reveal one of the main risks of moral enhancement. I will argue that human morality not only stands on a fragile balance between the two tendencies Agar mentions but that it contains an overabundance of tendencies that are in fragile balance.

Given many of the conceptual issues that would need to be settled before properly pursuing moral enhancement, Beck (2014) concludes that "moral enhancement is not very likely to be made sense of – let alone realised – in the medium-term future" (p. 2). Sparrow (2013) believes the main reason that moral enhancement will not help with our global problems is that "we don't have any such technology and we are unlikely to develop it within the time available to us" (p. 407), and this is so because our global problems are of a

political/social nature and not due to neurochemical imbalances on our individual morality. While Harris (2011) concedes that certain types of technologies could, in principle, improve human morality, he contends that the necessary fine-tuning is unlikely to be achieved: "I for one am sceptical that we would ever have available an intervention capable of targeting aversions to the wicked rather than the good" (p. 105).

Adding to the claim that most proposed moral enhancements will lack the necessary fine-tuning for a moral intervention, Chan & Harris (2011) analyse the case of proposed moral enhancements that would work by decreasing aggression or increasing empathy. Merely deploying an emotional moral enhancement that decreased aggressiveness would not result in an overall moral enhancement, as aggressive reactions towards unfairness are an important component of moral behaviour.[5] Increasing overall empathy would be ineffective as what needs to be improved is with whom we empathise and not the overall feeling. The problem is precisely that we feel too much empathy for those who are close and too little for those who are far from our social group.[6] Increasing empathy for those who are already its targets would actually exacerbate the problem. The manipulation of these predispositions needs to be extremely context-dependent in order to count as moral enhancement; some believe only moral education can deliver such a degree of specificity. Both parties agree fine-tuning is necessary, but Harris, Chan, Agar and others seem to hold that such fine-tuning will not happen in any relevant future time-frame.

Douglas (2014a) also entertains the possibility that more brute and direct forms of moral enhancement could be less reliable as they would be less context-dependent. He concludes that traditional forms of moral improvement are also contingent and depend on

[5] I take Chan & Harris' (2011) meaning of aggression towards unfairness to be any action (not necessarily violent) that is initially costly for both parties engaging in the behaviour but that can lead to a later resolution of an unfair situation, such as a political protest, strikes or punishment for norm violations in general.

[6] For more on the case against empathy in general, see Prinz (2011) and Bloom (2016). Persson & Savulescu (2018) also offer a discussion on how empathy needs to be properly directed, but has an important role in motivating moral behaviour.

specific circumstances in order for the right type of deliberation to take place (e.g. absence of distractions and temptations); and that moral enhancement would not be less reliable. As long as we make sure that moral enhancements are accompanied by a set of specific and suitable circumstances, in the same way moral education is, then they are likely to be no less reliable than conventional moral education. He mentions Wasserman's observation that with time, direct emotional interventions could lead to moral understanding: for instance, making someone feel less subconscious racial aversion could make him see that racial group in a new light and lead to the more general understanding that they deserve equal treatment, thus resulting in a more refined form of moral improvement.

Furthermore, there is much disagreement as to what degree of optimism is appropriate regarding the implications for the moral enhancement project of the current state of scientific research on the neuroscience of morality. For instance, consider the case of SSRIs. Crockett (2010) has claimed her studies provide evidence that SSRIs could promote pro-social behaviour, and this is routinely cited as preliminary evidence that one day moral enhancement will be feasible. Meanwhile, critics have gone so far as saying that SSRIs could be moral de-enhancers (Chand & Harris, 2011), or unreliable, unsafe and ineffective compared to alternatives (Wiseman, 2014).

Persson & Savulescu (2014b) reply that it is not a part of their argument that moral enhancement is necessarily feasible; they hold only that if it could be feasible, then we ought to pursue its development. They also present several studies that indicate that we are gradually understanding the neurobiological pathways of moral behaviour and that if such development continues, then neurochemical manipulations of morality will follow. Moreover, Douglas (2014a) notes that there is no nomological violation in the concept of moral enhancement, i.e. it does not seem to violate any known natural law, hence it is at least physically possible; which could be sufficient for Persson & Savulescu's argument. Finally,

Persson & Savulescu (2014b) argue that any form of human behaviour, whether its source is social or moral, is mediated by brain activity and thus modifying brain activity has to produce behavioural modifications, be that behaviour primarily social/moral or not.

3.2 Moral enhancement will not result in an improved morality

Some authors consider that the interventions being proposed will not result in improved morality. Instead of claiming there is something empirically wrong with moral enhancement, rendering it unfeasible, these objections claim that there is something *conceptually* wrong with it. Thus, according to these authors, even if these technologies are developed and work as technically intended, those changes will not amount to actual moral improvement.

One of the main debates is centred on whether direct modulation of morally undesirable emotions would in fact amount to moral improvement. Douglas argues that inhibiting morally undesirable emotions such as racial aversion and aggressiveness, even without any direct modulation of cognition, would increase the moral desirability of an individual's motives or behaviours as it would lead to better conformity to common moral expectations of fairness, equality and harm-avoidance. On the other hand, Harris and others believe that morality cannot be improved without the direct improvement of moral reasoning, which they believe to be a central faculty of human morality. Further, they are highly sceptical that moral reasoning could be improved by means other than conventional forms of moral education, or, if enhancement is required, by cognitive enhancement. In particular, Harris (2011) asserts that racial biases are the result of mistaken beliefs and focusing on emotions would miss the target.

While both Douglas and Persson & Savulescu believe the goal of moral enhancement is ultimately to lead to better behaviour or motives, others are of the opinion that morality

cannot be reduced to behaviour while disregarding beliefs, values and cognition (Jotterand, 2014). Harris (2013a) states:

> "It seems to me that moral enhancement, properly so called, must not only make the doing of good or right actions more probable and the doing of bad ones less likely, but must also include the understanding of what constitutes right and wrong action." (p. 172).

As noted by Baertschi (2014) and Sorensen (2014), these disagreements seem to be rooted in a more fundamental meta-ethical disagreement on the nature of morality. Baertschi draws attention to the fact that the real disagreement between Douglas and Harris is whether emotions or cognition lie at the foundation of human morality. Although Baertschi classifies Douglas as an emotionalist who apparently presupposes that emotions are the base of morality, he merely needs to presuppose that manipulation of emotions can manipulate morality. Meanwhile, Harris would only need to presuppose emotions cannot, or can barely, influence morality. The latter is a much bolder assumption. There is also no reason to assume that moral enhancement can only be done via emotional manipulation.

Sorensen has attempted to formulate possible reasons as to why moral enhancement might be intrinsically wrong. For instance, he draws attention to the fact that if there is a cross-temporal effect on moral value, then certain types of morally negative desires leading one to perform moral enhancement (e.g. the desire to receive moral appraisal) could diminish any higher moral desirability achieved through moral enhancement. If value is cross-temporally dependent it means that value at a time t_2 could be affected by a previous time t_1, independently of any causal role t_1 has on t_2. The same event X at t_2 could have more or less moral value depending on whether Z or Y happened at t_1. For instance, you live now in t_1. If suddenly in t_2 you were replaced by an alien individual with the same amount of value as you would otherwise have in t_2, then t_2 may not have the exact same amount of value as it would otherwise have, simply because in t_1 you were alive and the alien's previous time slice was not. (I have only found this terminology being directly used by Sorensen himself, but

others have considered similar effects: Velleman, 1991; Vellentyne, 1988). Besides morally negative desires, candidates for diminishing cross-temporal effects when performing moral enhancement are its relative effortlessness, psychological discontinuity,[7] and lack of achievement or desert. However, he concludes most of these effects would either not apply to moral enhancement, also apply to conventional moral education, or have counter-intuitive consequences if true. Nonetheless, there could still be other forms of cross-temporal dependence that reduce all future goodness if it were to be achieved through moral enhancement, perhaps even to the point of outweighing any higher moral desirability moral enhancement would confer.

Douglas (2014a) himself contemplates the possibility that effort might play a role in moral worth and that moral enhancement might be less conducive to moral worth than other more laborious forms of moral improvement, such as education. His conclusion is that even if it were the case that effort would lead to increased moral worth, in situations where one could follow a less effortful path such as moral enhancement but instead chose the laborious route, this effort will not confer additional moral worth as it seems counter-intuitive that gratuitous effort exerted without any good reason could be considered worthy.[8] Therefore, the potential additional moral worth conferred by effort would have no bearing on the choice between an effortless moral enhancement and an effortful, conventional moral improvement. Nevertheless, I am convinced that it could still be argued that the creation of an effortless path would render all effortful and previously non-gratuitous paths gratuitous and hence less

[7] For example, a person who forms a detailed plan to become morally better and works hard to be better could acquire more value than if that person that simply happens to take a pill and instantly becomes a better person – even if she became that exact same person. This is not necessarily because effort is intrinsically valuable, but because of personal continuity. There are more intentions, deliberations and desires connecting the two stages of the person who changed through effort than there are connecting the two time-slices of the person who changed by taking a pill. Even though both persons become equally morally valuable in isolated terms, they do so from different paths, which affects their final value.

[8] If we deny such a claim then it would mean adding unnecessary effort to an action would always increase its moral worth, which seems absurd. See Douglas (2014c).

worthy; which would in turn decrease the overall moral worthiness of all available paths. For instance, prior to developing an effortless moral enhancement, we only have an effortful path towards moral improvement; as that effort is non-gratuitous it confers additional moral worthiness. After developing such a moral enhancement, both choices available would not have this additional moral worthiness: one because it was effortless and another because it was gratuitous. Although this would not be an argument for choosing the conventional route, it would be an argument for not developing the effortless moral enhancements in the first place. However, even if this argument were fully developed, I believe that it would lead to a much weaker objection to moral enhancement than the concern that moral enhancement would be intrinsically harmful overall due to undermining effort. The decrease in moral worthiness of both choices could be outweighed by the increased likelihood of generating morally better behaviour or motives. Additionally, there are cases in which the effortful path is not available or requires an extraordinary amount of collective effort – this might be the case with creating better large-scale co-operation.

I believe that a stronger concern exists when moral enhancement is carried out to such an extent as to be likely to entail an increase in moral status. In Chapter 3, I argue that in these cases enhancement could undermine psychological continuity, and thus personal identity, due to this approach having a higher potential for being abrupt than traditional means of self-improvement. In such cases, the value of a person at a certain time would be influenced by whether this same person existed in a previous time or not. In section 4.3 of Chapter 3, I explore how cross-temporal effects could be at play at the generational level. Although related to cross-temporal effects, these are not cases in which moral enhancement will not result in an improved morality, but cases in which moral enhancement might not result in (all) the desired social effects.

All criticisms based on moral enhancement being intrinsically harmful seem to fail. What is more, even if there were something conceptually wrong with moral enhancement, this might not prevent it from being attempted. Shook (2012) contends that these technologies will be developed and commercialised, regardless of all potential conceptual problems. The labelling, distribution and use of such drugs will depend more on supply, demand and marketing than on whether these drugs do perform moral enhancement or not. He envisions several catchy labels for pills for enhancing thoughtfulness, moral beliefs, intentions, willpower and sensitivity, which could be targeted at specific groups with different conceptions, and expectations, about morality. Irrespective of all conceptual discussion on moral enhancement, such drugs – or compounds that claim to be such drugs – will be developed and their impact on individuals and society will largely depend on factors such as social perception and marketing. This point has been central to research developed in parallel to this thesis, motivating the construction of models for the spread of potential moral enhancements (Sandberg & Fabiano, 2017).

3.3 Moral enhancement is a threat to freedom

A few authors, chiefly Harris, posit that human freedom is harmed by moral enhancements that prevent us from performing morally bad actions. The greatest source of concern is emotional manipulation that, once put in place, no longer requires any deliberation; in those cases, it seems a person would not choose the right action but rather be emotionally compelled towards it. As mentioned, Harris believes morality is about knowing and understanding the good, and that merely increasing the likelihood a person will perform a good action, regardless of her understanding, would not amount to improved morality. Rather, by removing the choice to do wrong it harms human freedom. Mirroring a theist's defence that argues that God allows the existence of evil for freedom's sake, Harris

(2011) contends we should be allowed to act wrongly to preserve our freedom: "Without the freedom to fall, good cannot be a choice; and freedom disappears and along with it virtue. There is no virtue in doing what you must." (p. 104).

There were two types of response to such a concern. One claims that moral enhancement does not necessarily decrease freedom. Savulescu & Persson (2012) argue that if freedom is compatible with our actions being determined, then moral enhancement determining our actions would not be a cause for concern. On the other hand, if freedom is not compatible with determinism, either we are already unfree, and moral enhancement could not reduce what we do not have, or we possess some indeterminist freedom, and moral enhancement could not limit this freedom because it would be inaccessible to any causal forces. Harris (2013b) has counter-argued that the fact that freedom is compatible with our actions being determined does not mean freedom is compatible with any form of determination. He insists his point is precisely that certain moral enhancements would determine our actions in ways detrimental to our freedom.

After presenting his analysis of what criteria determine an action to be free, DeGrazia (2014) notes that there is no reason to suppose moral enhancement would make one fail those criteria. Likewise, Douglas (2014b) argues that a person can undertake a moral enhancement that enables her to act according to what her true self considers right, helping her to overcome the influence of social pressure. Such enhancement would then actually increase her freedom.

Persson & Savulescu, Douglas and DeGrazia also present a second type of response, which claims that there are situations where the benefit of preventing evil would counterbalance the loss of decreasing freedom. For instance, it is clear that preventing a murder by restricting the murderer's freedom is preferable to allowing the full exercise of his freedom. While freedom matters, it is not all that matters. When the evil is great enough,

freedom might have to suffer. Preventing the ultimate harm would seem to be sufficient grounds for restricting freedom, if necessary.

3.4 Reasons to morally enhance are flawed

There is explicit disagreement with some of the claims used by Persson & Savulescu to support their thesis that moral enhancement is necessary to avoid global catastrophes. Persson & Savulescu ground the duty to pursue the development of moral enhancement on the contention that, given our present moral inadequacy, current scientific development would only further increase our capacity for self-destruction. In this scenario, cognitive enhancement could aggravate such risks given it would fuel scientific development. However, the idea that further scientific development – accelerated by cognitive enhancement – would have a negative net effect is contested. According to Harris (2013b) scientific development fuelled by cognitive enhancements will increase our ability to deal with almost any kind of global catastrophe, whereas moral enhancement, if successful, would only alleviate the risks of intentional global catastrophes. Hence, it would be irrational to foster moral enhancement while inhibiting cognitive enhancement. Moreover, Fenton (2010) believes Persson & Savulescu have underestimated the current benefits of scientific advances for populations living under the threat of diseases and poverty. Even if these advances increase the probability of self-destruction, this could be outweighed by the present increases in well-being that they bring about. Persson & Savulescu's (2011) response has been that they do not need to underestimate the benefits of current scientific development in order to conclude that further scientific development has a negative net effect; instead, their argument is that even small increases in the risk of a great catastrophe would suffice in order to outweigh present gains. Additionally, they observe that some cognitive biases could lead us to underestimate these risks.

The claim that it is vastly easier to cause harm than to cause good is another premise for moral enhancement, as it is used to advance the conclusion that the probability of global catastrophe is severe enough to justify radical technological interventions. Persson & Savulescu exemplify the claim by saying that any individual with a car in a populated area could kill dozens within minutes, while few people could save the same amount of lives in the same period. However, Harris (2013b) offers a couple of counter-examples where, provided the right conditions are in place, it seems equally easy to save many lives in minutes by simply donating large amounts of money to charities or preventing a terrorist from detonating a bomb in a plane. I am not convinced this issue could be settled merely by offering general examples in one direction or another given that the claim being argued/contested is if it is easier *in general* to cause harm than good. This question can only be properly addressed by means of a general argument. In that regard, Persson & Savulescu argue that intelligent life is extremely complex and improbable and hence there are more ways of destroying such a configuration than there is of enhancing it.

Buchanan & Powell (2018) concede that our present morality might be unfit for dealing with modern challenges. However, they argue against a strong version of the moral enhancement project that claims that biotechnological interventions are necessary in order to effectively overcome our moral unfitness. They reason that the assumption that our evolved morality is inflexible enough to frustrate traditional attempts towards moral progress is unfounded. I believe the defender of moral enhancement need not claim that traditional moral progress is unlikely, only that it can be helped by non-traditional technological interventions. Moreover, there is significant room for disagreement regarding how easily we can overcome obstacles to moral improvement without directly interfering with our evolved biology. Even one of the most notorious past instances of successful traditional moral progress mentioned by Buchanan & Powell (2018), the abolition of slavery, has not been

carried out without its shortcomings. I will return to an careful analysis of the reasons for moral enhancement in Chapter 4 as well as why dismissing moral enhancement in favour of strict traditional moral improvement might not be justified.

3.5 Moral enhancement will not produce the desired social and political effects

Many critics argue that while the problems moral enhancement proposes to solve are mainly political and social, moral enhancement focuses solely on the individual. To increase an individual predisposition to act pro-socially, or to emphasize or diminish her innate racial biases, would do little to solve problems that are structural, or deeply rooted in our institutions, economic policies, political ideologies and culture (De Melo-Martin & Salles, 2014). Murphy (2014) mentions the fact that a great many people already oppose nuclear weapons to no effect on international disarmament agreements. Sparrow (2013) observes that from a certain perspective a suicide bombing can be considered an ultimate act of altruism. Therefore, correcting individual moral deficiencies would be attempting to solve the wrong problem. Harris (2013b) states: "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. It is clear we cannot provide health care 'free at the point of need' by private altruism" (p. 290). This social objection to moral enhancement will be duly taken into account during section 3.2 of Chapter 2. I will attempt to analyse whether increasing individual tendencies towards co-operation and altruism would entail increases in group/social overall tendencies towards co-operation and altruism. It seems clear the critics are right in pointing out that the problems that could lead to Persson & Savulescu's ultimate harm – e.g. nuclear war or global warming – all arise from social interactions between groups and that targeting individuals might not be sufficient. However,

positive modifications of the social tendencies of individuals could also have positive effects on the group level.

A further set of issues is raised by authors such as Agar (2013b) and Sparrow (2014), who believe that if we were effective in producing individuals with higher moral capacities, the inequality created by this scenario could outweigh its benefits. Agar – who believes we could have moral status enhancement through cognitive enhancement – asserts that moral status enhancement will create a future with two different classes of persons: the mere human persons who exist today, and a class of supra-persons whose higher capacity for engaging in moral reasoning will confer on them a higher moral priority over mere persons. In this setting, it will often be morally desirable to sacrifice mere persons to benefit supra-persons. Agar considers *the creation* of a scenario wherein such sacrifice is desirable – and not the sacrifice in itself – to have an extreme negative moral value, and thus concludes increasing moral status to be undesirable. The claim that supra-persons will possess higher moral status draws support from the fact we already have a hierarchy of moral status that seems largely dependent on moral reasoning, going from inanimate objects, to complex animals, to human beings. Others such as Buchanan (2009) believe that our higher moral status is a threshold concept, such as that all other entities possessing the necessary characteristics for personhood, to whatever extent, should be equally regarded. Douglas (2013a) argues that even if we reject the idea that supra-persons would have higher moral status, it is still plausible that the creation of supra-persons would harm mere persons who remain unenhanced. Nevertheless, he reasons that this does not actually render the creation of supra-persons undesirable unless, among other things, we adopt unjustifiable partiality to the unenhanced or ignore the gains of becoming a supra-person. Persson (2012) argues that if supra-persons' harm to persons were permissible, it would not be permissible to prevent their existence. However, Agar believes that any harm that can eliminate the set of experiences

specific to *Homo sapiens* should be prevented, even if this harm is permissible otherwise. I will develop a more detailed analysis of the risks arising from the creation of supra-persons in Chapter 3.

In contrast, Sparrow (2014) is primarily concerned with the political inequality created by the fact that the morally enhanced would have to have disproportionate political representation if moral enhancement is to be effective at all. Persson & Savulescu counter-argue that we already live in a situation wherein we trust political decisions to people believed to possess higher moral capacities. Marshall (2014) observes that moral enhancement only entails increased conformity with morality, not greater knowledge of the common good, and hence would not be grounds for higher political representation. On the other hand, Morioka (2014) and Shook (2010) note that the morally enhanced could be easier targets for domination due to decreased aggression and increased tendency to co-operate.

Wasserman (2014) draws attention to the fact that a universally morally enhanced society might not be functional. Many vital social roles can only be filled by those with the predispositions some universally deployed moral enhancements intend to eradicate. Police officers, intelligence services, surgeons and rigid leaders all possess traits that in other professions could be considered moral failings – such as emotional detachment, aggressiveness, paranoia and narcissism – but that enable them to perform their vital roles in society.

That moral enhancement might not generate the desired social effets will be the basis of my own arguments against moral enhancement in the next two chapters. I believe this is the strongest set of area of criticism against moral enhancement. In Chapter 5, I will argue that using virtue theory to guide moral enhancement can help address these concerns.

5. Conclusion

There is a large body of academic papers on moral enhancement, which has proven to be fertile grounds for practical ethical reflections. A critical review of the literature reveals that the common arguments against the project of moral enhancement can be divided into five groups. However, there is a lack of a unified conceptual framework and robust risk assessment of the project. Several critiques do not stand scrutiny in light of the original proposals advanced for moral enhancement, and others leave an increasingly large and disparate set of questions open. The rest of this thesis will aim to build a stronger criticism of moral enhancement and to resolve some of these questions by showing how moral enhancement can be pursued in a way that takes these criticisms into account.

Some of the critiques are based on pessimism over the empirical feasibility of moral enhancement. But the original argument made by Persson & Savulescu does not require assuming it is certainly feasible, only that it is possibly feasible; arguing it is entirely unfeasible seems highly implausible. Despite pessimism about empirical feasibility being a defensible position, it is a position that when assumed can lead to ignoring potential risks if the project turns out to be feasible. A lack of proper risk assessment is especially concerning given that we already make use of technological interventions that have a (supposedly positive) impact on moral behaviour.[9] Arguments against moral enhancement that claim that it will not lead to an improved morality seem to be based on the claim that there is something conceptually wrong with moral enhancement. Attempts to argue that the emotional manipulation proposed in specific forms of moral enhancement cannot lead to an improved morality seem to require bolder assumptions than assuming otherwise. What is more, moral enhancement need not involve emotional manipulation alone. Justifying the belief that there

[9] Antidepressants, ADHD medication and anxiolytics all have an impact on moral reasoning and moral behaviour (Levy et al., 2014).

is something conceptually wrong with moral enhancement requires finding something intrinsically wrong with the concept itself. Nothing has been found in this critical literature review, despite my attempts to develop some past efforts. Section 3.3. of Chapter 2 will make the case that there are intrinsic risks of forms of moral enhancement that target motivation, but these risks can be avoided, and moral enhancement need not focus on motivation.

Freedom and related concepts such as autonomy or liberty are central human values. Perhaps some strong objection can be made against moral enhancement based on its being a threat to freedom. However, the extensive literature already written on the topic concludes the danger is not substantial. At the end of Chapter 5, I will offer an example of a type of potential moral enhancement aimed at enhancing liberty as a virtue, which would not be a threat to freedom and indeed probably remove real obstacles to it.

A proper analysis of the reasons to pursue moral enhancement will be carried out in Chapter 4. There I will detail which are the sorts of extreme risks that moral enhancement can alleviate, and why traditional means of moral improvement should be used in conjunction with moral enhancement and not to the exclusion of it. Moreover, the concern that the reasons to morally enhance are flawed will also be taken into account throughout this thesis by choosing to focus on forms of moral enhancement likely to produce changes that would, in fact, give us strong reasons to justify the project; that is, moral enhancement that would significantly decrease the probability of extreme risks. Furthermore, this choice will shift the focus from the open empirical question of whether moral enhancement is feasible to the ethical question of whether it could lead to undesirable outcomes if targeted at the problems it hopes to solve. Simultaneously, this will help alleviate concerns that moral enhancement would involve only superficial emotional manipulations and would fail to result in an improved morality. I will specifically focus on what I will define in the next chapter as *deep moral enhancement*, technological interventions targeted at human traits

expected to lead to better moral behaviour or motives. I will then show that there are valid concerns about deep moral enhancement by developing further the argument that moral enhancement might not lead to the desired social effects; both Harris' private altruism critique and matters related to the creation of supra-persons. Finally, if indeed there are some still unknown reasons for moral enhancement being conceptually flawed, this focus could help reveal in which ways such misconceptions could materialise in the form of real risks.[10]

Moral enhancement needs to be taken seriously in its aims, and its risks adequately investigated. If one follows the original proposals for moral enhancement found in Persson & Savulescu and Douglas closely, then moral enhancement is hard to oppose. In the next two chapters, I will show that strong and rigorous critiques can still be made by developing some of the arguments mentioned above against moral enhancement. In the next Chapter, I will argue that if we take the bold aims of moral enhancement seriously (Persson & Savulescu's goal to decrease extreme risks for humanity in particular), moral enhancement can be catastrophic if done improperly because it might produce unexpected effects due to the complexity and fragility of the human traits that are likely targets of moral enhancement. Many of those effects are undesired social consequences of the kind mentioned in this Chapter. I will also argue that moral enhancement is likely to target motivation and consequently be intrinsically prone to be self-reinforcing and irreversible. In Chapter 3, after restating in more detail how many of the concerns about the creation of supra-persons are unfounded, I will argue that creating supra-persons presents risks which are related to the detrimental effects of deep moral enhancement on personal identity.[11]

[10] The discussion made in section 3.1 of Chapter 5 can been seen as an attempt to make moral enhancement safe even in the face of the possibility of conceptual mistakes.

[11] As I will focus on psychological continuity, the arguments will be cogent even for those who subscribe to the view that personal identity is not what matters. See paragraph 6 of section 3 of Chapter 3.

Chapter 2: Complexity and fragility of moral traits

1. Introduction

1.1 Deep moral enhancement

One simple line of reasoning in favour of moral enhancement proceeds as follows. Altruism, generosity, co-operation and non-aggressiveness are all human traits that most people would agree we should increase. If some technological intervention such as moral enhancement[12] could easily strengthen those traits, then we should implement the intervention. The thesis I will introduce here will challenge this by arguing that it is not the case that increases in currently morally desirable traits would be themselves desirable. More generally, I will argue that moral enhancement directly targeted at moral traits, which I will call *deep moral enhancement*, is particularly risky. Improving altruism or co-operation through traditional means of moral development such as education is undoubtedly desirable, and there might be no substantial difference between deep moral enhancement and the effects of education. Nonetheless, given that by using deep moral enhancement we could bring about changes in our levels of altruism and co-operation unachievable through conventional means – if we could not, then there is no potential advantage in trying to develop it in the first place[13] – I will argue such deep changes could actually bring about moral decay.[14] The limitations of traditional moral education create a safe boundary, which

[12] As in Chapter 1, here I follow Douglas in defining moral enhancement as any intervention that is expected to lead to morally better behaviour or motives.

[13] One might argue that moral enhancement could be cheaper and faster than moral education, thus that it should be preferred even if it produces the same results. However, I believe it is clear that the development costs and implementation risks of such technologies would outweigh these other factors. It would be hard to argue for the development of a technology that might or might not be feasible, and might or might not be risky, just to achieve the exact same results we already have with moral education, only because it might or might not achieve those results faster.

[14] There is a clear difference in degree regarding deep moral enhancement and conventional moral education. Persson & Savulescu and others have pointed out that this difference does not *need to* entail a substantial

deep moral enhancement will be likely to breach. In this Chapter, I will mainly focus on developing the thesis that our moral traits – the target of deep moral enhancement – are complex and fragile and that attempts to enhance them have a relatively high amount of risk. With this framework, I will then spend my final Chapter analysing the possibility of safe paths for deep moral enhancement.

Before introducing this Chapter's thesis and its background, let me briefly mention one example where seemingly straightforward enhancements of desirable things can lead to undesirable outcomes. Take the case of the artificial stimulation of the brain to produce pleasure (a procedure known as *wireheading*). Although pleasurable experiences are highly valuable and most can agree that increasing overall pleasure would be morally desirable, fundamentally changing our pleasure levels by using a brain implant to activate regions associated with it artificially, such that someone would live in a state of constant bliss regardless of all other aspects of life, will arguably lead to a life similar in certain respects to that of a heroin addict. Real life cases confirm that reasoning. Individuals who for medical reasons find themselves with the control of a brain electrode connected to regions associated with pleasure will compulsively self-stimulate and rapidly develop apathy towards everything else (Portenoy et al., 1986). Likewise, a civilisation of individuals constantly experiencing bliss while being fed and maintained by machines would have nearly maximum amounts of happiness, but probably no art, love, scientific discovery, or any of the other things humans find valuable. While pleasure is optimised, all other morally significant aspects are set to undesirable states. Many other paradoxes in moral philosophy would seem to be derived from this same more general problem.[15] The complexity of deeply enhancing

difference between the two and have argued for a blurred line between them. However, it could also be the case that differences in degree eventually do entail substantial differences, which I believe is the case with deep moral enhancement.

[15] According to this view, defended by Alan Carter (2011), the repugnant conclusion is the result of solely optimizing the total amount of happiness while dismissing any other value; it produces not just a scenario that

morally significant things seems to indicate that partial enhancements can lead to not only partial but also undesirable consequences. In wireheading, one gets undesirable outcomes from increasing just one morally significant, and otherwise desirable, dimension (sentient pleasure).

I will not defend the idea that every form of moral enhancement is undermined by the complexity claim. Only *deep moral enhancement* is. Deep moral enhancement will be defined as follows:

> *(1) Individual definition:* Significant changes, brought about via technological interventions, directly targeted at human traits (e.g. co-operativeness, empathy, altruism, etc.) primarily expected to lead to morally better behaviour or motives; **or**
>
> *(2) Societal definition:* Changes, brought about via technological interventions, in the normal human variation of these human traits primarily expected to lead to morally better behaviour or motives, even if brought about by small changes in the traits of individuals.

By human traits I mean general and stable patterns of behaviour or cognition such as empathy, honesty, aggression or extraversion.[16] Let us consider a few examples to clarify the rest of my definition. Regarding the individual definition, a drug working primarily by erasing racial biases in the decision making of a single judge during racially sensitive cases (Douglas, 2013b) would not be a deep moral enhancement. Although it would be a significant change leading to morally better behaviour, it would not significantly alter a

---

is only partially valuable but one with disvalue. The utility monster scenario is the result of solely optimizing average happiness; this scenario would be otherwise ideal if it were the case that there were no morally relevant variable other than average happiness. More importantly for my argument, if it were the case that optimizing only one value would not come at the expense of other values, both scenarios should be at least partially desirable – instead of undesirable.

[16] I take it as an intuitive matter that such patterns exist. Nonetheless, philosophical arguments made against their existence will be addressed in Chapter 5, section 3.2.

human trait due to its narrow scope.[17] Meanwhile, a drug that would significantly decrease this judge's racial biases across all domains, making – ceteris paribus – racial consideration in itself irrelevant, would count as a deep moral enhancement on account of (1). Regarding the societal definition, a drug that would only modestly decrease in-group favouritism but that is given to a large share of the human population would count as a deep moral enhancement. Most current forms of moral education are not a deep moral enhancement primarily because they do not change traits, although they can be administered to a large share of the human population. Moral education might increase the likelihood that individuals will be more altruistic or co-operative in given circumstances, such as offering their seat to an elderly lady on the train or donating to alleviating the suffering of the poor.[18] However, it would not immediately increase the likelihood that an individual would co-operate across all social interactions; the fundamental setting of human co-operation would remain relatively untouched. Moral education might decrease or inhibit racism, but if there is a biological tendency towards implicit racial biases (Ito et al., 2015), then we could only eliminate it – instead of merely frowning on or punishing it – by changing an innate biological predisposition. Both (1) and (2) exclude other types of enhancement that would lead to morally better behaviour or motives indirectly (as a secondary effect); causing better moral decision making by increasing short-term memory would not be deep moral enhancement. Environmental changes that cause morally better behaviour without changing any traits are also not deep moral enhancement.[19] The type of moral enhancement that

[17] Such a drug could work by merely preventing his brain from processing someone's skin colour during a trial. His traits and overall propensity to discriminate in other contexts would have remained unchanged.

[18] More efficient and widespread forms of moral education might produce changes in traits. If carried out across generations, moral education can even act as a selective pressure that modifies traits. However, direct biochemical interventions should more easily achieve the same result. I address these questions and the argument that traditional moral education should be strictly preferred over technological interventions in section 4.2 of Chapter 4.

[19] It might be the case that radical moral education or persistent and drastic environmental changes will cause significant changes in traits primarily expected to lead to morally better behaviour or motives, but for simplicity

Persson & Savulescu advocate would clearly satisfy condition (2), and quite possibly also (1). It would have to be widespread across society, and it would fundamentally change the way we co-operate. It would not only make it less likely for nations to engage in a nuclear arms race or more likely to care about the common pool of resources, but it would have to work across a wide range of social interactions. I will be focusing mainly on the societal definition of deep moral enhancement. It might be that a drug that would increase a single individual's likelihood to co-operate by 10% would produce very few consequences. In fact, the propensity to co-operate naturally fluctuates and such an individual might naturally exist within normal variation. Nonetheless, if a significant portion of the human population uses this drug, then the range of normal human variation would have been breached.

My analysis will be novel in that I will not suppose moral enhancement is unfeasible. I will assume we can, eventually, develop technologies that will fundamentally change human morality. I will argue that the risk lies specifically in the deep changes that moral enhancement could bring about. This charitable assumption is by no means shared among the opponents of moral enhancement. For example, Sparrow (2013) believes "we don't have any such technology and we are unlikely to develop it within the time available to us"(p. 407). Harris (2011) asserts that certain types of technologies could, in principle, improve human morality, but he contends that the necessary fine-tuning is unlikely to be achieved: "I for one am sceptical that we would ever have available an intervention capable of targeting aversions to the wicked rather than the good" (p. 105). Due to his concerns that moral enhancement will unbalance the delicate equilibrium present on our current morality, Agar (2013) states: "[It is] unlikely to be any pills or injections that directly produce in us morally superior judgments or motivations" (p. 1). Given many of the conceptual issues that would

---

I will be mainly focusing my arguments on cases of technological innovations directly targeted at traits primarily expected to lead to morally better behaviour or motives.

need to be settled before properly pursuing moral enhancement, Beck (2014) concludes "moral enhancement is not very likely to be made sense of – let alone realised – in the medium-term future" (p. 2). However, the technical scepticism towards moral enhancement that pervades most of these critiques is rarely properly and explicitly justified by these authors. I am convinced that the assumption that we will simply not develop deep moral enhancement is controversial at best,[20] and, more importantly, it makes us unable to assess important risks if it is indeed possible to modify human morality to such an extent.

1.2 Moral normalcy and complexity

Nick Agar presents a similar (but weaker) argument. As mentioned in Chapter 1, Agar believes that moral enhancement can destabilise the delicate balance of human morality. The balance consists of an equilibrium that he calls moral normalcy, existing between both rational cognitions and emotional intuitions (Agar, 2015). He also claims such moral normalcy "…emerges from the interplay among normal moral motivation, normal moral insight and normal moral behavioural capacities" (Agar, 2014a, p. 369). Having a normal human moral capacity depends on the balance and input of other moral faculties, which are, supposedly, set at very specific configurations, which if tampered with would lead to catastrophic imbalance. Agar argues that moral normalcy, although far from perfect, is the condition under which we test our moral theories and our various conceptions of what is good; it is a reference point and thus should be preferred over other configurations. He then asserts that "Moral improvements are unlikely to result from biomedical interventions in people who achieve moral normalcy" (Agar, 2013, p.1).

I will argue for a similar but stronger version of Agar's moral normalcy thesis.[21] This

[20] For instance, because individual neurochemical manipulations ought to have societal effects and because they are nomologically possible (Persson & Savulescu, 2014a, p. 42; and also Douglas, 2014b, p. 11).
[21] That means, among other things, there are likely to be enhancements that the moral normalcy claim would consider safe and desirable but that the complexity claim would consider risky and undesirable.

enriched thesis will contend that human morality not only stands on a fragile balance between the few factors Agar mentions, but also that there is an overabundance of factors contributing to this fragile balance. I will defend the *complexity and fragility of moral traits* thesis: (1) moral traits are extremely complex and the process of understanding them faces major epistemic difficulties – this can be called the *complexity claim*. In this thesis, I will also argue that (2) deep moral enhancement could have major unexpected consequences, risking severe catastrophes – this can be called the *fragility claim*. By *moral traits* I will mean all fundamental human traits that are primarily involved with human morality in a descriptive sense and, more importantly, that would reasonably constitute a target of deep moral enhancement.[22] In other words, moral traits will thus be general and stable patterns of behaviour or cognition primarily involved with human morality and that would reasonably constitute a target of deep moral enhancement. But before further detailing my two claims, we should revisit the following example. Perhaps a seemingly simple moral enhancement would be to increase one's moral disposition towards increasing the pleasure of others. Nonetheless, this might make one seek artificial means to induce pleasure, such as the aforementioned direct brain stimulation or the use of drugs, disregarding other potentially important dimensions such as truthfulness and variety of experiences. The disposition towards increasing others' pleasure is likely to be a worthwhile human trait that we both have and should have; however, excessively and fundamentally enhancing this disposition will lead to results that are not only incomplete but also detrimental.

[22] By descriptive sense I mean in the sense of being a description of human morality as an empirical matter (e.g. moral psychology) and not in the normative ethics sense. There is a multiplicity of terms used in the moral enhancement literature such as moral dispositions, moral traits, moral behaviour, moral attitudes, moral reasoning and human morality. Human morality is sometimes used to refer to the collection of these terms. However, it is also used to refer to a set of norms accepted by a society or to normative ethics. Given I wish to refer to individual human features that could be targeted by technological intervention, these two latter meanings are unfortunate. Moral traits will refer to the collection of general and stable human dispositions *in so far as they are reasonable targets of deep moral enhancement.* The relationship between moral traits, moral character and virtue will be explored in the last chapter of this thesis.

2. Complexity claim

2.1 Introduction

The complexity claim states that moral traits are likely to be relatively more complex than most other targets of human enhancement such as human physiology, short and long-term memory and mood. I will present as support the intricate causal history of moral traits, its relatively high frequency of contingencies, that moral traits are widely distributed across the brain, and epistemic difficulties arising from the first-person view of human morality. By complexity I mean descriptive complexity, the length of the shortest possible description. This definition can be formalised, but I will be using it in an intuitive sense here.

2.2 Complex aetiology

This complexity claim draws support from looking at the causal history of how our moral traits came to be. We should expect our moral traits, taken as a whole, to have a fairly high complexity given that they are the product of many contingent and accidental events with random processes involved. If we assume our moral traits were (at least partially) shaped by our natural history and human history, then it is easy to see how those two histories were populated with contingencies that happen for no good reason and that have shaped our moral traits. Firstly, one of the major processes influencing natural history is natural selection, which is often characterised as a bricolage that uses pre-existing traits with unrelated functions to produce new traits that are sub-optimal – merely good enough to survive and reproduce – under certain accidental evolutionary pressures, using random mutation as its source material. Because of this reliance on pre-existing traits, there is a great deal of influence of past evolutionary pressure on present design, a phenomenon known as evolutionary hangover. For instance, nearly half a billion years ago Earth conditions were radically different (e.g. the atmosphere had less than half of current oxygen levels) and

selective pressures at that time helped shape the current body template for all vertebrate animals. Our current body architecture would be radically different if those conditions half a billion years ago had been different. In the same manner, had our hominid ancestors not been driven to live in small, geographically isolated, co-operative communities, then it might be that our intuitions about the permissibility of not alleviating the suffering of humans who are in underdeveloped faraway countries would be different.[23] Our current body architecture or moral intuitions are packed with features whose causal explanation contain facts about the Earth's atmosphere half a billion years ago or density and spread of human population in the Palaeolithic. Another example, one of the explanations of why we seem to have different modes of moral reasoning that manifest at different situations and ages is that each one of those modes evolved at different times, under different pressures, of our evolutionary history. The old cognitive processes, rather than being deleted in favour of new ones, formed a complex base whereupon a new process would be built resulting in a kludge of cognitive strategies (Krebs, 2015; and also the entire volume on the subject Krebs, 2011).

Secondly, natural history is also largely dictated by sudden non-selective random processes such as population bottlenecks, mass extinctions, and founder effects. As an example of a founder effect, when a small subset of a population migrates to a previously uninhabited area, this new settlement's genetic diversity will be largely capped and unrepresentative of the original population. Whichever subset came to migrate will dramatically shape this new population gene pool. For another example, whether or not a

[23] If the groups had been even more geographically isolated by geographical accidents – ceteris paribus – we would be more strongly inclined to think it is permissible to ignore the suffering of those far away because our minds would have adapted to live in an environment where we could hardly affect those people far away and thus would have adapted to live in an environment where people far away simply did not matter. On the other hand, if the groups had been less isolated by geographical accidents – ceteris paribus – we would be more strongly inclined to think it is impermissible to ignore the suffering of those far away, for analogous reasons. Admittedly, the influence of those innate intuitions over extensive ethical reflection might not be so straightforward but my goal here is to analyse moral traits, not ethical theories.

particular hominid population was hit by a hurricane will dictate whether or not a particular brain structure, with a particular way of reasoning about morality, will continue to exist.[24] It should be noted that these kinds of extinction events and founder effects would not influence traits that are relatively homogenous in the initial population. Oxygen-transporting for instance, which can be easily enhanced by the use of doping, would not be affected.

Thirdly, one central aspect of the study of human history is the observation of contingencies: single events that led to a particular series of outcomes, which would not have occurred in the absence of that event (Andrews & Burke, 2007). The very idea of important historical events means that these events had an important causal role in history, i.e. history would have been different if they had not occurred. The spread of a particular cultural belief is the result of a complex series of historically contingent events. Moral traits are influenced by cultural beliefs, which adds to their complexity.

Therefore, given that the causal history of the way we currently reason about morality was populated with numerous contingent, complex and unrelated events, then it is to be expected that moral traits have, as a whole, a high level of complexity. It could be argued that causal history would have no influence on the final complexity of moral traits. However, this fails to account for the fact that a causal history is constructed as the most probable explanation of an event (Lewis, 1986). If an event is simple and probable enough that it could have been brought about by a very simple process, then its causal history will be simple. The fact we need to evoke complex processes to explain moral traits indicates they have high complexity.[25] Furthermore, past contingencies add to its complexity. Every time

[24] For instance, Japan escaped being under the Mongol Empire's sphere of cultural influence due to a typhoon frustrating a Mongol invasion that would otherwise have been the second greatest naval invasion in human history (Turnbull, 2010). This extraneous weather phenomenon had a defining impact on the history of the Japanese nation and its norms, and although it had predominantly cultural implications, it also had an impact in maintaining a modest genetic difference between the people of Japan and of the neighbouring mainland.

[25] Of course, it might be that we need not and that evolutionary explanations are completely unnecessary. My arguments rely on the assumption that we need evolutionary theory, even if sparingly, to explain human morality.

a contingent effect happens, the complexity of the end product is necessarily raised as it contains aspects that cannot be attributed to the general process that generated it – if they could, there would be no reason to assume a contingency in the first place. When we say that such and such in the Cambrian period had a causal role in our current vertebrate body template, it means that without assuming such and such happened in the Cambrian, it would be hard to explain the current vertebrate body template. It seems to be the case that the list of contingencies we have to evoke to explain moral traits is higher than to explain vertebrate body template. Causal histories will be revisited when defending the fragility claim, in section 3.4. Moreover, the fact that human history and natural history are two relatively independent processes also means the final complexity should be higher. Absence of correlation between processes leads to higher complexity; essentially because if there is a correlation it means there are shorter – less complex – descriptions of those processes.

On the extreme of triviality, much of what was said above could be said even of a rock on the beach. However, the degree of complexity found in the sedimentary process forming a rock on the beach is much lower than the degree of complexity of the processes that generated complex life. Supporting the claim that a complex aetiology leads to higher final complexity, the chemical organisation of a rock is far simpler – i.e. has a shorter description – than most life forms. To a minor degree, we could trivialise my arguments by noting that human physiology was also the result of random and contingent evolutionary processes and thus should be equally complex. However, one can observe that most of the basic aspects of our bodies remained relatively stable across human evolution. Even if we compare our bodies with those of our close phylogenetic relatives, there is very little variation. This stability is so extreme that we can successfully transplant animal organs into human beings; whereas we have never made any successful attempt of brain regions transplants and are unlikely to do so in the foreseeable future. In comparison, our moral traits

are extremely varied and have undergone drastic changes. For instance, while all human societies across time share the same oxytocin receptor, the norms concerning human mating vary from fostering/forbidding polygamy, to polyandry, to monogamy, to promiscuity. The degree to which moral traits were subjected to complex, contingent and random processes is higher than for other human traits.

2.3 Epistemic complexity

Besides having aetiological reasons to suspect that moral traits are extremely complex, we also face epistemic difficulties when trying to access this complexity. There is a natural tendency to search solely for simple, efficient and elegant causal explanations, making it hard to emulate the random and chaotic process involved in the creation of our moral traits correctly. Moreover, several parts of our moral traits that were shaped by evolution are not adapted to our environment any more; thus it is hard to understand their current dynamics and effects. Some of our desires are evolutionary adaptations for reproduction and survival; in that regard, they merely provide motivation for us to achieve sub-goals that will contribute to the ultimate goal of increasing survival and reproduction. Nonetheless, we often experience them as decontextualized emotions; so inferring their hierarchical and functional organisation correctly is challenging. As a result, we cannot distinguish between desires that are ends and desires that are means in the first-person view, making it hard to single out the most important and fundamental desires from the accidental ones. Furthermore, we are the only moral agent we know of, so we cannot assess how improbable our morality is within the context of all logically possible moralities. By assuming our present morality is somehow universal or a convergent point, we are likely to overestimate its robustness. These reasons above all indicate that moral traits are extremely complex, in the sense that our mind cannot easily draw general principles that would account

for all of them, i.e. they cannot be easily captured in a reasonably short description (a similar point has been made with other terminology by Ross, 1930).

3. Fragility claim

3.1 Introduction

The fragility claim states that our moral traits are fragile under deep moral enhancement. Fragility will be understood as proclivity to unexpected disturbances brought about by a change. We can define *fragility* as a positional measure of unexpected counterfactual variance under a certain modification:

> If modification M's actual outcome is farther from the reasonably expected outcome when performed on trait A than when performed on trait B, then trait A is more fragile than trait B with regards to M.

Thus, the fragility claim is: the actual outcome of performing a modification is farther from the reasonably expected outcome when performed on moral traits than when performed on physical or cognitive traits. One can specify this further using possible worlds. The definition can be translated to: A is more fragile than B in regards to action M, if (and to the extent to which) the world one reasonably expected to result from performing M on A is further from the world one reasonably expected to result with M on A than the world resulting from performing M on B is further from the world one reasonably expected to cause with M on B. Thus, the fragility claim translates to: possible worlds where we perform deep moral enhancement are (on average) more distant from the world we reasonably expect to cause with deep moral enhancement than worlds where we perform, for example, memory or physical enhancement are from the world we reasonably expect to cause with them. The particular setting of moral traits is such that all the possible deep changes to it would entail an unusually high amount of unexpected variance. If moral traits were not fragile, we could

perform deep moral enhancement and there would be little change other than the change that we expected to bring about. Note that this is not a measure of expected unintended consequences; it does not measure the extent of side-effects that we know will happen. It measures the scope of unexpected consequences, of the possible effects we currently do not know will happen. One may not know specifically how a porcelain dish given to a hyperactive toddler will break, but one knows there are more possibilities for it to break than there are for a stuffed animal. Just as one can attempt to list and prevent some of the possible ways the dish will break, a careful investigation of the risks of deep moral enhancement might lead to their mitigation.

I will make the case for the fragility of human moral traits along three lines. Firstly, I will present an instance where a *prima facie* simple and safe path towards moral enhancement, individual co-operational enhancement, could go wrong because it would have big, *prima facie* unexpected consequences – particularly, consequences in the opposite direction to that expected in the literature. I will then generalise this case based on emergent properties and argue that there are likely to be more unexpected consequences. Secondly, I will argue that general features of moral traits make any fundamental changes to them likely to be self-reinforcing and irreversible; thus, small but deep changes to moral traits could have significant unplanned results. Thirdly, I will contend that the aetiological argument given for the complexity claim also supports the fragility claim.

3.2 Fragility of human co-operation

3.2.1 Parochialism

Given that co-operational enhancements have been deemed both morally desirable and scientifically feasible,[26] one should expect that they will be developed eventually. Nevertheless, scientific research on understanding and increasing co-operation has largely focused on an individual level, though it is the group level that is problematic (Ostrom, 1994; and also Greene, 2013) and in need of enhancement (Persson & Savulescu, 2013). Although one might perhaps expect that an increase in our *individual tendency* towards co-operation *between individuals* would entail increased co-operation *between groups*, it should be made clear that the latter is most desirable. Co-operation between countries is more crucial than co-operation between citizens, and the same is valid for ethnicities, political orientations and cultures. The problem deep moral enhancement hopes to resolve is large-scale co-operation between nations, cultures or ethnic groups.[27] Hence, the question is: are increases in moral dispositions conducive to co-operation *between individuals* guaranteed to promote co-operation *between groups*? I argue that the answer is no. There is a large body of empirical evidence showing that co-operation between individuals does not lead to co-operation between groups and sometimes will actually cause increased competition. Additionally, I will present theoretical reasons to expect that co-operation between individuals is tightly coupled with competition between groups.

The fact that co-operative tendencies are mostly limited to in-groups is known in the scientific literature as parochialism or in-group favouritism. Parochialism is the tendency to prefer to co-operate with members of your own group over out-groups, sometimes even if

[26] For brevity the long list of recent studies on this area was omitted. For further information, refer to the bibliography in Crockett & Fehr (2013).
[27] This may often be in the form of co-operation between individuals of these different groups.

this comes at the expense of harming out-groups (for a meta-analysis Balliet, Wu & De Dreu, 2014; for a conceptual overview Hewstone, Rubin, & Willis, 2002). This tendency is generally proportional to how co-operative individuals are inside their own groups; that is, groups that are highly co-operative internally will tend to be the least co-operative with other groups (Bornstein, 2003; De Dreu, 2013). The relationship also holds in the opposite direction: competition between groups leads to increased contribution to the public good within-group and to increased group effectiveness (Cardenas & Mantilla, 2015; Puurtinen & Mappes, 2009; Burton-Chellew, Ross-Gillespie, & West, 2010; Bornstein, Winter & Goren, 1996). Men tend to exhibit higher levels of parochialism, co-operating more than women inside their group, but also have higher proclivity to conflict with out-groups (Sidanius & Veniegas, 2000). Besides direct empirical evidence, there is a cogent case for a strong link between individual co-operativeness and group conflict when we note how the policies and practices that promote co-operation inside a group can easily – unexpectedly or not – lead to a conflict between groups. In an overview of the literature on solidarity mechanisms, the psychologist Gary Bornstein (2013) observes:

> "Collective group goals and common group identity are emphasized, norms of group-based altruism or patriotism are fortified, punishment and rejection of defectors are increased, and the shared perception of the out-group is manipulated (Campbell, 1965; Pruitt & Rubin, 1986; Sherif, 1966; Simmel, 1955; Stein, 1976). Whereas the foremost function of these structural and motivational processes is to facilitate co-operation within the groups, they inevitably contribute to the escalation of the conflict between them" (p. 130)

There is also a theoretical reason to expect that between-individuals co-operation is tightly connected to group conflict. Many theories have been proposed to explain why non-kin co-operation evolved, and several of them establish that this type of co-operation could only become evolutionarily stable if it had coevolved with aggression towards out-groups. For example, Bowles & Gintis (2011) attempted to model the evolution of co-operation using our best estimates regarding group-size and food-sharing during the Palaeolithic. Even

when using the most unfavourable estimates to this conclusion, their results show that parochialism and co-operation could only have evolved together.[28] Therefore, it is reasonable to assume that the brain networks responsible for these traits are strongly interconnected and that a moral enhancement that targeted only one of them would be hard to develop. As further evidence of this assumption, oxytocin – one of the drugs cited as preliminary evidence that we could one day develop a moral enhancement – seems to increase altruism, co-operation and generosity (De Dreu & Kret, 2015), but it is also known to produce in-group favouritism, leading to ethnocentrism and parochialism (De Dreu, Greer, Van Kleef, Shalvi, & Handgraaf, 2011; De Dreu, 2012). So far, there is little hope for uncoupling its effects on individual co-operation from the in-group favouritism side-effect.

3.2.2. Problems with private altruism

It should be noted that the issues I have presented in section 3 so far reflect a common objection to moral enhancement. As mentioned in section 3.5 of Chapter 1, critics argue that moral enhancement seems to target the individual but the problems it wants to solve are mainly political and social. Increasing individuals' tendencies to co-operate, to emphasize or to correct innate racial biases will not directly address problems that are structural, or deeply rooted in our institutions, economic policies, political ideologies and culture. This can be summarised by this criticism of private altruism from Harris (2013a):

> "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. It is clear we cannot provide health care 'free at the point of need' by private altruism" (p. 290).

It seems clear the critics are right in pointing out that the problems that could lead to Persson & Savulescu's ultimate harm – e.g. nuclear war or global warming – all arise from social

[28] Their model revealed that: (1) groups with non-parochial co-operators have a disadvantage over other groups and thus would not have evolved in the first place, however; (2) groups with parochial co-operators, that are willing to sacrifice themselves fighting against out-groups in order to benefit their peers, have an evolutionary advantage and; finally, (3) merely parochial groups have a general disadvantage.

interactions between groups and that targeting individuals might not be sufficient. However, none of them have presented counter-arguments against the intuitive assumption that positive modifications on an individual's social tendencies would also have positive effects on the group level. Besides supporting the fragility claim, the two last sections have also presented those necessary counter-arguments. It is plausible that the case of human co-operation is just one example of several paradoxical emergent effects credibly to be found in deep moral enhancement. In many current economical and sociological theories, human society is a highly complex system whose organisation is partially (or primarily) determined by individual patterns of behaviour, changes in which can affect the system in unexpected ways and which may plausibly be altered by technologically intervening with moral traits.

Another case of a potential paradoxical effect would be enhancements targeted at decreasing parochialism itself. Firstly, as I have already argued, parochialism is so intrinsically connected with co-operation that decreasing parochialism while disregarding other traits will be likely to lead to less individual co-operation, which would not necessarily be desirable or intended. Secondly, if we pursue the eradication of parochialism by making group membership irrelevant, this is likely to lead to extreme individualism. In the human population, those individuals with low levels of parochialism are actually individualists and non-parochial co-operators are rare (Aaldering, Greer, Van Kleef & De Dreu, 2013; Aaldering, 2014). Disregarding group membership is a behaviour expressed by those for whom groups do not matter and for whom the only relevant factor is their own payoff. This indicates that the enhancement of a more inclusivist morality will not be trivial. It also indicates that attempts at increasing inclusivism with traditional means of moral education will face even more difficulties as the strong connection between parochialism and co-operation is unlikely to be overcome without substantial, and precise, technological modification.

3.3. Deep moral enhancement might be self-reinforcing and irreversible

According to some views, moral traits would be dispositions consisting of certain higher-order desires – the desire to desire *x* – producing motivations (e.g. Smith, Lewis & Johnston, 1989), which then generate behaviour. In many views, there is an important connection between moral traits and motivation (e.g. Zahn, de Oliveira-Souza & Moll, 2011). I contend that if there is a link between motivation and the target of deep moral enhancement, then it seems it could be subject to irreversibility and self-reinforcing effects. For instance, a hypothetical drug causing individuals to place a higher value on truthfulness could: (1) make them unwilling to reverse the change since they now place an even higher value on telling the truth and (2) make them become iteratively more prone to being truthful through further similar enhancements, dismissing all other relevant values. One seemingly small mistake when performing deep moral enhancement could have large and irreversible unexpected consequences, which offers further support to the fragility claim – assuming some small mistakes will be made.

3.3.1 Self-reinforcing

Steve considers all forms of violence wrong.[29] He also considers life on Earth worthwhile albeit spoiled with aggression. He will confess that in his darkest moments, when confronted with extreme injustice, he has impulses of committing aggression as a means to achieve justice. He wishes he did not. Steve wants to enhance morally. He takes a pill to become less aggressive and now considers all forms of violence even more wrong. Occasionally, his mind still entertains whether the severity of violence can be outweighed by other goods – but he now entirely despises these considerations. Steve wants to enhance himself morally. He takes a pill to eradicate these thoughts completely. Eventually, Steve

[29] I will use a fictitious person with oversimplified values here, but I expect similar examples can be found whenever there are competing moral dispositions that can be enhanced unevenly.

will be willing to take a pill that would make him willing to sacrifice all else for peace. Life on Earth will not look worthwhile anymore; no amount of happy, fulfilling lives can outweigh the violence it contains. Steve wanted to be morally better and to have less aggressive dispositions; Steve did not want to become someone in favour of human extinction. He would never take a pill that would cause him to consider life on Earth as not worthwhile. But morally enhanced Steve, eventually, would.[30]

Deep moral enhancement is likely to produce *self-reinforcing chains of modifications*.[31] Increasing one's moral dispositions to be motivated towards X will increase the perceived value of increasing one's moral dispositions to be even more motivated towards X and eventually lead to extremism towards X. Arguably, if the enhancement is deep – targeted at fundamental human traits and widespread – but sufficiently moderate (and we ignore other problems mentioned in this Chapter), then the result of one small enhancement interaction will be likely to be morally desirable. If we agreed that becoming more utilitarian is morally desirable, mapped the neurological structures and neurochemical pathways related to utilitarian behaviour correctly and developed a drug that increases utilitarian motivation, then using this drug would be likely to bring about moral enhancement. Furthermore, if use of the drug becomes widespread – and perhaps ideal conditions develop in which everyone could be convinced by and act in the light of utilitarianism – then we would have produced a humanity with a different set of values, goals and motivations. Being more utilitarian, they would be likely to be more prone to wanting to develop and take new drugs, which in turn would make them desire those values and have

[30] Such a scenario is not solely the result of mistakenly thinking decreasing violence is the only relevant dimension to be improved. Even if the initial intention was to decrease the inclination towards violence just a moderate amount, Steve would not be able to stop with just one single modification.

[31] Although one might rightly claim such chains would not be properly considered moral enhancements any more, the initial step that led to them would still be seen as a straightforward path towards moral enhancement. One can see the arguments of this section as showing that these intuitively desirable modifications would not be, in fact, moral enhancement.

those goals to an ever-greater extent. Eventually, a large number of iterations would produce individuals who would be considered morally undesirable by the ones who first engaged in the enhancement process; they might even feel morally disgusted by them, seeing them as the alien and immoral products of far too many radical modifications. However, if it is then likely that the enhancement operation will change individuals to make them want and value radically different things, to want to enhance more towards these values and finally to become immoral from the perspective of the initial individuals, then why would it be morally desirable to embark on the first enhancement interaction to begin with? The most desirable outcome would be the one to be found in a middle step, but we would be unable to stop there since it would entail further iterations.

Alternatively, such reasoning could be a fallacious instance of a slippery slope argument. A slippery slope argument concludes that a present course of action – considered desirable now – is wrong because it may produce a line of causation leading to a future undesirable consequence. This reasoning is frequently deemed unsound (Douglas, 2010; Volokh, 2003). Nonetheless, not all instances of slippery slope arguments are deemed fallacious; due to the strong motivational self-reinforcing aspects at play in moral enhancement, the use of such an argument might be sound. Douglas (2010) argues that slippery slope arguments that claim that performing a currently desirable *mild* version of action now is wrong on the basis that it will entail that future persons perform an *extreme* version of that action, can be self-defeating. If the mild action is desirable, and if future people consider it desirable to perform the extreme action after experiencing the effects of the mild action, then perhaps we should take the willingness of future people to undertake the extreme action as evidence that the extreme action is desirable.

As Douglas points out, however, this would only be the case if future people were to have epistemic access to the moral desirability of performing those extreme actions equal to

or better than the access we currently have. It might be that deep moral enhancement in a certain direction will necessarily lead to a bias for the desirability of more moral enhancement in that direction. Douglas argues that experience with the effects of the mild action will typically give future persons better epistemic access to the moral desirability of performing the extreme one. But this would not be the case for deep moral enhancements; if the mild action itself could bias future persons' epistemic access then experience with it is detrimental. Morally enhanced Steve has worse epistemic grounds for deciding to enhance morally than non-enhanced Steve had. If we currently only value a certain level of utilitarianism but frown on extreme utilitarianism, we would perform moral enhancement in order to produce individuals only mildly more utilitarian than ourselves. But these persons would not have the same epistemic access to whether extreme utilitarianism is morally desirable or not.[32] It might well be that for future persons, extreme utilitarianism is morally desirable, but the fact it is not for present persons has a greater bearing on the question of what we should do now than the potential moral inclinations of future persons (unless we are specifically enhancing epistemic access to moral statements). We want to be morally better according to our conception of the good, not according to enhanced persons' conception of the good. If we let the values of future enhanced persons matter more than our current values, then we will lose a great deal of value heritability. We want to fix our failings to realise our current values, not to alter our values themselves. We might want to improve our instrumental goals or accidental values, but we want to improve these in order to achieve our fundamental values more efficiently. We do not want to change our fundamental values, and this should make us wary of deep moral enhancement.

[32]There is the possibility that deep changes to our motivational system will not change our fundamental moral values in a significant manner. Either way, increasing motivation towards utilitarian actions will also increase motivation towards being able to perform more utilitarian actions regardless of whether utilitarianism increased its moral value.

Furthermore, even if we take the fact that future persons would consider it morally desirable to become extreme utilitarians as admissible evidence of the moral desirability of extreme utilitarians, such evidence – as Douglas admits – could be countered by a strong present belief that extreme utilitarians are morally undesirable. I contend there are many moral inclinations that we consider morally desirable to have to a greater extent, but that we strongly believe would be wrong to have at an extreme level. My initial arguments against overly increasing morally desirable traits offer support to that contention.

3.3.2 Irreversibility

Iterative deep moral enhancement was bad for Steve. Let us say we fixed that problem by simply committing to not continuously enhancing.[33] Steve morally enhanced to become less violent. But moral traits are extremely complex. A moderate decrease in Steve's aggressiveness made him less likely to be outraged by injustice. World poverty seems less revolting now; he takes fewer aggressive actions against it. Steve should reverse the change. But morally enhanced Steve does not want to become more aggressive or to be more revolted by world poverty; moral outrage looks too close to violence to him. Initially, Steve would never want to not be revolted by world poverty; now he is stuck with apathy.

Deep moral enhancement will probably be *irreversible*. Increasing one's moral dispositions to be motivated towards X will decrease the perceived value of decreasing one's moral dispositions towards X – that is, of reversing the increase. Presumably, prior to the enhancement, the perceived value of decreasing one's moral dispositions towards X was already lower than the perceived value of increasing it, hence decreasing it even further will mean such an action will become more unlikely. If we shift a motivational structure in a certain direction, this plausibly creates a chain of motivations that would function to

[33] Note that enforcing such a commitment would be hard if deep moral enhancement is widespread.

maintain that motivational structure, causing it to be irreversible. The new value structure – ascribing less value to the previous structure – would naturally be opposed to reverting. When Steve takes a drug that causes him to value pacifism more, he will be less willing than before to become less pacifist or to take any drug that would cause him to commit violence. Higher-order desires, or our values, motivate one against any action that would deeply change these desires. There is a strong reason for being unwilling to change one's goals; in the absence of an intelligent agent with a certain goal in the future, there is little reason to expect such a goal will be fulfilled. Humans might not be considered fully efficient rational agents and thus allow for manipulation of their goals. But if we perform deep moral enhancement so that we can act more efficiently to realise our values, then it becomes more probable that we will not want to change our goals and reverse this change.

Irreversibility concerns not wanting to go backwards. Self-reinforcement concerns increasingly wanting to go forward – in this sense it indirectly implies irreversibility. I have separately argued both of these features occur in deep moral enhancement. Any possible mistake from performing moral enhancement wrongly is either amplified by self-reinforcement or made unfixable by irreversibility. Although extremism could be seen as a mistake in itself, these two features of deep moral enhancement are not in themselves wrong. Nonetheless, considering the complexity claim, it is plausible that there will be mistakes. Therefore, self-reinforcement and irreversibility do imply increased fragility by increasing the scope of each one of those mistakes. Deep moral enhancement might be performed without directly changing higher-order desires or motivation; in that case, the two effects explored here would be less likely. On the other hand, shallow forms of moral enhancement might modify motivation. Nevertheless, in so far as some motivations are connected to fundamental aspects of human morality: (1) deep moral enhancement is likely to change

these motivations, and (2) changes primarily expected to alter these motivations will fall under the definition of deep moral enhancement. Finally, even if motivations were not seen as an important aspect of human morality, it would still be the case that deep changes to them would be susceptible to self-reinforcement and irreversibility, which could be morally undesirable regardless of the role of motivation in morality.

3.4 Aetiological complexity increases fragility

The aetiological argument laid down for the complexity claim can also be used to support the fragility claim. The fact that our current moral traits can be explained by assuming they were partially shaped by many past contingent events with considerable enduring effects implies we should expect that moral traits have a high susceptibility to such contingencies. One account of contingency is that of frozen accidents (Gell-mann, 1995), small random events that produce long-lasting consequences by putting in place an irreversible course of events (Bennett & Elman, 2006). Suppose your dog stole, played with and buried one of your books. Upon finding the book, it is chewed, covered in mud, wet, stained, it smells bad and so on. The causal history that accounts for all those marks will be full of contingencies: the dog stole the book, chewed it, dragged it down the stairs, urinated on it and so on until you found it a few weeks later. Suppose he did the same with your wedding ring. After a quick wash the ring will be as good as new and its causal history will be: the dog stole the ring, you found it a few weeks later.[34] Robust objects, subjected to a long history, will necessarily have simple causal histories and simple descriptions. Fragile objects, subjected to a long history, will have complex causal histories and complex descriptions. It is worth making a comparison to physical traits. For instance, the ability to

[34]Arguably, whether or not the book's marks are relevant will play a role on whether or not we need to evoke a complex etiology. It might be the book is susceptible to contingencies that do not affect it in a relevant manner for the purpose in question. However, it does not seem that the full range of human moral traits is something we can ignore when implementing moral enhancement.

transport oxygen by red cells has little relative variation in the human population and is well understood. There is a multiplicity of founder effects, population bottlenecks and so on that could have happened in the past that would have left no mark on our oxygen-transporting mechanism. As my arguments predict, this ability is relatively simple and robust, and can be currently enhanced with the use of the drug EPO. The same cannot be said about moral traits; it resembles a pile of frozen accidents more than oxygen-transporting does. Hence, we should be more careful when trying to enhance human morality than when trying to increase human endurance.

4. Objections and responses

4.1 Redundancy, modularity and canalization

A significant line of counter-argument against my fragility claim can be found in Allen Buchanan's chapter "Conservatism and Enhancement" in his 2011 book *Beyond Humanity*. Buchanan (2011) argues against the idea the human physiology is akin to a house of cards where only one small apparent improvement could bring the entire system down. According to him, generally speaking, the evolutionary processes that resulted in our current traits tend to produce very robust end products. There are three causes of this robustness. Organisms often have more than one feature to perform the same function: usually, new adaptations evolve without the old ones being replaced, creating redundancy. He does not mention this case, but redundancy often evolves as a survival adaptation against losing essential features for the organism (Zhang, 2012). Features are costly to maintain; if two features perform exactly the same function and conferred no advantage of being duplicated, then one of them would be selected against or undergone functional divergence. However, for instance, some of the cases of enzymes' redundancy can be explained by a selective pressure to preserve essential functions of the organism even when mutations or diseases

affect the expression of one of the redundant enzymes.[35] Moreover, Buchanan observes that organisms are extremely modular and removing or altering one system does not entail a change to any other system. Finally, there is the fact that often small variations in the genotype or environment do not produce any variation in phenotype – a feature called canalization. For instance, if the last nucleotide in a DNA codon unit of three nucleotides is changed, the sequence will normally still codify the same amino acid. This protects the organisms against the common misreading of the last nucleotide. As such, this is an adaptation against fragility. Redundancy, modularity and canalization all significantly reduce the chance that one localised change, even if disastrous, will harm the whole organism. As Buchanan points out, these features have evolved exactly to prevent excessive fragility of organisms. But what is not acknowledged in his discussion is that they have evolved to prevent fragility from the types of threats that were recurrent throughout evolutionary history. For instance, the codon canalization example exists because on one of the final steps of translating the genetic code into proteins the interaction with the codon is weaker at the last nucleotide, which means it is more susceptible to being paired with the wrong amino acid, which could lead to the wrong protein being synthesized.[36] However, when only the last nucleotide is wrong, it does not affect which amino acid it pairs with, thus preventing the error from propagating to the protein synthesis. In the same way, the redundancy of enzymes is an adaptation to prevent losing essential functions due to recurrent types of mutation or diseases harming the expression of an important enzyme. But as the changes brought about by human enhancement are outside of the scope of natural selection, there was never an evolutionary pressure to create organisms that would be robust to

[35] There seems to be no redundancy in the case of enzymes that are essential but in whose operation redundancy would cause an deleterious imbalance in the body due to several enzymes performing the same role (Kelso, 1994). Redundancies can also provide a fine-tuning of a certain functions (Kafry, Levy & Pilpel, 2006). These findings indicate redundancy can be an evolved adaptation and not just a by-product.

[36] Redundancy would happen regardless because there are more codons than encodable amino acids. But this does not explain why the third nucleotide is often the redundant one.

modifying themselves with the use of technology. Moreover, even with this level of robustness, this has not prevented over 99.9% of all species that have ever lived from going extinct, mostly due to changes in the environment too drastic for these species' levels of evolved robustness (De Vos, Joppa, Gittleman, Stephens, & Pimm, 2015; Wills, 2008). For instance, one of the most devastating mass extinctions on our planet happened when our atmosphere became rich in oxygen, thereby extinguishing most obligate anaerobic organisms from Earth. Organisms cannot evolve to have any level of robustness against completely new modifications such as the ones entailed by deep moral enhancement because they have never been exposed to such selective pressures. Finally, I have argued here that moral traits, in particular, are fragile; not that every human trait is fragile. One peculiarity of the cognitive processes involved in human morality is that they rely on multiple systems from various sort of brain areas with different functions. Human moral traits are, therefore, particularly non-modular and overly interconnected; thus Buchanan's modularity argument does not hold for moral enhancement, although it might hold for most other targets of human enhancement. I will address the modularity objection in more detail a few paragraphs below.

4.2 Further objections

I have presented an aetiological argument for the complexity of moral traits due to historical contingencies. On the contrary, it would seem that there is not as significant an amount of historical and geographical variation in moral reasoning as the complexity claim would predict. There are many moral principles shared across most cultures and historical periods, but if moral traits were subjected to so many historical contingencies, the variability should be more drastic. Responding to these objections, it should first be noted that the existence of a simple set of cross-cultural moral principles is an empirically open question

at best.[37] Secondly, even if there are simple cross-cultural moral principles, this does not preclude that the larger share of human morality, not subsumed under these principles, could present greater variability and complexity. Indeed, it seems that the present picture of human morality contains evidence of some cross-cultural similarities in some areas, but also evidence of a wide variability in other areas. Thirdly, it might be that historical contingencies are responsible for the complexity of moral traits to a minor degree and contingencies in natural history – which is populated with far greater variability – are the main factors. In this case, the argument would lose some force but would still stand. Furthermore, one might counter-argue that even if I present evidence that there is a lot of variability, then according to the fragility claim it would also have to be the case that most of those variants were unsuccessful. However, it should be noted that the scope of the fragility claim is deep modification. It might as well be that there are not as many cases of moral decay as one would predict if moral traits were fragile under conventional modifications of morality; however, we do not know empirically whether this would still be the case for deep modifications, which could be possible with future technological advances.

I have presented evolutionary arguments for the complexity of moral traits. However, subscribing to the view that natural selection shaped moral traits would be likely to also mean subscribing to another view commonly held in evolutionary psychology; namely, the view that cognition should be processed by highly specific, independent evolutionary brain modules, and thus be fairly simple and robust instead of complex and fragile. Firstly, it should be noted that evolutionary modules are expected to be robust under the recurrent challenges present in our ancestral environment – our *environment of evolutionary*

[37] There is extensive research finding commonalities and overlaps between the morals of different cultures, but there is no simple set of values that are uncontroversially accepted as universal. There is also plenty of variation in how and which of those norms are applied in any given culture. Interestingly, most of the cross-cultural studies attempting to find a universal morality have arrived at different conclusions. For instance, compare Schwartz (1992) and Dahlsgaard, Peterson & Seligman (2005).

adaptedness (EEA). This would agree with my claim that conventional means of moral improvement – some of them present on our EEA – are safe. Nevertheless, this would not mean that moral traits should be robust under non-conventional means of deep moral enhancement. Rather, given that it was never an evolutionary challenge to survive and reproduce in an environment where we could have the freedom to set our moral traits according to our will, we should expect that our brain modules for moral traits are not robust under deep moral enhancement. Secondly, although each module might be specific and simple on its own, this does not entail that they should be simple taken as a whole, or that modifying them without their full description – which, even if simple, we do not currently possess – should be considered safe.

I have conceded that my evolutionary assumptions mean moral traits should be robust under the EEA, but fragile otherwise. However, our current modern environment is radically different from the EEA; hence, moral traits should be already in a risky and fragile setting. Although I believe deep moral enhancement could further increase that fragility, I am willing to admit our current situation is also fragile. I do not disagree that our present morality is unfit for a world of unprecedented technological powers. Instead, I claim that deep moral enhancement, as an unprecedented technological force on its own, although possibly beneficial, will also bring about increased fragility.

It could be argued that given that our current understanding of moral traits is limited, it is bound to be the case that we see the matter as extremely complex. However, this was true of several other fields before we found a very simple law or principle.[38] Firstly, I have given arguments that indicate that moral traits are actually more complex than other traits; thus we would have reason to suspect they are not only contingently complex; that even after

[38] For instance, the astronomical laws explaining the apparent retrograde motion of planets became substantially simpler after Kepler's laws were proposed.

the simplest possible theory is produced, such a theory will still be very complex. Secondly, regarding the epistemic complexity of moral traits – that is, the difficulty in accessing our final values and so on – I am willing to admit it might be the case that we could overcome this difficulty one day. However, until such a day the strength of this argument remains in force. It would not violate this specific worry to prescribe that we should not attempt deep moral enhancement until a full account of moral traits is provided.[39]

Finally, the overall reliance on moral traits could be seen as unconvincing from the standpoint of philosophical positions where they never seemed particularly relevant in the first place. For instance, several utilitarian moral philosophers such as Joshua Greene and Peter Singer believe moral intuitions, and maybe moral traits in general, are not the main guides for building a normative moral theory (Greene, 2014; de Lazari-Radek & Singer, 2012).[40] However, even if we discard our current moral traits as a primary guide for normative theory, it is still the case that our moral traits are the starting material that moral enhancement will attempt to improve. It might be that the morality we should have bears little relation to the morality we currently have, but the complexity and fragility of the former is a reason to suspect that morally enhancing towards the latter is risky.

5. Conclusion: complexity and fragility lead to increased risks

In order to improve or enhance a certain human trait, we have to be provided with two reference points: the current state of the trait and its desired state. The complexity of moral traits means that the current state required by moral enhancement is hard to define. The fact that complex processes and contingencies dictated the aetiology of our moral traits means that the simplest possible theory explaining them should be fairly complex. Moreover,

[39] As argued in Chapter 4, it would raise other concerns related to the fact we would be dismissing a possible solution to the extreme risks facing humanity.
[40] For a compilation of several instances of misguided moral intuitions see Beckstead, 2013, pp. 25-53.

given that we are not fully aware of the final goals of our moral traits, it is hard to access the hierarchical organisation of our values that would be needed to produce such a theory. These arguments lead to the conclusion that the starting point of deep moral enhancement is hard to provide, and thus deep moral enhancement is a difficult enterprise.

Furthermore, not only would a full account of moral traits be complex, but also any deep alteration of them would be more likely to cause harm than benefit. Moral traits have a particularly high proclivity to unexpected disturbances, as exemplified by the co-operation case, as amplified by its self-reinforcing and irreversible nature and finally as its complex aetiology would lead one to suspect. Even the most seemingly simple improvement, if only slightly mistaken, is likely to lead to significant negative outcomes. Unless we produce an almost perfectly calibrated deep moral enhancement, its implementation will carry large risks. However, if we cannot even provide a simple account of moral traits, a perfectly calibrated deep moral enhancement will be hard to produce. When put together, the complexity and fragility claim entail even greater risks from deep moral enhancement.

It seems clear that these two claims lead to the conclusion that deep moral enhancement posits significant risks and will be much harder to develop safely than other types of human enhancement. However, it should be emphasized that those claims only lead to the conclusion that deep moral enhancement is likely to be hard; they do not suggest it should be impossible nor undesirable. Given that deep moral enhancement could prevent extreme risks for humanity, in particular decreasing the risk of human extinction, it might as well be the case that we still should attempt to develop it. I am not claiming that our current morality is well suited to dealing with global problems. On the contrary, there are certainly reasons to expect that there are better moralities that could be brought about by enhancement technologies. However, I believe my arguments indicate there are also much worse, more socially disruptive, moralities accessible through technological intervention; or at least that

we are likely to produce worse moralities. During the last Chapter of this thesis, I will outline more positive conclusions from the arguments laid down here, attempting to foresee safe paths towards fundamentally enhancing human morality. One of the tasks there will be to make deep moral enhancement safer by attempting to avoid possible unexpected consequences of the type explored here. For instance, self-reinforcement and irreversibility seem to suggest there should be limits to prevent runaway deep moral enhancement and that there should be a reversibility mechanism in place to prevent it being permanent. The conclusion I have argued here, however, states that these safe paths will be hard to find and that pursuing deep moral enhancement while failing to find them can have severe consequences.

Chapter 3: Moral status enhancement and individual interests

1. Introduction: supra-persons and moral enhancement

Persson & Savulescu argue that humans need to enhance their moral traits if they are to deal with problems on the scale of global warming and nuclear proliferation. These problems deal with an almost blind spot of our moral dispositions: large-scale cooperation. Douglas argues that an enhancement that leads someone to have morally better motives is clearly advantageous for society. In this Chapter, I will primarily address concerns and responses targeted at societal level moral enhancement, such as that proposed by Persson & Savulescu. I will tangentially address some issues related to the desirability of single individuals performing moral enhancement in section 3. The motivation behind these concerns is the possibility that the morally enhanced could form a new class of superior beings with higher moral status, putting those who do not enhance in a precarious position. These are valid concerns that, I will argue, have been met with unsuccessful responses.

I have defined *deep moral enhancement* as significant changes, brought about via technological interventions, directly targeted at fundamental human traits (e.g. co-operativeness, empathy, altruism, etc.) primarily expected to lead to morally better behaviour or motives. By fundamental human traits I meant general and stable patterns of behaviour or cognition such as empathy, honesty, aggressiveness or extraversion. Moral traits are all fundamental human traits that are primarily involved with human morality in a descriptive sense[41] and, more importantly, that would reasonably constitute a target of deep moral enhancement. As noted in Chapter 1, although the feasibility of such radical manipulation

[41] That is, in the sense of being a description of human morality as an empirical matter (e.g. moral psychology) and not in the normative ethics sense.

of human moral dispositions in the near future remains uncertain, recent studies have demonstrated some aspects of moral reasoning are subject to pharmacological manipulation. Therefore, it is reasonable to assume we might come to develop and use drugs that will enhance our moral capacities, and, importantly, categorically assuming otherwise can lead to dismissing real potential risks.

Moral status, for the purposes of this thesis, will be defined as the property an entity has such that its interests matter morally for the entity's own sake. I will also assume that moral status can vary and that the interests of entities of higher moral status matter more than the equivalent interests of entities of lower moral status.[42] In particular, the interest in continuing to live of entities with higher moral status has a higher degree of inviolability than the corresponding interest of entities with a lower moral status.[43] It is generally assumed that persons have a higher moral status than non-human animals, on the basis of their higher morally significant psychological capacities such as moral and non-moral reasoning.[44] Therefore, in so far as the likely targets of deep moral enhancement correspond to these capacities, deep moral enhancement would produce beings of an even higher moral status, i.e. supra-persons. This inference draws further support from the fact that non-human animals also seem to possess a higher moral status than, for example, inanimate objects, by possessing a certain level of some morally significant psychological capacities that are primarily moral in kind.[45] Perhaps we could achieve moral status enhancement through cognitive enhancement alone. However, deep moral enhancement seems more likely than other types of enhancement to affect whatever capacities confer moral status. These

[42] Some brief arguments for the assumption moral status varies are given shortly.

[43] Here I follow a similar definition as found in DeGrazia (2008), McMahan (2005), Warren (2000) and Korsgaard (2018). The last uses the term moral standing instead of moral status, but actually shares the most similar definition.

[44] Even those who assume animals have no moral status whatsoever will tend to agree with this claim.

[45] This argument is developed (among many other arguments) in the fifth chapter of Korsgaard (2018).

capacities might include sufficient cognitive capacity to perform moral reasoning. Arguably, the overlap between moral traits and morally significant psychological capacities should be extensive. For instance, empathy, practical reasoning, moral dispositions, moral conformity, and co-operation would all be possible targets for moral enhancement and candidates for morally significant psychological capacities. Hence, significant changes brought about via technological interventions, that are directly targeted at fundamental human traits (e.g. co-operativeness, empathy, altruism, etc.) and are primarily expected to lead to morally better behaviour or motives (i.e. deep moral enhancement, as defined) are likely to lead to enhancement of moral status.

As introduced in Chapter 1, Agar (2013) and Sparrow (2014) believe that if we are effective in producing individuals with such higher capacities, then the inequality created by this could outweigh the benefits. Agar asserts that moral status enhancement will create a future with two different classes of persons: the mere human persons who exist today, and a class of supra-persons whose greater morally significant capacities will confer a higher moral priority over mere persons. By contrast, Buchanan (2009) believes that we cannot increase moral status. For him, having a higher degree of whatever characteristic confers personhood does not confer higher moral status. Buchanan argues that only a threshold could explain the intuition claiming that all humans share the same moral status equally. On the other hand, McMahan (2010) believes that if Buchanan is right about enhancing existing capacities not leading to higher moral status, then enhancements of various sorts might produce a new emergent morally significant capacity, which could plausibly warrant a higher degree of inviolability of the right to life, i.e. higher moral status.

Douglas argues that if we were to reject the idea that supra-persons could have higher moral status, it is still plausible to assume that the creation of supra-persons would harm mere persons who remain unenhanced (Douglas, 2013). This reasoning supports a general

motivation for this thesis, stated in section 5 of Chapter 1, which is that a risk assessment of moral enhancement is needed regardless of conceptual disagreements. All the same, Douglas reasons that this does not render the creation of supra-persons undesirable, unless, among other things, we adopt unjustifiable partiality to the unenhanced, or ignore the gains of becoming a supra-person. He contends that there is another view of moral status that does not deny the existence of a threshold and that fits our intuitions equally as well as Buchanan's, but that nevertheless can accommodate higher moral status: the Plateau View. This view states "there is a range of mental capacity within which all beings have the same moral status, and within which all currently typical adult humans lie, but above which moral status rises, either gradually or in steps" (p. 479). Supra-persons with higher capacity for morally significant properties would then seem to enjoy a higher moral status in the same way that cognitively normal adults enjoy higher moral status than that of cognitively complex non-human animals, and perhaps even foetuses or severely cognitively limited human adults. Furthermore, Douglas notes that it is hard to pinpoint a morally significant property that would provide all members of the human species with the same level of moral status, excluding all non-human animals. Even all-or-nothing properties fail to provide criteria for personhood in which all humans enjoy the same moral status. The case for equal moral status is further complicated by the fact that most candidate properties such as cognitive complexity admit degrees.

Persson (2012) reasons that it would be morally permissible to prevent supra-persons from coming into existence only if the harm they cause to mere persons would be morally impermissible. In his example, an athlete has no valid reason for preventing a competitor from training on the grounds that this training will increase the likelihood of the competitor beating him, because beating someone in a competition is a permissible harm. Only when the harm is great enough does it become impermissible. Given that Agar himself concedes

that the harm caused by supra-persons would be morally desirable, then it must also be morally permissible. However, Agar holds that moral rightness is relative to species. He calls his position species-relativism, asserting that whether a moral judgment is right or wrong is relative to a species, depending on a particular set of species-specific experiences. "According to species-relativism, certain experiences and ways of existing properly valued by members of one species may lack value for the members of another species" (Agar, 2010, p. 13). Thus, he holds that if we eliminate the possibility of certain experiences specific to *Homo sapiens* by becoming supra-persons, we would have a different form of morality (Agar, 2013). Therefore, although harm to persons would be morally permissible from the standpoint of supra-persons, it might be morally impermissible according to persons, because they would not value the benefit gained by supra-persons who have not had specific human experiences.

I will make a similar case here, but I will not draw support from any form of species-relativism and ascribe any fundamental moral significance to the human species – Bernard Williams' defence of this position is discussed in the last paragraphs of section 4.2. Besides leading to unsatisfactory conclusions when considering the moral status of the severely cognitively limited and higher non-human animals,[46] I believe using the concept of species to judge moral status enhancement would be considered inappropriate even by biologists. Although widely used in biology, the concept of species is known to be unsuccessful in borderline or artificial cases. Moral status enhancement is precisely both a borderline and artificial case; therefore, subscribing to a moral view whereby what has value is tightly tied with such a rigid concept would lead to the same problems as this concept faces in biology. Moreover, belonging to a different species does not entail having radically different values,

[46] It would ascribe a brain-dead human full moral status due to still belonging to our species, but would give no moral status to a highly intelligent chimpanzee.

only being reproductively isolated. Finally, species-relativism implies that the act of an individual intentionally exiting the human species *due to* valuing other experiences will always be considered mistaken,[47] which seems counter-intuitive. I will ultimately defend some form of partiality towards the kind of beings we are, but I do not believe that we are – in a morally relevant sense – beings of the kind *Homo sapiens*.

In the next three sections I will analyse how these concerns about and defences of moral status enhancement are applied to three different possibilities for the creation of supra-persons. I will argue that many of those defences are flawed. The possibilities will be increasingly favourable. First, I analyse the case in which supra-persons come to suddenly co-exist with current persons, who become susceptible to harm; I argue that this is intuitively problematic. Next, I look into the case of creating supra-persons by enhancing existing persons; I argue that this can be detrimental to certain individual interests. Lastly, individual interests at play in generational continuity discussed by Samuel Scheffler (2018) reveal that even the less problematic case of generational replacement of persons by supra-persons contains risks so far ignored in the literature.

2. Harming persons

The most concerning possibility would be if we created supra-persons *de novo* within one generation leading to an increased likelihood of harm to persons.

> *Harming persons* – We engineer beings with significantly higher cognitive, emotional and moral capacities, which as a consequence have supra-personal moral status.[48] They mature at an accelerated rate. Within a decade, adult supra-persons *co-*

[47]One may still wish to exit the human species for other reasons, e.g. due to having disvaluable experiences and preferring to have no experiences at all or some other species' experiences.

[48] Here I do not want to specify which of these capacities are necessary or central for producing higher moral status. Instead, I merely commit to the claim that the orchestrated radical enhancement of all of them would plausibly be sufficient to produce a higher moral status.

exist with unenhanced persons, who become susceptible to being denied resources, harmed or killed for supra-persons' sake, perhaps in a manner analogous to persons' treatment of non-human animals.

The creation of beings with a higher degree of inviolability of their interest to continue to live would mean that persons, who have enjoyed so far the highest degree of inviolability, would be more likely to be harmed.[49] There is likely to be no explicit desire to harm, but supra-persons would find justification to sacrifice beings of lower moral status if it benefited them. Given that these beings will have higher moral capacities by stipulation, some might think that if we develop technology that enables us to create such beings, then we must do so. Ethical analysis of the question ought to stop and we must engineer supra-persons, who will then be better able to inform us whether we should have done so. I will argue against this view.

To be sure, the concern that they would surely harm us for immoral and selfish reasons is unfounded, for they will have higher moral capacities. If they sacrifice us to their ends, these beings will be likely to be right in thinking that they are justified in doing so; I will not address the possibility in which we create beings who are likely to be wrong by accepting such a belief, but who mistakenly arrive at that conclusion nonetheless.[50] Alternative courses of action that prevented the sacrifice of persons would be likely to be preferred by supra-persons, but these do not ensure that resource disputes would necessarily be prevented. For instance, when resources are scarce we already prioritize patients with greater life-expectancy in public health care without any desire to harm those with lower

[49] Some might claim inviolability is either absolute or nonexistent, similarly to arguments against higher moral status. But arguments against absolutism of inviolability (McMahan, 2009, pp. 599-600) seem even stronger than against the Plateau View.

[50] As noted earlier, even if one rejects the plausibility of supra-personal moral status, these enhancements would possibly produce beings who can harm persons and who behave in a way compatible with possessing a higher moral status. For those taking this position, the investigations here can still be justified as a matter of risk assessment.

life-expectancy. Conversely, it would seem that if we were to believe possessing a higher moral status would involve having higher moral capacities, then it could be argued that although persons would become susceptible to being harmed, it is unlikely that beings morally superior to us will want to harm us. For instance, one might point out that persons already have a strong moral inhibition against extinguishing or harming beings of lower moral status, such as non-human animals; thus, supra-persons, possessing even higher moral capacities, will feel even more inhibition against extinguishing or harming beings of lower moral status, such as persons. It is possible that supra-persons will be so abruptly more efficient than persons in gathering and using resources that they will find no compelling reasons to dispute resources with persons. However, it seems unlikely that enhancement will suddenly produce these beings without first producing beings that are only moderately more efficient than us. If it is to be possible at all, suddenly creating agents with astronomically greater efficiency is more likely to be a result of Artificial General Intelligence technologies, which lie outside the scope of my analysis here.[51] Moreover, I will argue shortly that sudden enhancements could cause more harm than gradual ones.

However, the claim that persons have a strong moral inhibition against harming or killing beings of lower moral status is dubious. It is true that most human beings with full moral capacity do feel inhibited about harming or killing non-human animals, but this inhibition has never stopped humans from harming and killing farm animals. In actuality, persons do not have sufficient moral inhibitions against annihilating (or inflicting unnecessary harm on) species of lower moral status than their own in order to refrain from doing so. Nonetheless, let us suppose supra-persons had higher inhibitions not to harm persons than persons currently have not to harm non-human animals and believed they had a duty to avoid the suffering of persons (which is perhaps analogous to the duty some

[51] For more on this subject see Bostrom & Yudkowsky (2014).

ethicists today argue we should have towards beings of lower moral status than our own, such as farm animals). Such a position regarding farm animals is seen as ascribing beings of lower moral status an exceptionally high moral relevance. The outcome of caring so much about farm animals is that, were they to be treated fairly, they would risk extinction. This is because fair treatment would incur unfeasibly high economic costs for us, the higher moral status beings; thus, we would conclude that ought to stop the farming of animals. The case of farm animals demonstrates that persons constantly harm and kill beings of lower moral status unnecessarily on an industrial and global scale. The best possible treatment that persons could currently envision towards farm animals is close to annihilation. Notwithstanding our higher moral capacity, we currently harm and kill non-human animals on a much greater scale than any animal with lower moral capacity does. The treatment that persons dispense towards beings of lower moral status does not seem to offer a basis for believing that supra-persons will treat us kindly. We might not be farmed for food, but we currently use a large share of the Earth's available resources, therefore it does not seem unreasonable to expect that we have some chance of being either used as a means to fulfil supra-persons' ends or of being denied resources.

Another form of argument claims that because supra-persons would make morally better judgments than persons then they would be right if they decided to harm us and we would be wrong. Therefore, this possible future harm should not offer us a reason against creating supra-persons. However, the fact that supra-persons will have improved moral judgment does not necessarily assume that they will have perfect moral judgment and thus necessarily treat beings of lower moral status fairly. It would be reasonable to expect that supra-persons would be less likely to have certain moral failings such as partiality to their own group. However, this does not entail that they would not have any degree of partiality. Arguably, their much higher capacities would mean that they would be more powerful than

persons,[52] and they would tend to be able to exercise this partiality more often, thus leading to outcomes where persons are more likely to be treated unfairly. In comparison to non-human animals, we have higher cognitive and moral capacities, which enable us to understand better that inflicting harm on other animals is wrong. That non-human animals lack this higher degree of cognitive and moral capacities has not been sufficient to prevent us from harming non-human animals more than they harm themselves. Our higher capacities seem to have increased our power to harm beings of lower moral status to a greater extent than they have increased our capacity to refrain from inflicting harm on them. Likewise, adults possess higher moral and cognitive capacities than children, but adults harm children much more frequently than they harm themselves because they can do so more easily. Therefore, there is nothing in the increase of moral status that guarantees treating beings of lower moral status better than they would treat themselves.

Moreover, we can consider the consequences if supra-persons do make a perfect moral judgment and conclude that from an impersonal perspective, the world inhabited by supra-persons would be better than the world inhabited by persons, even though there would be persons whose lives were made worse in the former in relation to the latter. Supra-persons might decide to annihilate persons if in impersonal terms this harm would be outweighed by supra-persons' gains, but those persons who would be annihilated would have an individual interest to prefer to live in the former world and thus to not create supra-persons.[53] The core intuition here is that we have strong reasons not to welcome our impending extinction, regardless of how morally superior our exterminators would be.

---

[52] Even if they are produced by the enhancement of a primarily moral capacity such as co-operation, their increased ability for co-operation, as observed by Buchanan, would radically increase their capacity to produce wealth, thus giving them more power. Moreover, they would be more likely to co-operate with other highly co-operative agents such as themselves, instead of with persons. Some form of partiality, even if in absence of aggression, is likely to arise in virtue of their existence as enhanced beings.

[53] This argument does not imply individual interests matter more than impersonal reasons nor does it suggest some measure of comparison between the two. It merely relies on the fact that there is still some uncertainty on how to balance competing individual interests and impersonal reasons.

Imagine we encounter an alien species who are much more intelligent and co-operative than ourselves. Suppose these aliens have evolved to lack any form of aggression whatsoever. They have studied human nature and history in detail and concluded the universe would be a better place if we go extinct. After having carefully explained how we are irredeemably immoral beings in comparison to them, they present us with an *extinction device* that would painlessly end all human life, thereby giving way to their advanced civilisation to flourish on earth. Would we activate the device? It seems most of us would be at the very least reluctant to do so. Bernard Williams (2008) has correctly observed that this kind of case shares a relevant similarity with the project of transcending humanity.[54] Creating supra-persons and being untroubled by the fact that they might sacrifice humanity for the greater good would be similar to being untroubled with activating the extinction device.

The intuition against activating such a device might be mistaken. Perhaps we have no moral basis to oppose being replaced by supra-persons and intuitions to the contrary are motivated by an ungrounded survival instinct. Nonetheless, the supposition that one should welcome one's death merely because one's killer has a higher moral status, and would thereby generate more impersonal gain, is, at the very least, debatable. Few people would think they should resign their right to life in these circumstances. Likewise, it seems controversial to suppose the whole of humanity should welcome its extinction because it would be carried out at the hands of supra-persons. The same kind of reasons we would have to fight for the survival of humanity if supra-persons attempted to extinguish us will give us reasons not to create them in the first place. Perhaps these reasons are significantly weaker than the reasons to create supra-persons; perhaps they are not. It is uncertain whether or not we would have to conclude that we ought to let them kill us all; therefore we should not act now based on this conclusion. After we create them, it might be too late to feel regret.

[54] See Section 4.2 of the present chapter for further discussion of Williams' arguments.

To conclude, there are three reasons why one would be mistaken in asserting that supra-persons' moral superiority over us necessitates that we create them first, so that they might then correctly decide whether or not they should harm us. Firstly, we do not have evidence for the assumption that increased moral status implies having the right set of reasons to prevent the infliction of impermissible harm upon beings of lower moral status. Secondly, it seems that increases in moral status might create more opportunities for acting on the wrong reasons, exerting impermissible harms upon beings of lower moral status to a larger degree, than for increasing the capacity to have the right reasons and to refrain from exerting such harms.[55] Thirdly, if supra-persons were to exert perfect impersonal reasoning and to conclude that we should go extinct, this would not eliminate the possibility that we still have reasons to oppose our extinction at their hands. We are uncertain as to whether we would not have countervailing individual interests to prevent our extinction in the event that there are impersonal reasons for us to go extinct. Creating perfect impersonal reasoners will not close this epistemic possibility but it might eliminate the possibility of us acting on it.

3. Enhancing persons

A less concerning possibility than the one that occurs when creating supra-persons *de novo*, would be if some persons willingly underwent various sorts of enhancements to achieve higher moral status.

> *Enhancing persons* – Through the right mixture of gene therapy and neurobiological interventions, within a few years, some persons can achieve significantly higher cognitive, emotional and moral capacities, and in consequence they come to possess supra-personal moral status. Unenhanced persons in turn will become susceptible to

[55] In principle, if we increase moral status by targeting capacities that are exclusively moral and do not entail increased power, then this relationship would not hold. However, it seems most forms of increasing moral status imply an increased power.

dangers such as the denial of resources and the infliction of harm or death upon them for the sake of the supra-persons, perhaps in a manner analogous to persons' current treatment of non-human animals.

Here, it might be argued that although the unenhanced persons would be harmed, there would still be a gain for those who underwent enhancement. For instance, Douglas' response to the concern that the creation of supra-persons will cause persons to lose relative moral status is that this loss will be matched by an increase in moral status for those who become supra-persons. He argues that the reasons against decreasing relative moral status of persons and the reasons for supra-persons gaining status seem equally strong. Therefore, from its effects on moral status distribution alone, there is no overall reason to oppose the creation of supra-persons.

One possible problem with this argument is that the desire of present persons to become supra-persons will lead to harm to future persons, whereas the desire of present persons to not be harmed by the loss of relative moral status will harm no one, given that it will merely result in the enhancement not taking place.[56] If some form of asymmetry between harming and benefiting holds when bringing people into existence (McMahan, 1981),[57] then the interest in not being harmed should count for more than an interest that creates harm. In the world where we create supra-persons, there are individuals whose lives are made worse in losing relative moral status. In the world where we do not create supra-persons, there are no individuals whose lives are made worse.

[56] Some might argue that frustrating persons' interests to become supra-persons would also harm them. But if we allow that harms can be brought about by not performing an action, then everyone is currently being inflicted with infinitely many harms from all the actions not being performed to their benefit, which is absurd.
[57] The view claims that a harm can be a strong reason not to bring someone into existence while a benefit does not constitute an equally strong reason to bring someone into existence. This problem would be more pronounced when creating supra-persons in the replacement case discussed shortly. However, I will not explore it further here.

However, I intend to pursue another reason for not counting the supra-persons' benefits equally to the persons' losses. The intuition behind such a reason is the following: the loss of persons' moral status would produce a possibility in which it is more likely for *persons* to be harmed to benefit supra-persons. The gain in status of supra-persons would produce a possibility in which it is more likely for supra-persons to benefit from the harm of *persons*. Why would a current *person* count a harm to *persons* as equivalent to a benefit for future supra-persons?

The obvious response, given that this case assumes that some of the original persons willingly enhanced themselves until they became supra-persons, is that the interests fulfilled by the gain in status of supra-persons are those of said original persons. If John becomes a supra-person, but Jim does not, then Jim's loss of relative moral status is matched by an equal gain in relative moral status by John. Douglas defends this view by arguing that there is no reason to presuppose that a person undergoing moral status enhancement would significantly alter the relevant psychological features that constitute her personal identity. Thus, for him, there are persons whose interests are such that they become supra-persons who gain moral status; this would balance the interests of present persons so that their future selves do not lose moral status. In cases of gradual enhancement done via moderate steps, Douglas' assumptions are likely to be correct, I will take no issue with such cases. But drastic steps are a plausible possibility for moral status enhancement; especially if we suppose it is done fast enough that supra-persons co-exist with current persons. I will argue that an individual interest implying radical changes to one's own psychological capacities would be likely to partially undermine its realisation.

I should clarify that I will take personal identity to mean whatever defines one's self over time and thus justifies our patterns of egoistic concern. An egoistic concern is the kind of concern we have about what happens to ourselves but do not have about what happens to

others; they are a kind of individual interest. Considering when these concerns are present in hypothetical cases of extreme longevity, fission and fusion reveals that what explains our patterns of egoistic concern is the presence of psychological connections. What grounds our patterns of egoistic concern is the continuity of sufficiently strong psychological relations with our future self, that is, the continuity of psychological connections. Examples of plausible psychological connections with a future self are having memories of the same past experiences, expressing sufficiently similar personality traits, having sufficiently similar desires, beliefs and intentions, the continuous flow of consciousness, and the pursuit of goals and projects that are later realised or frustrated. There are many psychological relations between moments of one's life – or a person's time-slices – that are plausible psychological connections and many come in degree. But identity is all-or-nothing. Consequently, Derek Parfit and others following his work hold the view that personal identity is not what matters for egoistic concern. Although I agree that psychological connections are the grounds of egoistic concern, I will not abandon the personal identity terminology. I believe that if I replaced all instances of personal identity with the continuous holding of the psychological connections grounding our patterns of egoistic concern, then little would change in this Chapter. The chief reason for using personal identity is pragmatic. Most of the literature I will use elsewhere in this thesis, both from philosophy and psychology, uses personal identity to mean the psychological properties that define one's self, and these do come in degrees. For those who hold the view that personal identity is not what matters, assume that by personal identity I mean the continuous holding of the psychological connections grounding our patterns of egoistic concern. This pragmatic choice does not necessarily imply a loss of rigour; there are philosophers who defend the view we can continue to make use of the term personal identity (for a compilation of those views see Shoemaker, 2016). A few paragraphs below I will examine one example of each opposing view and observe that both

conclude that the realisation of individual interests is partially undermined by drastic changes in psychological capacities, thus strengthening my argument.

There seems to be no indication that moral status enhancement will interfere with the continuous flow of consciousness or episodic memory,[58] so two of the proposed psychological connections that ground egoistic concern would be preserved. But there are also reasons to suppose that other grounds for egoistic concern, such as other modes of psychological connectedness, would be undermined. The targeted psychological capacities of moral status enhancement are deeply implicated in forming psychological connections in one's life. For instance, higher-order desires, moral conformity, and social preferences for co-operation are all used to define one's identity and are possible targets of moral status enhancement. The fact alone that the products of moral status enhancement would cease to belong to the class of persons is an indication that we should suspect that their personal interests would not remain intact.[59] That they would belong to a new class of beings is also what seems to support the intuitive appeal of questioning why current persons would count a harm to persons as equal to a benefit for supra-persons.

The changes we would have to bring about in order to create supra-persons are assumed to be so extreme that the feasibility of moral status enhancement is often put into question. For instance, Buchanan (2009) asserts that the changes we would have to bring about are inconceivable, supporting his contention that we are unlikely to enhance moral status beyond that of persons. Even Agar, who wishes to argue that we can create supra-persons, has at times conceded that this might be an inconceivable change, though this *prima facie* inconceivability should not be grounds to assume it is impossible. McMahan (2010)

[58] Although it is likely that moral enhancement would alter the interpretation of past memories if it causes any significant changes to personality traits.

[59] In order to pose this issue more rigorously, one would have to define, within the broader class of individual interests, two stricter classes of personal interests and supra-personal-interests which would apply only to persons and supra-persons, respectively. The intuitive appeal of the question then would be the fact that moral status enhancement would be so radical we would have had to create such new classes.

mentions the possibility of increasing the various sorts of capacities significant to moral status in conjunction in such a way that this would produce a new emergent capacity to bestow upon those enhanced with higher moral status. Hence, if we are to produce supra-persons at all, it would have to result either from a radical form of enhancement, or from the conjunction of various enhancements that leads to a currently unseen morally significant capacity. Thus, it seems the intuitive appeal of my question could be well grounded. The unenhanced persons could share significantly more psychological connections with their past selves than the supra-persons would share with their past unenhanced selves.[60] I will, therefore, concentrate my arguments on the plausible, but not inevitable, possibility that moral status enhancement would produce significant damage to personal identity – that is, to the continuous holding of the psychological connections that ground our patterns of egoistic concern. Moral status enhancements that do not significantly affect personal identity would escape most of my objections in this Chapter.

McMahan (2002) offers a fruitful framework to disentangle further the balancing of individual interests when there are changes in the degree of psychological connectedness. He contends that the strength of someone's present interest in a certain future depends not only on her well-being if that future obtains, but also on the degree to which she would be psychologically related to herself in the future; he calls this her time-relative interest. One must consider not only how good one's life will be, but also the degree to which the individual enjoying that life is related to their own self. He is not alone in arguing that the degree of psychological connectedness matters in itself. David Lewis (1983) has also defended a similar but perhaps stronger view, wherein the moral significance of personal identity is translated entirely in terms of psychological continuity.

[60] I do not mean the unenhanced would necessarily share more. If proper care is taken and the change is gradual, supra-persons might share even more connections with past persons by virtue of having more mental states to connect with and because those changes could be intentional.

Instead of accepting a deflation of personal identity into stricter forms of psychological connectedness, then grounding our other intuitions of egoistic concern on other forms of psychological connectedness, Lewis asserts that we can continue to capture our all-or-nothing intuitions regarding personal identity if we let facts about identity depend on a further terminological fact. That fact would be the minimal degree *x* of psychological interconnectedness that could predispose us to claim that a set of time-slices belong to the same person – a fact that would then pinpoint what we mean by sufficiently similar when referring to some psychological connections. Lewis sees personal identity as the property two continuant persons hold between each other if, and only if, their corresponding time-slices belong to the same maximal set of psychologically interconnected (to degree *x*) time-slices. In other words, there is a one-to-one mapping between the set of identical (to degree *x*) continuant persons and the corresponding set of time-slices whose psychological relatedness have a greatest lower bound *x*. There is no survival whenever no latter slice is psychologically connected to at least degree *x* to former stages and hence no latter constituent person holds the property of personal identity with any former constituent person, i.e. that person ceased to exist. More importantly, disruption in the form of abrupt decreases in the degree of psychological interconnectedness (e.g. a severe stroke) is seen as a greater loss of what matters than a gradual change over a wide range of time (e.g. Methuselah's 900-year life-span). The first is closer to sudden death; the second is merely longevity. For Lewis, the degree *x* of psychological interconnectedness would track the degree to which someone should be egoistically concerned with her future self. He attempts to define a form of psychological continuity that captures both everything that is at stake in cases of life and death and all the significance we attribute to personal identity. McMahan seems more inclined to consider some individual interests as time-relative in the sense of being tied to the person's psychological unity during the time she held that interest. Either way, both

frameworks support a discount rate on individual interests (or egoistic concerns) proportional to the degree of loss of psychological connectedness at the time the interest is realised.

Suppose John wants to undergo moral status enhancement and Jim does not. If John performs the enhancement, Jim's individual interest to not be harmed by losing relative moral status[61] will be fully frustrated while John's individual interest to gain relative moral status will be only partially fulfilled. This is because the gains will be enjoyed by the resulting Supra-John (John after substantial moral status enhancement), and John's individual interest will be fulfilled to a discounted degree proportional to the loss of psychological connectedness. Therefore, ceteris paribus, there will be more frustration than realisation of their individual interests. When balancing the level of fulfilment or frustration in the individual interests of those considering moral status enhancement, we should not count the benefits to supra-persons to the same degree as the harms to persons.

McMahan (2002) himself considers the implications of applying his framework to radical enhancement:

> "Imagine the possibility of becoming vastly more intelligent and developing a vastly richer and deeper range of emotions, including emotions of which one cannot now form any conception. One would be as different from oneself now, in terms of psychological capacities, as one is now from a dog (or, more to the point, as different from oneself now as a dog would be from itself if it were to become a person). One would be, in short, so utterly psychologically remote from oneself as one is now that one may now have little or no egoistic reason to want to become that way" (p. 231).

Massive increases in intelligence, emotional richness, and emotional depth could lead to differences as significant as the ones currently separating humans from non-human animals, and seem to be what would be entailed by the types of enhancements necessary to bring about a supra-person. As McMahan (2002) observes, if we were to enhance a dog so that it would become as intelligent as a human, such a radical process would be likely to

[61] I am not suggesting this interest is the motivation for Jim not wanting to enhance. Supposedly, most people have an interest to not be harmed by losing relative moral status.

destroy its identity substantially, in such a way that even if its life would become vastly better, we might not be able to say this was in its interest to become as intelligent as a person. Likewise, greatly enhancing a person's moral capacities so as to produce a higher stratum of moral status might cause a better life than not enhancing, but for whom? Such a life might be so radically different that it would not be in this person's interest to become such a being. Moreover, when compared with traditional means of self-improvement, moral status enhancement would not only cause undesirable decreases in psychological connectedness but also would lead to an even greater loss of psychological continuity, given that it would presumably work in a much shorter time-frame. Moral self-improvement through technological modification is more disruptive than currently available forms of moral education. The trajectory of a racist who through reflection, life experiences, and moral deliberation gradually starts to care equally about people from different ethnicities contains a richness of intimately connected stages. Taking a certain combination of neurochemicals to achieve the same result might contain only a handful of stages; the racist is suddenly replaced with his non-racist version leading to a bigger loss of psychological continuity. Arguably, it might be better than continuing to be a racist, but worse – ceteris paribus – than abandoning racism through gradual moral education.

According to McMahan (2002), we must consider the loss in psychological continuity when enhancing a cognitively limited foetus in order to achieve cognitive normalcy in adult life, discounting the foetus' interest in having a normal adult life due to the psychological differences between a foetus and a normal adult. For analogous reasons, we should also discount a person's interests being realised by becoming a supra-person.[62]

[62] Some might argue that since we should still bring about normal adults even if it does not benefit the foetuses, we should also bring about supra-persons. However, this ignores the fact that the correct analogy is between the foetuses' interest in being replaced and our interest in being replaced; both should want to preserve their identities. The disanalogy would be that foetuses may not have enough cognitive complexity to have interests about the future.

Certain interests can only be said to be wholly fulfilled if the bearer of the interest comes to enjoy the full realisation of that interest in the future. It should be noted that while not all individual interests are of this kind, a significant amount of them are. John's interest in living in a more co-operative and efficient society can be wholly fulfilled by John living in such a society; it can only be somewhat fulfilled by supra-John living in such a society. In the case of radically enhancing persons until they become supra-persons, it seems that the future unenhanced persons who would be harmed would share more psychological continuity with those past persons who decided to pursue (or with those past persons who decided to allow) the enhancement of moral status than they would with the benefited supra-persons created by those past persons, therefore rendering this creation undesirable from current persons' perspective.

The moral status enhancement advocate might bite the bullet on identity change, claiming that although one would become a different individual by becoming a supra-person, the interests of those future supra-persons are still being represented by the interests of current persons who wish to become like them. However, if supra-persons' interests are being represented by present persons who wish to become supra-persons, this representation should decrease the force of those interests. If Alice wishes to become like non-existent Hannah, this does not imply that we should count Alice's interests as if they were Hannah's if she were to exist.

4. Replacing persons

Instead of claiming that the sacrifice of persons to benefit supra-persons is necessarily morally permissible, or that the gain in moral status to supra-persons outweighs the loss to persons, the moral status enhancement advocate can respond that we need not sacrifice or harm persons at all; we can simply replace them.

Replacing persons – We engineer beings with significantly higher cognitive, emotional and moral capacities, which as a consequence have supra-personal moral status. We intentionally make them mature slowly enough that they would only achieve supra-personal moral status once all currently living persons, who have been willingly sterilised, are dead.

This possibility would resist the objection that we should not cause our own deaths and would seem to be immune to both the arguments I presented against the two previous possibilities. However, while this third possibility eliminates direct harms to existing persons and continuity loss at the individual level, it does not prevent continuity loss at the generational level. I will argue this loss matters from the perspective of the shared individual interests of humankind in there being continuant generations.

4.1 Value fulfilment and psychological connectedness

The interest that current persons have in a more co-operative and efficient society can only be wholly fulfilled by beings sufficiently like us (even if they are from future generations) and not by the product of drastically enhancing persons. Current persons are all human beings. Their combined individual interest in there being future generations can, therefore, be framed as the human interest in there being future generations. [63] Of course, if some of our interests can only be achieved through our replacement by radically enhanced persons, then perhaps this is the best we can achieve in respect to the fulfilment of those interests. But it is still not clear that humankind's interest in a more co-operative society can only be achieved by the radical forms of enhancement implied in the enhancement of moral status. Additionally, suppose the most favourable case for the creation of supra-persons; that is, that mere persons will absolutely never be able to co-operate more and that replacement

[63] I do not wish to mean the biological species *Homo sapiens* here, but a more culturally enriched conceptions of humans.

with supra-persons is the only way to achieve our interest in a more co-operative society. Under this circumstance, every interest that is sensitive to who is fulfilling it – without requiring the creation of supra-persons – will forever only partially be fulfilled because of the discount rate from the loss of psychological connectedness. The risk is not only that the individuals enjoying the benefits of radical moral enhancement will not, to a relevant degree, be the same kind of individuals who wished for those benefits, but also that every other interest held by those individuals will no longer be enjoyed by their kind. As has been observed by Scheffler (2018), the realisation of our values depends not only on the right state of affairs to obtain but also on the existence of the right kind of evaluators to enjoy that state of affairs in the first place. Producing a future with superintelligent ultra co-operators might result in a risk-free society with an abundance of economic wealth and technological capability but would also be void of the types of individuals who originally wished for such a world.[64] Perhaps supra-persons will be able to appreciate such a world even more than persons could, but it seems to matter for current persons that no person (or being of his kind) will be able to enjoy the wonderful results of moral status enhancement.

Whether in the case of enhancing current persons so that they become supra-persons or in conducting generational replacement of persons by supra-persons, there is a loss connected with the interests of persons being fulfilled by radically enhanced persons who would have radically different moral judgments, emotional profiles, and levels of cognition, thus having significantly different identities from the persons who held those interests. For solely impersonal interests this may not matter, but for individual interests that are relative to a specific individual enjoying an outcome, or a specific kind of individual in the future

[64] A similar but more restricted point is made with regards to consciousness in Bostrom (2004) "There may be an abundance of economic wealth and technological capability in such a world, yet it would be of no avail because there would be nobody there to benefit from it."

enjoying an outcome, attempting to fulfil them by means of radically changing the person or group of persons who had those interests can be self-defeating.

Suppose that the far-future contains at least a considerable amount of potential interest-realisation and this realisation depends on the existence of a certain kind of evaluator. In this possibility, if some form of preference utilitarianism were right, it would give us a good reason to bet that our annihilation at the hands of supra-persons would not be such a great moral loss, and would instead represent our salvation from Persson & Savulescu's ultimate harm. If the sole source of value in the world is interest fulfilment or sentient pleasure, then it seems our fear of annihilation at the hands of much more efficient value fulfillers is some form of prejudice. At the individual level, this form of prejudice would manifest itself as our desire to continue existing and our right to do so even if our death would somehow allow resources to be spent more efficiently elsewhere. But it also would be absurd to suppose that someone loses the right to self-defence merely because he is sure his killer would thereby be able to allocate resources more efficiently. In the same manner, but perhaps to a lesser degree, it seems absurd to suppose we lose our right for there to be future generations of our kind simply because another type of being would be a better interest-fulfiller. In the same manner as an individual has the right to defend his life against a more efficient resource allocator, humanity has the right to defend its continuity against more efficient resource allocators who are not human continuants. If these beings are better at fulfilling our interests, then perhaps we would have grounds to want to be replaced by them, but this cannot be in the case of our interests that are dependent on the kind of being that we are existing in the future; these will always be frustrated by our extinction.

There is a great deal of uncertainty and vagueness in defining the morally relevant class of "the kind of beings we are" in this context. The existence of any kind of future persons does not seem to be either a sufficient or necessary condition for the complete

fulfilment of our interests. The existence of any kind of supra-persons does not either, even supra-persons who share many of our interests. The existence of humans leading lives of the highest possible value, on the other hand, does seem like a sufficient condition for the complete fulfilment of our interests; however, we seem to be incompatible with this realisation in many dimensions, not least because of our moral failings. Perhaps, then, the existence of some kind of supra-persons is a necessary condition for the full realisation of our values.

The problem lies in that there is nothing inherently human in supra-personhood. A world of supra-persons might consist of merely highly efficient and co-operative beings. For instance, Bostrom (2004) considers that the future might consist of ultra-specialized and highly co-operative agents, each dedicated to a task currently executed by only one of our many cognitive modules. Such a possibility might contain more value in the form of interest realisation and would be bound to contain more co-operative behaviour as agents would depend highly on each other, but it also would lack many of the inefficient human activities we consider intrinsically valuable. There is an important question of what and how much would be missing from a future consisting of beings of completely alien consciousness and high levels of well-being compared to a future containing human beings who share the same high levels of well-being. I do not intend to try to respond to this question here; for my purposes it seems sufficient to note that there is a strong intuitive case for there being something missing. A future with human continuants with extremely high levels of well-being seems better than a future with aliens who have that same high level of well-being. If there is really a choice to be made between enhancement or extinction, as contended by Persson & Savulescu, I argue we should prefer forms of enhancement that preserve some continuity of our shared individual interests. The continuation of our human values does not seem to be entailed by creating radically superior people. There would have to be a careful

preservation of the shared individual interests of humanity when carrying out moral status enhancement in order for supra-persons to fulfil our interests.

4.2 Subjectivist arguments

Although I have argued that the continuity of human values matter, the reasons I have presented might all still be merely individual interests shared by the currently living human beings and as such might not be impersonal reasons. I believe this reason alone should make us wary of the *replacement* case. However, we should note that for a subjectivist the fact that they are merely individual interests would not mean they could not be moral reasons.

For instance, for a cognitivist subjectivist[65] who holds a dispositional theory of value – such as the one explored by Smith, Lewis & Johnston (1989) – values are defined based on one's dispositions in ideal circumstances. According to this theory, valuing something is defined as desiring to desire it, or higher-order desires.[66] Moreover, under ideal circumstances, valuing something will lead to a chain of implications that leads one to pursue it as effectively as possible in the absence of conflicting values. For someone holding this kind of theory, it would be enough to demonstrate that if under circumstances near enough to ideal, people are disposed to act in a manner that reveals that they value the continuity of human generations. I am convinced that if presented with a *replacement device* similar to the extinction device I mentioned earlier but that would put in place the *replacement* possibility, many people would not, after careful reflection, activate it.

One need not take my conviction concerning such an empirical fact to conclude that we value the continuity of human generations. Scheffler (2018) has extensively argued for this disposition. He subscribes to a dispositional theory of value whereby valuing something

[65] One who believes moral statements are either right or wrong but that what makes them so are not objective, independent, moral facts.

[66] For instance, a repenting drug addict desires the drug but does not desire to desire the drug. He values being addiction-free and thus desires not to desire it even though he desires it.

involves either attachment, investment, or engagement with it, all of which involve some kind of emotional vulnerability to its loss, and a disposition of having reasons for action which one would not otherwise have. He asserts that if one were to become aware that humanity would be extinct after one's death, this would compromise one's capacity to lead a good life more than the awareness, which most people already have, of one's own mortality. There seems to be a greater attachment to the survival of humanity than to our own personal survival. Scheffler explores four different reasons why we value the continuity of human generations within that framework; reasons of love, interest, value, and reciprocity. Many human beings would react to the prospect of the extinction of humanity with great sadness and despair, and, likewise, with great joy to the prospect of continuous human flourishing into the indefinite future. Thus, we display similar reactions to humanity as we do to any other thing or person we love. Besides this emotional attachment to humanity, most people would find that their own lives and activities would lose meaning and value if they believed that humankind would not go on after their deaths. Therefore, we also have a self-interested reason for wanting humanity to continue to flourish because our flourishing partially depends on believing it will. Moreover, the disappearance of humankind would also mean the demise of all the things we value and whose existence depends on human activity. Intimate and fulfilling personal relationships, music, dance, and science would all cease to exist. Not only that, but the beings who could enjoy those things would also cease to exist. The extinction of humanity would mean the disappearance of both valuable things and of the act of valuing itself from Earth. Hence, we have reasons of value to care about the future of humanity. Finally, given that the value of our current activities partially depends on our confidence in humanity's survival, our present flourishing depends on their future flourishing. Additionally, if we were to believe they would survive but would also lead lives we find despicable, then this too would compromise our current capacity to lead valuable

lives. This mutual dependency creates a form of reciprocity with future generations wherein as we make sure to safeguard their flourishing so that they will be able to lead lives we consider valuable we thereby safeguard our own flourishing. Thus, we have reasons of reciprocity to care about future generations. As Scheffler observes, people will have these reasons and corresponding dispositions to various degrees. Some might not be affected at all by the knowledge of human extinction as long as it happens after they are dead; others might find the replacement device more terrifying than the extinction device so they would live agonizing lives under the awareness humankind would soon end instead of suffering a painless death.

Like Scheffler, I believe our dispositions towards the future are complex and rich, and I also believe that there are likely to be several other reasons behind our valuing of the continuity of humankind. The reasons explored by him all reveal a common thread that I suspect is shared by many, though not all, in our valuing of the future. They are all reasons that depend on the future existence of human beings or beings like us in some relevant dimension, whose full specification lies outside the scope of this thesis. It seems clear, nonetheless, that beings of superior moral status will not necessarily be sufficiently like us to be considered human continuants and thus fulfil the role that future human generations have in any of the four reasons presented in the last paragraph. As Scheffler states, "…the importance to us of future generations lies partly in the fact that they are our *successors*, that their existence extends the chain of generations in which we ourselves are participants" (p. 16). Supra-persons might not necessarily be our successors. Many people would display no emotional attachment to supra-persons if they were sufficiently different from us and this kind of attachment is fundamental for our concern over the future, according to Scheffler.

Williams (2008) has argued that human beings form the morally relevant class of beings we should care about and preserve because "they are us," a reason he calls an

inescapable human prejudice without which we would be hard-pressed to justify our moral intuitions. Therefore, we should fear our replacement by morally superior aliens or morally enhanced humans. Savulescu (2009) has raised several criticisms of Williams' view, some of them targeted at the fact that this human prejudice is no different, and no more justifiable than any other form of prejudice, like racism. Any white supremacist could argue he is right in valuing whites more than other races because, to him, "they are us" as well. Williams would counter that humans seems to matter more in their own right than other groups matter in their own right. "Because they are humans" has greater force than almost any other independent justification of the class "Because they are *x*". But Savulescu's response is that one would only have morally valid reasons to be partial towards members of one's group if this group shared a morally significant property justifying this partiality. The human species has no such property. Anencephalic newborns,[67] for instance, can be considered members of the human species but lack any of the proposed morally significant capacities such as consciousness and complex reasoning. Savulescu suggests that persons are a better candidate for justifiable partiality given that personhood often involves possessing some of the proposed morally significant capacities. I expect he would also include supra-persons in this group given they would, by definition, possess even higher morally significant capacities.

It seems clear that the human species, and any species-based classification, lacks intrinsic moral relevance. Being a member of the human species in no way guarantees the possession of any morally significant capacity. I doubt, however, that Williams would continue to identify human beings with the human species if pressed to scrutinize his view. I imagine he would prefer to mean a much more culturally than biologically charged definition of human beings. Regardless of his intended meaning, human beings certainly can

[67] One should note that one classic definition of species membership is determined by the capability to generate viable offspring in natural conditions. Anencephalic newborns can be argued to lack such capability.

be defined as a more morally relevant class than human species. For instance, we can define human beings as beings capable of human flourishing or beings capable of fulfilling the role of human continuants in Scheffler's four set of reasons. A future with an advanced civilisation of genetically modified human beings is a better prospect than a future with permanently non-conscious *Homo sapiens*, even if the former and not the latter, could no longer be members of the human species. I suspect Williams would be more inclined to call the former human beings than the latter.

But suppose we reject the idea that there can be a sensible definition of human beings that would give an objective moral reason to be partial towards humanity; Savulescu has conceded that we could still have valid reasons to be partial. For instance, a father might have special individual interests in saving his child rather than a stranger from drowning, even though he might not have any moral reason to do so. His partiality towards his children is based on a special relationship, perhaps based on family membership. Savulescu has equally conceded deaf parents can have valid reasons to select for a deaf child even if being deaf leads the child to be worse off. Likewise, one could claim humans stand in a special relationship with all other humans and hence have reasons to be partial towards human beings by not activating the replacement device.

4.3 Objectivist arguments

Suppose now that one believes all moral reasons should be impersonal reasons. As I mentioned, one would still have to concede that individual interests do offer a basis for action and hence reasons against replacement. Additionally, I believe there is one argument for continuity at the population-level being valuable in impersonal terms.

Consider the Period Independence assumption, which holds that:

> "By and large, how well history goes as a whole is a function of how well things go during each period of history (…), the extent to which it makes history as a whole go

> better or worse is independent of what happens in other such periods" (Beckstead, 2013, p. 59).

If Period Independence holds, moral value cannot be cross-temporally dependent. If value is cross-temporally dependent it means that value at t_2 could be affected by t_1, independently of any causal role t_1 has on t_2. The same event X at t_2 could have more or less moral value depending on whether Z or Y happened at t_1.

For very large periods of time, it seems there is no cross-temporal dependence and Period Independence is intuitively true at that scale. However, it cannot, and never was intended to be true for periods as short as one's lifetime, given that most of us would ascribe value to psychological continuity. If psychological continuity matters, then however many psychological connections hold between time-slices also matters, not just how much value each time-slice independently has. Suppose Steve lives now at t_1. If suddenly in t_2, Steve is replaced by a radically different alien individual with the same amount of value as Steve would otherwise have in t_2, then t_2 may not have the exact same amount of value as it would otherwise have, simply by virtue of the fact that in t_1 Steve was alive and the alien's previous time slice was not. One thousand individuals with one-day lives do not seem to amount to the same value as a single individual living through one thousand days.

Now, suppose a super-intelligent alien race were playing God with a planet. They consider two options for that planet. They could divide the next ten millennia into one hundred periods of one century, and place in each period one infertile generation with the same overall total and average momentary well-being, with each generation being entirely unaware of and unrelated to the other. Each lasts for one century and then is gone forever. As a second option, they could pick one of those generations and enable it to reproduce and continue living in that planet for ten millennia, but amounting to the same constant levels of momentary well-being. Should the aliens be indifferent regarding these options? The option with many continuant generations that flourish across ten millennia – albeit with constant

momentary well-being – seems better than the option with isolated generations. The value of a given generation seems to be affected by whether or not there will be future (or there were past) continuant generations. If this intuition is correct, then we have impersonal reasons to prefer a very gradual replacement, possibly spanning many generations, of persons by supra-persons.

This would mean that there are certain individual reasons that, when accounted for in terms of impersonal reasons, retain their intrinsic cross-temporal dependency originating from the value of continuity at the individual level. Perhaps continuity is an intrinsic aspect of being an individual and hence individual interests are intrinsically sensitive to the loss of psychological connectedness. Bostrom (2012) and Omohundro (2008) have independently argued that value-coherence, the stability of an agent's value structure, can be a universally shared instrumental value of all sufficiently intelligent agents. This would help explain why continuity is a basic objective value. I used the term momentary well-being to distinguish the conception of well-being in play from possible conceptions of well-being that attempt to take into account cross-temporal effects. These richer conceptions would have to allow matters related to having a good life over extended periods to affect how well things are going in that life at a given moment inside that period. Establishing this relationship, however, is not a trivial task because having a good life might depend on specific psychological connections between moments and these connections can only be determined to exist when looking at extended periods. For instance, having a good life might depend on whether one lives a life of improvement or decay and even on narrative relations as proposed by Dorsey (2015) (see also Velleman, 1991).

There is one response that would avoid most of my objections. I have been basing my arguments on the moral significance of psychological continuity, either at the individual

or generational level. I argued it is plausible sufficient moral enhancement will harm psychological continuity. But it will not necessarily do so. Both the objectivists and subjectivists would accept a very *gradual replacement* without abrupt changes leading to loss of continuity. Eventually, a moral status enhancement advocate could propose the possibility of gradual replacement, which none of the arguments presented here would oppose. Supra-persons can be created by gradually enhancing our psychological capacities throughout a long period of many generations. So many that the difference between them would not be more significant than the difference we would expect without enhancement. The arguments here make it clear that gradual replacement is a superior alternative to any of the three possibilities listed before. The previous literature on this subject overlooks issues related to psychological continuity and thus fails to reveal the higher desirability of gradual replacement.

5. Conclusion: consideration of individual interests leads to increased risks

Any kind of drastic enhancements, even when impersonally desirable, may undermine certain individual interests due to being detrimental to psychological continuity. A person's psychological continuity is preserved during a period – i.e. the person survives during the period – whenever all neighbouring pairs of time-slices of that person during that period preserve a certain minimal amount of psychological connectedness between themselves. The level of psychological connectedness a person's time-slices hold with each other is a function of how many and how strong psychological relations are between them.

It is not wise to attempt moral status enhancement while being exclusively guided by the question of what sort of value we want more of in the future, because we risk forgetting that whatever sort of evaluators, i.e. whatever sort of people, we want in the future can also be relevant. I contend it would not only be relevant, but safer, to ask the latter question.

Sufficient moral enhancement with an immediate eye on creating whatever dispositions realise that we think has value risks extinguishing the evaluators, and they might be the only ones able to tell if those values were mistaken. It will be safer, not to mention more pragmatic and intuitive, to centre the question around what sort of people we want to become using our current moral traits as a starting point. Certain defences of moral status enhancement fail because they do not account for interests whose realisation depends on the existence of the same individuals, or type of individuals, who held those interests. At the individual level, one must consider that there is a loss of psychological continuity after substantial moral enhancement, and at the societal level one must consider the risk of creating beings that are not human continuants in the relevant sense.

Chapter 4: The need for deep moral enhancement

1. Introduction

I will argue that the type of harm deep moral enhancement could help prevent – namely, extreme risks – provides a strong reason to pursue it. However, considering the possible sources of risks investigated in the second and third chapter, deep moral enhancement is also a source of extreme risk, which provides strong reasons against pursuing it. I will thereby outline a conflict in the project of moral enhancement which I will aim to solve in the next and final Chapter of this thesis.

In the preceding chapters, I have used extreme risks as a roughly defined umbrella term in order to describe risks that could harm human well-being on a global scale,[68] risks that could cause the end of intelligent life on Earth,[69] or other risks of comparable negative moral value. I have mentioned two sources of extreme risks throughout this thesis. Firstly, I briefly mentioned extreme risks arising from co-operation problems and moral failings in general that moral enhancement is proposed to solve (Persson & Savulescu's ultimate harm). Secondly, I explored in more detail the possible extreme risks that could arise from deep moral enhancement.

In the following section, I will summarise some arguments regarding the moral relevance of these risks. I aim to present the theoretical arguments for the moral relevance of these risks in order to strengthen one of the main motivations of my thesis: assessing and presenting solutions for the extreme risks of deep moral enhancement, while also reinforcing

[68] Global catastrophic risks, as defined in Bostrom & Cirkovic (2008): "a risk that might have the potential to inflict serious damage to human well-being on a global scale" (p. 1).

[69] Existential risks, as defined by Bostrom (2002): "Existential risk – One where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential. [...] For there to be a risk, given the knowledge and understanding available, it suffices that there is some subjective probability of an adverse outcome, even if it later turns out that objectively there was no chance of something bad happening. If we don't know whether something is objectively risky or not, then it is risky in the subjective sense. The subjective sense is of course what we must base our decisions on" (p. 2)

the moral importance of a widespread reduction of these risks whatever their sources may be. Then, a subsequent section will introduce the sources of extreme risks mentioned in the literature, as well as the sources surveyed in this thesis. I aim to list the sources of these risks in order to later clarify the fact that most of them either arise from or require overcoming the moral failings that deep moral enhancement is proposed to solve, thus strengthening the case for deep moral enhancement. I will also observe, as argued throughout this thesis, that this proposed solution is itself among those sources if safety concerns are ignored. Next, I will clarify how the most frequently proposed aim of moral enhancement, namely improving large-scale co-operation, would produce a widespread reduction in extreme risks. Moreover, I will consider the objection that traditional forms of moral progress are an effective means of fixing our moral failings and addressing extreme risks. I will concede there are cogent arguments showing traditional means have some effect, but I will argue they do not support dismissing technological interventions.

The tension that this Chapter aims to expose will thereby become clearer. Deep moral enhancement is both extremely morally desirable as a widespread solution to a large class of problems with unusually high negative moral value but also extremely dangerous as a source of these same problems. The next Chapter proposes to solve this tension by steering the development of deep moral enhancement away from the extreme risks I have surveyed. There, I will develop the desiderata of a safety framework for deep moral enhancement, argue that such a framework is likely to be a form of virtue theory and then compare available frameworks.

2. Extreme risks and their moral relevance

The moral value of the future is crucial in evaluating the desirability of deep moral enhancement. For instance, if we should primarily be concerned about the interests of people

alive today and not the interests of distant future generations – as some philosophers with person-affecting views in population ethics hold – then a moral enhancer with immediate positive value as well as huge far-future negative value might be considered desirable. On the contrary, if one does not consider anyone – future or present – to be less important, then we should probably focus solely on the far-future value when evaluating this technology. Excluding person-affecting views, one could commit to simple marginal discounting or time caps in the future.[70] However, in this section I will agree with Parfit (1986), Bostrom (2013) and Beckstead (2013) and argue that defensible positions in which the value of the future does not dominate the assessment are hard to find. This will offer an additional strong justification for having conducted a careful risk assessment of deep moral enhancement in the preceding chapters. On the other hand, given that deep moral enhancement would be likely to reduce a wide range of plausible extinction scenarios, this same consideration will also provide strong reasons to pursue it if we can sufficiently address the severe risks that could arise.

At the end of *Reasons and Persons,* Derek Parfit (1986) asks us to compare three outcomes for mankind: peace, a nuclear war killing 99% of the world's existing population, or a nuclear war killing 100%. He contends that the difference between the third and second scenario is, counter-intuitively, much greater than that between the first and second. The complete extinction of humans would mean the destruction of all future humans expected to flourish on Earth for at least another billion years in which the planet remains habitable. Killing 99% of currently living humans would have more restricted consequences, as it would not curtail the continuity of humankind on Earth. Nick Bostrom (2002, 2013) has developed this argument in more technical detail, arguing that even conservative estimations

[70] For marginal discounting see Beckstead, 2013, pp. 159-162. For time caps see see Temkin, 2012, pp. 238-262.

in which technological development remains stagnant would project that our future will contain a total of 10^{16} human lives. Assuming continuous technological development and the colonisation of other planets, he estimates 10^{52} human lives. Whether we assume those lives will be at least as good as ours, and use aggregation over lives or some other theory that ascribes value to future human flourishing – or things that depend on human lives – it seems most of the moral value lies in the future, and curtailing that possibility has an unusually extreme negative value. Bostrom concludes that this value is so great that it should dominate our normative thinking, and that the reduction of our risk of extinction should take moral priority over other actions. Beckstead (2013) has made a more modest argument, relying on the simple assumption that the moral value of a history containing large periods of time is a function of independently aggregating these periods. Given that the far-future contains so many more of these periods than the present, most of the potential moral value is in the far-future. Beckstead extensively argues that even if the periods cannot be fully independently aggregated due to a limit in the total amount of good periods that matter, or diminishing value as good periods are added, then the far-future would still contain most of the potential value for reasonable limits or discount rates. Discounting the consequences for future persons at reasonable rates would also have the same effect; it is intuitively easy to see this considering what sort of discount would have to occur so that 10^{16} are outweighed by the current 10^{8}. Beckstead concludes that determining how well the far-future goes has overwhelming moral importance.

One of the only possibilities in which Beckstead concedes the far-future could not have overwhelming importance is one in which we are to take a hard-line position on the moral significance of future people, claiming that whatever happens to people who do not currently exist and might not come into existence should be outside of the scope of moral deliberation. Suppose then that we were to make such a concession. As has been explored

in Chapter 3, Scheffler (2018) arrives at a similar conclusion on the value of the future of humanity by being mostly concerned with what present humans are disposed to value. Each of these lines of argument comes to different conclusions about the value of the future, and each departs from different and general background assumptions, but they also conclude that the continuity of humankind has an extremely high moral relevance. If one were uncertain about these background assumptions, then the subjective probability of their conclusion being right could not be low enough to render it a trivial matter; otherwise, it would not be uncertain. It is safe to assume then, that the value of a proposed radical technological intervention ought to be heavily influenced by its negative and positive impact on the long-term future, with even more special attention directed to eventual scenarios threatening to extinguish the possibility of future value by means of annihilating humankind.

Suppose that future people do not matter morally, and that Scheffler along with whoever finds his arguments cogent is wrong about what human beings are disposed to value. In that case, present persons would be our only concern, so whatever inclination they had towards future people would have to be dismissed. Beckstead, Parfit and Scheffler's conclusions would not obtain, but one slightly altered version of Bostrom's arguments still would. If we only cared about the seven billion humans currently living, their annihilation would be one of the greatest possible harms, and thus we ought to give priority to the reduction of existential risks. It is even safer to assume then, that the value of a proposed radical technological intervention ought to be heavily influenced by the consideration of scenarios in which it brings about our extinction or a global catastrophe.

There is also a series of biases that all seem to lead us towards underestimating or lacking the appropriate response to extreme risks. The probability that a civilisation has previously encountered an extinction event is always zero, irrespective of the probability of such an event happening in the future. Therefore, in contrast with other risks, we cannot

learn from experience or use history to inform us of the likelihood of such events. Our natural tendency to do so is likely to lead to a severe underestimation of such risks, leading to an anthropic bias (Cirkovic, Sandberg & Bostrom, 2010). As Bostrom (2013) concludes:

> "The reactive approach – see what happens, limit damages, and learn from experience – is unworkable. Rather, we must take a proactive approach. This requires foresight to anticipate new types of threats and a willingness to take decisive preventive action and to bear the costs (moral and economic) of such actions" (p. 27).

Many empirical studies have found that our response to a harm does not scale proportionally to the number of people affected. For instance, one study found that if the declared deaths from chlorinated drinking water were increased by a factor of 600, participants would be only willing to spend four times more in order to prevent this level of risk (Carson & Mitchell, 1993). Other studies have found that humans tend to be irrationally willing to undergo a constant low probability of an extreme loss in exchange for a constant high probability of a moderate gain. Participants were willing to make an investment that earns $10 with 98% probability but loses $1000 with 2% probability, a net loss mathematically but which psychologically looks like a good bet. Hindsight bias also affects the estimation of unlikely but extreme risks. If whenever a big unforeseen catastrophe happens and we both retroactively and mistakenly assume we knew it was going to happen all along, then we are always left with the impression that all big catastrophes are easily foreseen, and so we tend to dismiss any risks that are extreme but difficult to assess.

The far-future consequences of a new proposed technology are, minimally, an extremely morally relevant consideration; if not the most pertinent or so relevant as to dominate any ethical analysis. Several biases lead to underestimating the extreme risks that significantly decrease the chance that the far-future will contain things we value. Therefore, the impact of deep moral enhancement on extreme risks is of the utmost significance.

3. Extreme risks and their sources

3.1 Expected sources in the literature

In the book *Global Catastrophic Risks*, a variety of experts assess the main risks which could cause the loss of human lives on a global scale (Bostrom & Cirkovic, 2008). The book separates risks into three categories: nature, unintended consequences and hostile acts. I will make an overview of these risks following the book and clarify how the main solutions to each of them involve increased global co-operation. Later, I will explain how this would be helped by deep moral enhancement.

3.1.1 Nature

Super-volcanism

The closest that humankind ever got to extinction was possibly due to a volcanic super-eruption 75,000 years ago in Toba, Indonesia. Earth temperatures decreased by 5-15C for many years alongside other significant persistent changes to global climate. The reproducing female population is estimated to have dropped down to 500 individuals during this time. Most of the expected long-term catastrophic consequences of super-volcanoes are caused by drastic climate change that drives a loss of crop productivity, resulting in mass starvation. These eruptions seem to occur on average once every 50,000 years and not much can be done to prevent them. However, their most drastic effects would only last a few years; thus, we could prevent most of the harmful effects of low crop productivity by increasing the global grain stockpile. Co-ordinating such a costly increase in order to prevent an improbable but extreme catastrophe is a classic co-operation problem, which moral enhancement could help us address by increasing our ability for large-scale co-operation.

Comets and asteroids

One of the most well-known mass extinctions is believed to be the result of an asteroid impact, which annihilated 75% of all species from the face of the Earth, including all non-avian dinosaurs. Impacts with asteroids ten times smaller would also cause a global catastrophe, with the dust from the impact blocking sunlight for several years and the collapse of modern agriculture leading to mass starvation. Once more, a greater stockpile of grain would help mitigate the effects, but the largest risk reduction, in this case, would be increased spending on mapping the trajectories of near-Earth asteroids larger than 1 km. There is a good chance we would be able to deflect such an asteroid with previous notice. Again, investment in mapping near-Earth asteroids is a large-scale co-operation problem to address a global risk.

3.1.2 Unintended consequences

Climate change

Anthropogenic climate change is one of the most well-known global risks. The likelihood of a sudden increase in global temperature and sea levels is low, but there is high uncertainty in the models, and even a linear modest increase might have catastrophic effects by the end of this century. The main mitigation strategy is an international effort to decrease the emission of greenhouse gases, which once again relies on global co-ordination. Considering that this is the global risk that has received the greatest amount of attention, and that we still have failed to produce an enforceable global strategy for reducing greenhouse emissions, the prospects of having enough large-scale co-operation to tackle this as well as all other global risks listed here seem dim.

Pandemics

Infectious diseases have killed more people in the last century than both world wars put together. The Spanish flu has killed 20-50 million people and smallpox has killed well over 500 million (Koplow, 2003). There are 15 million deaths per year due to infectious

diseases. In the past, infectious diseases reduced the population of Europe by 25 to 60% (World Health Organization, 2017). Globalisation has made it possible for a disease to spread across the world within days. Even our flawed responses to relatively localised epidemics of a deadly disease such as Ebola have evidenced the need for proper global co-operation on the matter (Moon, 2015).

Artificial intelligence

Long-standing trends in the advancement of artificial intelligence seem to indicate that we will be able to achieve human-level artificial intelligence in the future. Once achieved, there is a reasonable argument for the notion that this artificial intelligence will be able to recursively improve its own intelligence, and that from then on exponentially high levels of above-human intelligence could be reached. Given that higher intelligence is what enables us to control the environment and non-human animals, the introduction of agents with significantly higher intelligence, who may not share our values, would be likely to bring about global catastrophes, to say the least. Countries are already rushing in an arms race to be the first to develop such a powerful technology, and safety concerns could be overlooked in the absence of an international agreement to prioritise safety in the face of rapid development. Any safety framework such as imparting human morality into an artificial intelligence would require global co-operation to be effectively adopted.

3.1.3 Hostile acts

Nuclear war and nuclear terrorism

Although most people would imagine the risk of a nuclear Armageddon dramatically decreased by the end of the Cold War, the reality is that the systems responsible for nuclear threat detection and response have remained the same, and various false alarms have been reported since the fall of the Soviet Union. The direct death toll of a US-Russia nuclear war would be upwards of 400 million people (Martin, 1982), with the possibility of a much

higher toll from the resulting dust creating the same type of climate change as was discussed with asteroid impacts. As far as the general public is aware, no terrorist group has ever gained possession of fissile material or an intact nuclear weapon; however, the possibility of a terrorist nuclear attack is still significant, with some specialists estimating a greater than 50% probability in a decade. Widespread substantial nuclear disarmament and radical anti-proliferation treaties would be the only way to significantly reduce these risks, but once again this depends on achieving a level of international co-operation we manifestly lack.

Biotechnology

Despite smallpox having killed well over 500 million humans across history until its final eradication during the middle of the last century, methods for recreating it are publicly available (Kupferschmidt, 2017). Compared with nuclear weapons, the means to synthesise a biological weapon are easier to obtain and the weapon easier to deploy. Necessary research into infectious diseases needs to be balanced against the risk of this research being weaponised. Global catastrophes might ensue as a result of just one rogue group of researchers either deciding to sell such deadly biotechnologies to the highest bidder – or foolishly making them publicly available. Biosecurity policies need to be extensively enforced, making this is contingent on high levels of large-scale co-operation (Hoffman & Behdinan, 2015).

3.2 A new source: deep moral enhancement

As I have explored throughout this thesis, deep moral enhancement can also be a source of extreme risks. I have contended that moral traits – the target of deep moral enhancement – are unusually complex, and prone to unexpected consequences. I use the aetiology of moral traits, their unusually high susceptibility to contingencies, and epistemic difficulties arising from the first-person view of human morality in order to support my contentions. Furthermore, I have argued that technological intervention on individual moral

traits will often lead to paradoxical effects on the group level and, finally, that in so far as moral enhancement targets motivation, it is prone to be self-reinforcing and irreversible. I have presented empirically informed concrete possibilities of apparently desirable changes in moral traits that may lead to catastrophic consequences, such as the increase of individual co-operation leading to a decrease in large-scale co-operation, which would be plausibly accompanied by a disruption of social order on a global scale. I have also argued moral status enhancement (a likely outcome of deep moral enhancement) could lead to losses of psychological continuity both at the individual level and generational level. Sufficient moral enhancement may lead to beings that no longer play the role of human continuants and to the loss of continuity in the chain of future generations.

Deep moral enhancement is therefore also a source of extreme risks. A sufficient decrease in our propensity to co-operate between groups could lead to a breakdown of international relations, wars, significant loss of economic output, or other scenarios causing significant harm to human well-being on a global scale. The loss of continuity of human generations could arguably amount to a loss of moral value comparable to a global catastrophic risk, because it would diminish the realisation of all interests depending upon the existence of human continuants. If we lose some capacity that is not only contributory but also necessary for the realisation of our values (e.g. sentience), then this loss would effectively amount to extinction. This sort of scenario is categorised as a "whimper" in Bostrom's seminal paper on existential risks:

> "A posthuman civilisation arises but evolves in a direction that leads gradually but irrevocably to either the complete disappearance of the things we value or to a state where those things are realised to only a minuscule degree of what could have been achieved" (Bostrom, 2002, p. 5).

4. Deep moral enhancement as a widespread solution

4.1 The argument for deep moral enhancement from extreme risks

I explicitly discussed how addressing all the global risks explored in the last section relies on large-scale co-operation between groups. Co-operation between groups on a global scale is one of the main proposed targets of moral enhancement and has been discussed frequently throughout this thesis. However, not only was large-scale between groups co-operation never a recurrent evolutionary pressure due to groups being small and scattered, but also some level of group competition was likely to have been selected for in our evolutionary past. In fact, studies have shown that groups behave in a more individualist and competitive fashion than individuals (Shi, 2014). Furthermore, most of these risks require both a global response to decreasing their odds in the form of globally enforceable regulations and strategies as well as a global response for if they were to materialise. For several of them, it would be enough for either one rogue nation or group of individuals to jeopardise the co-operative effort. Several of them are present risks that ought to be addressed as soon as possible and require a radical shift in our co-operative dispositions. For instance, they would require an unnatural disentanglement of individual co-operation and parochialism. Traditional means of moral education might not be either widespread, fast or radical enough to help address these global and existential challenges in any meaningful way. Moreover, taken as a whole the reduction of extreme risks can be seen as a public good of a global scale, whose benefits will never be properly observed. No one can directly observe the event of a risk that fails to materialise due to sufficient prevention measures. Therefore, co-operation problems exist not only for solving each particular risk but also for focusing attention on extreme risk itself as a worthwhile cause. Agents might feel safe to take no action given that the number of agents that could act is so high that one single agent

defecting would not be noticeable. Moreover, agents have no incentive to act given that if humanity is successful in preventing those risks, their actions might go unrecognised.

There have been other proposed widespread solutions to extreme risks, but several of them also rely on a level of global co-operation that we currently lack. For instance, one prominent solution is differential technological development, as proposed by Bostrom (2014, pp. 229-237). According to this strategy, humanity's focus should be not on developing technological capabilities faster, but instead on developing them in the right order. We should slow the development of potentially harmful technologies while accelerating the development of safe technologies. Increasing technological progress overall might not be necessarily beneficial if it means we develop technologies able to cause human extinction before developing means to control these technologies. But co-ordinating technological development on a global scale seems impossible, given our current levels of parochialism. Taking the lead on developing a powerful harmful technology would confer any group an immense strategic advantage over others. One past instance of this is the development of nuclear weapons by liberal western democracies. In fact, the differential technological development strategy would prescribe the development of technologies focusing on solving global co-operation problems before developing any other powerful technology that could be unilaterally used to annihilate a large share of the human population, i.e. the development of deep moral enhancement should take precedence. Co-ordinating the order of technological progress in synchronicity is unfeasible in the absence of efficient means of guaranteeing large-scale co-operation. The presence of a single rogue advanced agent would be enough to jeopardize the strategy as it requires full compliance.

4.2. Why not traditional moral progress over moral enhancement?

At this point, one possible objection is that moral enhancement could potentially solve these problems, but it is not necessary in order to do so, and therefore other less radical

and more traditional forms of moral improvement should be preferred. I will examine one promising argument for this position defended by Buchanan & Powell (2018), which was briefly summarised in Chapter 1 (section 3.4). As mentioned there, they agree that humanity is faced with challenges that our present morality is unable to address but do not hold that biotechnological interventions are necessary in order to effectively overcome our moral unfitness.[71]

Buchanan & Powell claim that moral enhancement advocates (whom they call evoliberals) and their opposition (evoconservatives) share the evolutionary assumption that our evolved moral traits are sufficiently innate and inflexible to make substantial moral progress via traditional means impossible, although the two groups draw opposing conclusions. Once we reject the assumption that human nature is sufficiently inflexible that justifies a pessimistic view of the prospects of moral progress, the positions of both moral enhancement advocates and evoconservatives are refuted. The two authors proceed then to argue against this assumption both in theory and in practice. Theoretically, the fact that our moral dispositions are evolved adaptations does not justify the claim that they are insufficiently flexible. An explanation can concern either the past history or the current state of its explanandum; explanations can be either diachronic or synchronic. The former provides factors immediately responsible for its explanandum, the latter the distant historical factors; the former offers proximate causes whereas the latter distal causes. Accordingly, Buchanan & Powell argue that evolutionary theories provide only a diachronic explanation, which relies on a distant history in order to draw conclusions about the evolved function of certain traits requiring a synchronic explanation to be fully understood. That is, evolutionary

[71] The central thesis of the book is not about moral enhancement itself but about moral progress in general. I will not be concerned with the central thesis. Moreover, I will only address the arguments that relate to moral enhancement and *only in so far as they do*. Some arguments that I argue to be unsuccessful in giving reasons for dismissing moral enhancement can still be successful in other regards.

accounts explain why a certain, otherwise unlikely, biological configuration becomes recurrent, by showing it is an evolutionary adaptation to recurrent past evolutionary pressures. It does not produce an explanation of all the current aspects of the trait. The fact that a trait has a function in order for it to be an adaptation to certain environmental conditions does not entail that the trait cannot be currently modified by culture. In their own words,

> "The question of moral malleability turns on the nature of morality's proximate (synchronic) causes, not on its distal (diachronic) causes. In other words, what matters for purposes of gauging the plausibility and durability of moral progress is the nature of the moral psychology we currently possess regardless of how or why morality originated.
> Put more technically, synchronic properties, which determine how moralities develop from a complex interaction of genetic, epigenetic and environmental causes, "screen off" diachronic properties in relation to the alterability of human moral psychology." (2018, pp. 351-352)

Empirically, history is full of examples of substantial moral progress overcoming our moral failings. The two authors cite the abolitionist movements and a decrease in inter-group wars as major examples of what they call the inclusivist anomaly, whereby we expand our sphere of concerns beyond our own group. According to them, the standard evolutionary account of human morality says our altruistic propensities are necessarily parochial, as humans were subjected to the pressures of inter-group competition, which selected only groups who were able to co-operate internally while competing with out-groups whenever necessary. An unrestricted tendency towards co-operation would have been disadvantageous and selected against (this has been discussed more extensively in Chapter 2, section 3.2). The fact that human societies have been able to develop higher levels of moral sensibility for out-groups, undergoing structural changes for their sake, shows that culture can produce moral progress despite going against our evolved inclinations. The authors also offer an evolutionary explanation for this inclusivist anomaly, claiming that under conditions of low environmental stress, such as the absence of inter-group conflict, violence, scarcity, or

parasites, it is actually advantageous to be able to co-operate with out-groups. Their argument is that our evolved morality is plastic and can be moulded by culture much more than both evoliberals and evoconservatives suggest. Therefore, it seems their main point of disagreement is over the flexibility of our evolved morality and not necessarily over whether or not it was shaped by natural selection.

Their argument claims that a belief in an insurmountable obstacle for traditional means of moral progress is flawed, because it is based on an inaccurate view of how evolution shaped human morality. Their conclusion is that

> "We agree with the evoliberals headline that there is an "Urgent need to enhance the moral character of humanity", but we do not think that BME [i.e., moral enhancement] is likely to be a very effective and plausible means by which to do so" (p. 373).

If we take this conclusion to mean that we should not pursue the development of moral enhancement, it requires a degree of certainty over which means of moral progress to pursue that we do not seem to have. The arguments for claiming that cultural progress alone can yield the unparalleled and much-needed level of moral progress in order to meet our radically new environment are not that certain. They present convincing arguments in favour of moral progress as an effective way of improving some moral failings, but their arguments do not justify the claim that traditional moral progress alone should be preferred to meet the unprecedented challenges posited by modern technology and society. I will list several reasons to be uncertain of their reasoning; I do not intend to offer a refutation of their arguments, but to cast doubt over how sure we can be of their conclusion that we should favour traditional moral progress alone. Their conclusion requires two claims, that technological moral enhancement is unlikely to produce the necessary moral progress and that traditional forms will be sufficient.

Regarding their first claim, one can make the argument for moral enhancement without the use of the assumption that human morality is extremely inflexible. An advocate

of the moral enhancement project need claim not that moral traits are inflexible, only that some of them might be sufficiently hard to change via cultural means as to justify the pursuit of technological modification. The abolition of slavery was a long process met with extreme resistance, even in the form of wars. Moral enhancement via technological intervention might present a less costly and bloody means to achieve moral progress. If abolitionists could have developed a technology that would help slavery advocates to have perceived slaves as human beings equal to themselves, this could have prevented the American Civil War.[72] Perhaps we can avoid the bloodshed that could be implied by enforcing an expansion of basic human rights in populations that currently oppose granting those rights to people of all genders, sexual orientations or races. Moreover, it seems that treating members of all groups with some basic concern for their freedom and well-being might indeed have been advantageous in those periods in the prehistoric environment that were relatively peaceful and disease-free. Historical examples might only reveal cases of moral progress that consisted of merely shifting which moral traits we exhibited rather than overcoming them. Under certain stable and peaceful environments, being able to co-ordinate trade and mating between groups would produce higher fitness and there is, in fact, plenty of evidence early humans engaged in such exchanges.[73] But, under this explanation, respect towards out-groups evolved only to the extent that it enabled such basic trading. The extremely high levels of large-scale co-operation necessary for overcoming the extreme risks listed in this thesis seem to go much further beyond a basic respect for out-groups and what would have been selected for in our ancestral environment. Therefore, not only would moral

[72] At the very least, it would have helped to advance its legitimate goals with less bloodshed. This bloodshed was by no means trivial even when comparing with other wars of the time, killing 3% of the American population and using particularly gruesome tactics. For instance, the strategy of total destruction during the Atlanta Campaign (Castel, 1992) or the futile massacre of blacks during the Battle of the Crater (Suderow, 1997).

[73] For the most recent findings that put the documented existence of exchange networks earlier than 300,000 years ago see Brooks et al., 2018.

enhancement remain a desirable option in place of traditional costly moral progress even when this moral progress is possible via traditional means, but also the extreme and unparalleled levels of moral progress required might be unattainable via traditional means, simply because a reason for them was wholly absent during our evolutionary history. Buchanan & Powell would have to offer a decisive argument to show that such levels of moral progress are not only feasible via traditional means, but also attainable with lower costs than via technological intervention. The fact that human morality is sensitive to cues present in our evolutionary past is no such argument. At best, it is an argument that certain types of moral progress seem to be achievable via traditional means only after much bloodshed, as observed in the case of the abolition of slavery.

Their second claim that traditional moral progress is enough is also not so securely established. It is plausible that much of our moral progress has come about via accidental and non-moral advances. For instance, the lower frequency of actual wars does not seem to be solely caused by moral progress. More rigorous work on nations' proclivity towards war argues that this has increased in modern times, and that the fact that wars have decreased is a consequence of nations now being more spread out (Braumoeller, 2013, 2018). That is, nations are avoiding war against each other not because they have become more peaceful, but because they have become more distantly placed. One may argue that what really matters is the actual frequency of wars. However, intuitively, peace seems to mean a state where people are intentionally not committing violence and not just accidentally. A prison might have lower violence than a certain neighbourhood, but it might still not be considered a more peaceful place exactly because the individual proclivity to violence is higher despite the fact that violence itself is not. Proclivity matters for morality. One of the most convincing examples the two authors cite might plausibly be an unintentional consequence of the technological capability to spread across the globe, not moral progress.

Although there is an overall trend of decreased violence throughout human history, from modernity until now conflicts have been more sporadic but also deadlier and more focused on non-combatants likely to be under-represented in statistics and whose killing is more atrocious. World War II has been the deadliest conflict in human history, with the highest death toll per year, both as a percentage of the global population and in absolute terms (Human Security Research Group, 2013).[74] Approximately half of those killed were non-combatants, which decreases the reliability of the data. Recent violence has taken a fat-tailed distribution, one in which there is a low probability of extreme events that are uncertain to estimate. Statistical analysis that takes this into account concludes we cannot assume our recent peaceful period after the world wars resulted from a real trend of decreased violence instead of a trend of more sporadic, but deadlier, violence (Cirillo & Taleb, 2016). Moreover, even the relatively peaceful period after World War II seems to have come accompanied by the new risk of nuclear annihilation and maybe have partially resulted from it due to nuclear deterrence.

The two authors suggest this recent decrease in wars was brought about by the rise of liberal democracies, which are less likely to engage in conflicts than more centralised governments. Suppose we grant this claim and reject the plausible possibilities conflicts have either become unlikelier but deadlier or unlikelier unintentionally. When it comes to technologies with the potential for global or existential catastrophe, it is not enough that on average societies have been shifting away from overly centralised power structures and closer to liberal democracies. It is enough, however, for one rogue nation to gain access to these technologies in order for a worryingly high risk of global catastrophe to remain. The claim by advocates of strong moral enhancement that moral enhancement alone is necessary

[74] There is an important moral question behind the decision to use either absolute or per capita deaths. Using the former paints a less optimistic picture about the decrease in war deaths.

and sufficient in order to solve our moral failings requires the assumption that everyone will take their moral pills. On the other hand, Buchanan & Powell's claim requires the assumption that every society with access to powerful technologies will be a liberal democracy.

The costs of spreading moral progress across the globe are not trivial. They cite the spread of liberal democracies as a primary example of this. However, one of the latest intentional attempts to spread democracy, the Iraq War, has cost the world over one trillion dollars (Amadeo, 2018) and hundreds of thousands of human lives (Hagopian, 2013). The total wealth possessed by all Irish nationals and their government is less than that value (Credit Suisse, 2017). Just in financial costs alone, implementing democracy in one rogue nation can cost more than the total economic value of one advanced nation. It is hard to estimate, but it is challenging to conceive how implementing a form of moral enhancement that reduces out-group aggression in one rogue nation would have cost more than one trillion dollars or incur more violations of basic rights than the deaths of hundreds of thousands of human beings (of whom a significant percentage were innocent civilians). One might claim that this is an extreme example of failure for traditional moral progress, but we cannot ignore such cases. As mentioned, it is enough to have one rogue nation with access to powerful technologies in order to jeopardize a global effort to reduce extreme risks to humanity.

Buchanan & Powell's claim that a trait's synchronic aspects (the current state of a trait) screen off diachronic aspects (the past history of a trait) does not enable them to conclude that the current state contains more possibilities than the causal/evolutionary history would suggest. If the diachronic aspects are constrained (screened off) by the trait's synchronic states, the latter must contain fewer possibilities than the former. For instance, when patients are screened off from a certain medical evaluation, fewer patients remain at the end. A constraining process, screening off, reduces the number of possibilities. In the case of evolution, the diachronic explanation entails that the current traits must be an evolved

solution to some recurrent evolutionary challenge, but there is a wide range of possible solutions to any given proposed challenge. A full synchronic account would provide a complete description of a current trait, but if evolutionary theory is correct, then it cannot give an account that is impossible according to evolution. If current states really contained more possibilities than aetiological history would suggest, then we would have to find violations of evolutionary theory everywhere. Rather, we find traits which are the realisation of one possible solution, out of many, to recurrent evolutionary challenges. Of course, most of these solutions are sub-optimal and complex, this is not only within the possibilities of natural selection but to be expected.

It is clear that diachronic explanations are of little relevance compared to detailed synchronic explanations when deciding if and how it is technically feasible to carry out a focused intervention. A surgeon intending to operate on a human heart would not know where to cut if relying on an evolutionary explanation of why the human heart was selected to have its current configuration. A competent surgeon needs a detailed physiological and anatomical model of the human heart, a synchronic account. Of course, evolutionary accounts can sometimes help elucidate the possible reasons behind the human heart's current configuration, but they are merely auxiliary. However, when it comes to interventions that will have intergenerational or societal consequences, we often look into history. The distant past might be irrelevant when determining if and how it is technically feasible to change a specific feature of our immune system, but the reasons this feature evolved matter for assessing the long-term consequences of this change. For example, suppose it is the case that dust and seafood allergies are caused by the misfiring of a feature of our immune system that evolved to defend us against being poisoned by certain animals that went extinct.[75] Then we

[75] This is merely hypothetical; for a more refined hypothesis attempting to explain links between substances found in both dust mites and shellfish and associated with poisonous organisms see Komi, Sharma, & Dela Cruz (2017) and Wong, Huang & Lee (2016).

have reasons to expect deactivating such a feature would be beneficial. If, however, we were to discover that this feature evolved to protect us against some dormant pathogen, then we have reasons to expect deactivating this feature would be dangerous.[76] When intervening in a feature that can have consequences over long time periods, diachronic explanations become more relevant because they can reveal how that feature is likely to change across time. In that respect, moral progress is more like intervening in the immune system and less like heart surgery. Moral progress does not consist of focused interventions whose consequences are immediate and easily identifiable; its repercussions are intergenerational and complex. Moreover, moral progress directly engages with safeguarding and enabling the flourishing of humankind. It is no surprise that more profound and ultimate explanations should be more heavily involved.

Given their extensive use of historical examples, Buchanan & Powell are likely to agree that moral progress is not akin to focused interventions. One way to make sense of their claim that synchronic explanations are what matters when intervening with moral psychology is that human history is so recent that it can be considered to contain only proximate events. Perhaps, for them, whatever happened before human civilisations emerged are distal events which are the concern of diachronic explanations such as human evolution, and whatever happened after are proximate events which are the concern of synchronic explanations such as human history. It seems counter-intuitive to place human history with the set of synchronic explanations, thus putting interventions with historical consequences closer to heart surgery than to changing our immune system. More importantly, their strict separation between diachronic and synchronic explanations gives insufficient weight to the fact that moral traits have been subjected to evolutionary forces from the Pleistocene period

[76] Marichal et al. (2013) and Sherman, Holland & Sherman (2008) propose evolutionary advantages of allergies.

to contemporary history, and that inclusivist moralities are also adaptive strategies co-evolving with more ancient moral traits, and as such are subject to environmental cues only to the extent that they had been selected for in our most recent evolutionary history. They correctly identify inclusivist tendencies that would arise in the absence of human and non-human threats, but do not mention that such a response, also being an adaption, cannot be completely freely shaped by cultural moral progress. As discussed, we may have evolved an adaptation that enables us to have a basic respect for out-groups in safe environments, but it seems unlikely that we could have evolved the degree of large-scale co-operation necessary to overcome the extreme risks listed here.

A significant part of the traditional moral progress achieved by human civilisation has also been carried out by changing our genes and, consequently, biochemistry. There is substantial empirical evidence that human evolution has not only continued since the Pleistocene (Courtiol, Pettay, Jokela, Rotkirch, & Lummaa, 2012; Voight, Kudaravalli, Wen, & Pritchard, 2006), but accelerated (Hawks, Wang, Cochran, Harpending, & Moyzis, 2007; Helgason et al., 2015). Scientists expect that this was primarily a result of increased sexual selection (Richards, 2016), which more heavily selects based on social behaviour than other forms of natural selection. There is preliminary evidence that several of those selective forces have, in fact, influenced genes connected to social behaviour (Harpending & Cochran, 2015); and hence to moral behaviour as well. Therefore, a strict preference for traditional methods instead of technological moral enhancement does not result in leaving the biological basis of morality untouched, but in letting it be manipulated via cultural means alone. If biological changes form part of the mechanisms through which past moral progress has been realised, then the argument against intervening in our moral traits cannot use past successful moral advances as its basis. Altering the biological basis of our moral traits has been one of the means through which moral progress has happened. If sexual selection has been one

primary driver of the recent changes in our moral traits, rejecting technological means of directly intervening in our moral traits amounts to a strong partiality for sexual selection over technological manipulation as a means of changing moral traits. The argument for preferring sexual selection over careful and intentional technological manipulation allied with traditional methods is difficult to make. It seems unlikely that our sexual preferences alone would be a factor more conducive to human flourishing or reducing extreme risks than sexual preference paired with intentional technological manipulation. In fact, Bostrom (2004) has argued that in the absence of a globally co-ordinated policy to control human evolution, uncontrolled evolution will lead to the elimination of the kinds of beings we care about (Bostrom, 2004).

The claim that traditional means of moral progress alone are sufficient to overcome our current challenges seems not significantly more reasonable than the claim technological moral enhancement is the necessary and sufficient solution. A solution to our moral failings leading to extreme risks is certainly necessary. Until we know a sure solution, no proposal should be deem sufficient and none of the possibly feasible solutions should be discarded.

5. Conclusion

A strict preference for traditional methods of moral improvement for solving the moral failings leading to extreme risks should be rejected. Addressing these risks is of major moral relevance, but solutions may not come from traditional moral progress alone. Properly considering the initial proposals for moral enhancement reveals that most arguments against it fail but that deep moral enhancement – i.e. moral enhancement carried out to the extent necessary to address extreme risks – can have extreme risks of its own. The arguments that moral enhancement is technically unfeasible, poorly motivated, conceptually mistaken or a threat to freedom are not compelling. Instead, moral enhancement has the potential to cause

significant unexpected consequences due to the complexity and fragility of moral traits. It might be so radical that it undermines interests sensitive to psychological continuity, both at the individual and generational levels. However, its probable effectiveness in decreasing known sources of extreme risks provides strong reasons for pursuing its development. If we can avoid its pitfalls, then there seems to be no substantial reason for not trying to develop deep moral enhancement. I will explore how to avoid those pitfalls in the coming Chapter.

Chapter 5: Virtue theory for deep moral enhancement

1. Recapitulation of the risks of deep moral enhancement

In my second Chapter, I argued that the plausible targets of deep moral enhancement are unusually descriptively complex and relatively prone to unexpected consequences when attempting to modify them – for instance, consequences related to the self-reinforcing nature of motivation and to the paradoxical effects emerging in society as a result of modifying individual behaviour. In the subsequent Chapter, I have argued that several defences of moral status enhancement fail because they do not account for interests whose realisation depends on the existence of the same individuals, or type of individuals, who held those interests – for instance, at the individual level one must consider that there is a loss of psychological continuity after substantial moral enhancement and in the societal level one must consider the risk of creating beings that are not human continuants in the relevant sense. In my fourth Chapter, I made the case that deep moral enhancement can solve a wide range of extreme risks for humanity by listing some major known extreme risks and noting how they often arise from large-scale co-operation problems. Moreover, I argued that recent proposals to address these same problems with traditional moral progress will face many difficulties that moral enhancement can help alleviate, provided it steers clear from generating extreme risks itself. The present Chapter will outline some straightforward positive prescriptions to avoid the risks I have explored so far. I will then argue that some form of virtue theory is likely to fulfil those prescriptions and to address other concerns present in the literature. Finally, I will compare previous attempts at developing a virtue theory framework for moral enhancement with other possible suggestions regarding addressing the many severe risks I have explored in the preceding Chapters.

There are at least three general levels of uncertainty creating risks for the project of moral enhancement. Firstly, there is theoretical moral uncertainty over which ethical theory is correct and hence which should be the goal of moral enhancement, i.e. the ideal moral agent to be pursued. This is expressed in the investigation of moral status enhancement. For instance, the risk of creating supra-persons lacking important human values is a consequence of being unable to guarantee that a proposed moral enhancement is being guided by a complete account of our values.[77] Secondly, there is descriptive moral uncertainty over how exactly human moral psychology works, i.e. the departing point of moral enhancement. This is expressed by the complexity claim: the claim that human moral traits are relatively more complex than other human traits. Thirdly, there is uncertainty over what the long-term overall effects of moral enhancement will be, i.e. how human moral psychology will react to proposed modifications. This is expressed by the fragility claim: the claim that moral traits are prone to unexpected disturbances when we attempt to enhance them.

2. A framework for moral enhancement

2.1 Desiderata

In order to address these three levels of uncertainty and potential risks, a framework to steer the development of safe moral enhancement should have the following properties. The safety of such a framework would have to be robust to moral uncertainty, so that even if the adopted ethical theory were incomplete, it would not steer moral enhancement in completely the wrong direction. Arguably, an initial focus on accounting for the multiple traits of human morality in preference to producing a unifying ethical theory would be more likely to do justice to the complexity of moral traits as evidenced by the complexity claim.

[77] For a longer and proper investigation of the problem of theoretical moral uncertainty, see MacAskill (2014).

Furthermore, the framework would have to be heavily informed by neuroscientific data on moral traits to avoid ignoring the complexity of our moral psychology and to fulfil the need to guarantee proper technical feasibility. Additionally, a proper consideration of identity or of all human traits that could be related to identity would be more likely to preserve individual psychological continuity and to produce human continuants over generations. Moreover, aiming for the proper balance between competing moral dispositions would help prevent the large unexpected consequences of failing to do so; for instance, it would help to avoid creating runaway self-reinforcing chains of enhancement of single traits leading to overall moral decay, as evidenced by the fragility claim. Finally, a broader consideration of the practical consequences of enhancing moral traits would help prevent the emergence of catastrophic effects in society.

In this section, I will briefly introduce virtue theory. In the next section, I will argue that a virtue theory framework is more likely to possess all these desiderata than alternative frameworks. In the other two subsequent sections, I will introduce five different possible virtue frameworks and then compare how well they comply with these desiderata. The primary objective will be that by indicating what a compliant framework would look like, I will illustrate the usefulness of a virtue framework, and not necessarily argue for or build a specific framework. The secondary objective will be comparing available frameworks and suggesting future directions.

2.2 Virtue theory as a candidate

I will follow Crisp (1998) in adopting Julia Driver's distinction between virtue theory as descriptive and virtue ethics as normative.[78] I will define virtue theory as any form of

[78] In Driver's own words, "Virtue ethics is the project of basing ethics on virtue evaluation. I reject this approach. This is an essay in virtue theory, since what I am trying to do is give an account of what virtues *are*." (Driver, 1998, p. 112). Crisp's endorsement can be found on p.5 in his chapter on the same book.

descriptive theory that focuses primarily on the virtues. Therefore, the framework I will propose is based on a descriptive account of the virtues, although I will propose we should use such framework to increase safety. Virtues are defined as those traits[79] conducive to the good; this is in opposition to a focus on actual/expected consequences of actions or the fulfilling of generalizable maxims.[80] Virtue ethics would be a normative theory about morality advocating the virtues; I will not be concerned with that in this Chapter. I intend to make the applied claim that when discussing moral enhancement, a virtue theory will be more likely to avoid the risks I have so far explored. In that sense, I will make the normative claim that one should use a theory about virtues to avoid crucial risks for deep moral enhancement – which might be simplified to the claim one should use a theory of virtues when attempting deep moral enhancement.

A virtue theory is primarily concerned with the moral agent and his traits in so far as they lead to the good. Moral enhancement is primarily concerned with interventions expected to lead to better moral behaviour or motives and one obvious way of achieving that is by intervening in one's traits leading to the good. Hence, it should be no surprise that moral enhancement and virtue theory are deeply connected. Regardless of how the good or the right is defined, moral enhancement is often concerned with instilling traits that are more conducive to the good; that is, it seems it is concerned with instilling virtue. At the very least, both are concerned with the agent and his traits. In fact, it would seem counter-intuitive to think one can improve an agent's moral behaviour or motives without a theory about which

[79] As in Chapter 2, I will define traits as general and stable patterns of behaviour, thought or emotion. Traits are meant to be more robust and general than mere dispositions. One may have the disposition to be frightened by a specific snake appearing inside one's house, but one might have the trait to be frightened by the sudden appearance of visual patterns resembling consistent evolutionary threats to survival.

[80] The good, in turn, might be defined as good consequences or fulfilling norms. For instance, for Driver (2003) virtues are character traits that systematically lead to more good consequences than otherwise. Although I will use external good consequences being systematically produced by a trait as a criterion, I will also consider virtues might be connected to internal goods – or at least necessary to avoid undermining certain external goods – when mentioning virtues relation to personal identity. I do not wish to settle which virtue theory is to be used in ethics.

of his traits lead to the good. Therefore, virtue theory, as a theory ascribing a more central role to the virtues, is a *prima facie* good candidate for a framework for moral enhancement.

Moreover, here I will make the more specific claim that virtue theory is uniquely equipped to deal with the risks I have so far explored. A strong version of this claim is that virtue enhancement is one type of moral enhancement unlikely to face the risks I have identified as particularly likely for deep moral enhancement. Virtue enhancement could therefore assist in producing the needed moral progress where traditional means are limited by human moral psychology. It would follow that virtue theory is not only a good candidate but possibly a mandatory framework if we assume we have an obligation to avoid these risks. The virtue theories I will explore are not meant to provide the basis for a satisfactory virtue ethics, but merely to provide the basis for a safe framework for moral enhancement. I will argue certain common features of a virtue theory are also desiderata for a framework, but some of these same features might prove undesirable in a virtue ethics.

Finally, given that most of the risks I discussed relate to significant changes in moral traits, one can still avoid the need for a safety framework by focusing on superficial changes. As I argued in the second Chapter, a weakness of this solution is that it does not then target the problems moral enhancement is proposed to solve, or confer moral enhancement with any substantial advantage over traditional moral education. Therefore, I focused on deep moral enhancement, which consists of significant changes brought about by technological interventions targeted at fundamental human traits expected to lead to morally better behaviour or motives. It is with this goal in mind that I argue that deep moral enhancement guided by a virtue framework is the best way to bring about these significant changes. Given that I defined moral traits as the fundamental human traits that are plausible targets of deep moral enhancement, my argument in this Chapter can be rendered as saying that if we want

to decrease the types of risks I have explored, then moral traits should be equated with the virtues as I have defined them and we should treat them within a virtue theory framework.[81]

3. General desiderata compliance

3.1 Practical robustness to moral uncertainty

A virtue theory makes no necessary initial commitment to what constitutes the good. Good actions might be those of the virtuous agent who possesses the traits conducive to human flourishing, but it is perfectly possible to have a virtue theory focusing on which traits, in the long run, produce actions with the maximum amount of good actual/expected consequences or actions with the greatest intentional conformity to a universal moral law. Therefore, using virtue theory as a guide for moral enhancement will not exacerbate the risks involved with using a specific theory of the good as a guide in the presence of theoretical moral uncertainty. Virtue theory, when applied, can withstand moral uncertainty without leading to catastrophic failures; I call this property *practical robustness to moral uncertainty*, i.e. a framework that is unlikely to generate catastrophic prescriptions, even if the guiding moral theory is wrong or incomplete. With this approach, we can start with an adequate empirical model of human morality based on well-researched traits that are intuitively morally desirable. Contrary to more theoretical approaches, we do not need to commit to an overarching moral theory from the start. Nonetheless, a theory of the good will still be necessary to solve potential counter-intuitive cases. Despite the need for such a theory being

[81] One may define virtue in a way to include more superficial or specific dispositions, thus rendering virtue theory useless as a safety framework for deep moral enhancement. However, I suspect that most of what is meant by virtuous dispositions are types of traits, or at the very least, are only virtuous in so far as they are the application of a trait. Hursthouse (2001, p. 12), for instance, draws attention to the fact that virtues are not just tendencies to act in a certain way, but are entrenched and reliable, affecting a wide range of emotions and preferences and are not easy to acquire or to lose. For instance, bringing back the example from Chapter 2, a judge's racial bias in racially sensitive cases would be a specific disposition whereas we might regard the whole set of dispositions leading to racial bias across a variety of racially sensitive contexts as a racist trait – by racist here I mean assumed natural proclivities towards in-group racial biases, of which impermissible or illegal instances are clear identifiable violations of racially-neutral shared norms or laws.

unavoidable, this approach will minimise the risk of using as a guide an incomplete moral theory that misses some important aspect of what we value. The potential risks of using an incomplete axiomatic theory as a framework are bigger than the potential risks of using an incomplete virtue theory as a framework. An abstract theory has fewer practical constraints than an empirically adequate theory and thus can be more drastically wrong; thus, if we wish to minimise severe risks from moral enhancement, virtue theory should be preferred.[82] For instance, suppose Jim is about to take a pill that will increase his tendency towards co-operation. The pill will make Jim more likely to pursue outcomes where the sum of everyone's benefits is maximised. A simple utilitarian framework would evaluate that Jim ought to take as much of this pill as required to increase his co-operative trait. On the other hand, a virtue framework would prescribe a more conservative approach; it will be more likely to consider how this increase in co-operation would impact other traits that may be morally relevant, such as parochialism or altruistic punishment. I return to this example in more detail in section 4.3. Of course, the utilitarian framework can be made empirically richer and account for the complexities of the co-operative trait – but this will come as a posterior addition for the utilitarian framework while it will be an expected feature of a virtue framework, because virtues are more empirically grounded than abstract utilitarian principles. It seems even a utilitarian would have good practical reasons to adopt a virtue instead of a utilitarian framework.

In the third Chapter, I explored the risks of attempting to produce future generations of supra-persons who are merely extremely efficient utility-maximisers but who lack any other human trait relevant to the role of being our continuants. As I will argue in section 3.4.3, I believe a virtue framework will be unlikely to lead to these catastrophic scenarios in

[82] Empirical theories, when poorly reasoned, can be incoherent and impractical. These should also be avoided, which I aim to do with my *empirical adequacy* desideratum.

part due to its better treatment of moral uncertainty and the consequent lower likelihood of ignoring morally relevant traits.

One common objection to virtue ethics is that it does not always seem to be able to identify the correct action in a given situation. While utilitarianism might produce counter-intuitive answers, it will in principle always sort better from worse actions if provided with perfect information about consequences. Deontological systems might produce apparently unresolvable dilemmas but are still able to generate action guidance in most scenarios. Focusing on the agent and his traits, however, does not immediately reveal which actions should be pursued. One solution given by Hursthouse (2001) is to extract action guidance from conceiving what a virtuous agent would do in that given situation. However, since there is no straightforward way to balance conflicting requirements of different virtues against each other, there is no guarantee it will be able to tell what a virtuous agent would do even with perfect information. In such an instance, we might say we do not know what a virtuous agent would do. It would seem this attack on virtue ethics reveals that what I have called desiderata fulfilment might actually be a fault in comparison to alternative frameworks. However, what might be a problem in moral theory can be an advantage when applying virtue theory as a safety framework. We can see this by analysing how an error would spread across the framework. If we still have an incomplete moral theory, that utilitarianism or other fully codifiable systems will always be able to prescribe which sort of actions an agent should perform (even when highly counter-intuitive) and hence which kind of dispositions a morally enhanced person ought to have (even when, as discussed in Chapter 3, this person looks more like a superefficient alien than a human being) will only aggravate the problem. But if instead our incomplete knowledge results in having no clearly discernible prescription,

it will prescribe we have no easy answer and ought to consider the question further.[83] This is a prescription that is vastly safer than prescribing enhancement towards the wrong target when taking self-reinforcement and irreversibility into account. Proponents of virtue ethics have provided responses to this theoretical problem, but it seems they would still not entail the same type of risk present with straightforward codifiable moral theories.

3.2 Empirical adequacy

Even for a strong defender of consequentialism or deontology as the right moral theory, it should come as no surprise that human beings are not the best utility-maximisers or searchers and executors of optimal universalisable maxims. Virtues might not be intrinsically related to the good, but human morality seems to be better described by an account of virtues than by an account of utility maximisation or moral rules in isolation. I call this feature of conforming to empirical data empirical adequacy. This is not to say a utilitarian or a deontologist could not develop an empirically adequate theory of moral psychology, but only that virtue theory is closer to doing so as these theories presently stand.

Nonetheless, one common critique of virtue ethics has been that it lacks empirical adequacy, a critique that was heavily influenced by psychologists' failure to find reliable patterns of behaviour suggesting the existence of core aspects of character traits such as honesty. For instance, they found that the likelihood of being honest in a certain situation seemed to be determined more by the situation than with having been honest in the past, suggesting there was no such thing as the trait of honesty (Isen & Levin, 1972). Several responses have been made to this critique, such as claiming deficiencies in the experiments used to evaluate character, the fact that competing virtues might lead to behaviour that does not obviously express any individual virtue, claiming virtues are rare but do exist, etc. (for

[83] For a more general defence of the idea that we are often better off arriving at a dilemma instead of an outright prescription, see the first chapter of Nussbaum (2001).

an overview see Miller, 2017; for one of the most influential critiques see Harman, 1999). I will, however, focus on two relevant responses for building a safe framework for deep moral enhancement.

3.2.1 General similarity between moral psychology and virtue theory

Firstly, the lack of empirical support for certain famously proposed virtues does nothing to undermine the fact that the virtue approach would still be the closest to how the human brain processes morality. Virtues help explain the integrative nature of moral cognition better than either a deontological or a utilitarian framework. As argued by Casebeer & Churchland (2003), if human morality were primarily concerned with utility maximisation or following maxims, we would expect that most of our moral faculties are correlated and centralised in one highly specialised brain area dedicated to either of those ends. However, neuroscientific data contradicts the notion that human moral traits can be reduced to any single task or cognitive module. Instead, moral cognition is highly distributed among brain regions responsible for a variety of tasks such as processing emotions, perception, memory formation and abstract reasoning. I will briefly introduce different neuroscientific approaches to moral cognition, each with different definitions of morality, but all closer to virtue theory than to alternative moral theories.

Casebeer & Churchland's approach sees moral cognition as being centred around norms, defining it as cognitive acts and judgments associated with norms. Notwithstanding this relatively deontological framing, after compiling an extensive list of empirical studies, they conclude these norm-related cognitions depended more on the identification of morally relevant environmental cues and the production of the appropriate emotions on the limbic system than on the direct processing and storing of moral rules. They cite lesion studies where damaging the brain area responsible for either identifying morally salient cues or creating the corresponding emotion on the limbic system severely compromises behavioural

conformity to moral rules. It seems, therefore, that the brain can follow moral rules only in so far as it has the right context-specific responses. Exclusively formulating pure, consistent maxims as in deontology or rational action forecasting as in utilitarianism do not seem compatible with the emotionally laden processing essential to moral cognition. Instead, the process Casebeer & Churchland describe of causing appropriate emotions as a response to environmental cues generates behavioural conformity to moral rules over many interactions. For example, the role of the moral brain is to inform the rest of the brain that someone's angry face after you stole his cookies signals enough negative value to counter the positive value signalled by the sweet taste in your month, instead of concatenating an *a priori* rule against stealing (Moll, Oliveira-Souza & Zahn, 2008). Having the appropriate responses to specific contexts is an idea that features more prominently in virtue theory than in the other two main alternatives.[84] Casebeer & Churchland go on to make the more controversial and bold conclusion that this makes deontology as a moral theory unrealistic. For the purposes of this Chapter, the more modest conclusion that this makes a virtue framework more empirically adequate will be sufficient. The only normative claim I wish to make from this is that we should prefer such a framework when attempting deep moral enhancement.

Molly Crockett's (2013) approach is based on the integration of various forms of moral thinking. It identifies the utilitarian perspective with an abstract search function over a forward-looking decision tree representing contingencies between actions and outcomes, and the values of those outcomes. The deontological perspective is identified with a search function over a backward-looking list of context, actions and values, i.e. state-action reinforcement histories. A third perspective is the simple Pavlovian harm-avoiding bias,

[84] One less well-known alternative that is even more context-dependent is particularism (Dancy, 2004), which relies on an infinite list of conditional rules given all relevant contexts. But this theory seems unlikely to produce a corresponding neuroscientific approach given it depends on the specification of possible state-of-affairs and not just mental states.

based on aversive predictions, which prunes the model-based search function whenever harm is produced. Finally, this overarching model of moral cognition works by adding up these three perspectives by "treating the three systems as separate experts, each of which 'votes' for its preferred action" (p. 364).

Moll, Zahn, Oliveira-Souza & Krueger (2005) see moral cognition as social cognition of a co-operative nature; it relies on an account of the formation of co-operative behaviour as a product of social cognition. On their model, motivation towards moral behaviour is the result of context-specific goals derived from the general moral value one wants to realise. Therefore, specific goals such as cooking dinner for a friend are intertwined in our brain with more abstract moral values such as cultivating friendships. They may depend, for instance, on reading the friend's reactions to the food and remembering his gastronomic preferences as well as on having a proper abstract concept of friendship. Moral motivations, according to this view, depend on the integration of brain regions responsible for abstract thoughts, memory and primitive emotions – more precisely, the authors suggest the involvement of the pre-frontal cortex and subcortical mesolimbic regions for context-specific goals, and integration across the temporal lobe and fronto-mesolimbic regions for abstract goals. This model is supported by anatomical evidence showing relevant connections between these areas and imaging studies showing they are selectively activated when these goals are elicited; lesions to these areas are predicted to selectively impair one of these goals.[85]

Just as in the case of Casebeer & Churchland's approach, Crockett's and Moll et al.'s approaches point to the idea that the most likely candidates for the neural correlates of moral cognition are scattered but deeply connected in the human brain. Emotions, social cognition, explicit and implicit memory are all involved when processing morally salient scenarios.

[85] Some of these correlations have been confirmed (Moll et al., 2018).

This interconnected nature of human morality is more characteristic of virtue theory than of any of the other moral theories because only virtue theory can account for the fact we use different competing moral thought processes, which are selectively activated, and then integrated, in different contexts.

3.2.2 Possible existing traits

The second part of my response to the empirically inspired critique of the virtues is that the failure to find *certain* stable and general patterns of behaviour does not prove there are no such patterns.[86] In fact, as the study of personality has advanced, psychologists have found stable and reliable traits such as the big five personality traits (O'Connor, 2002; van der Linden, Nijenhuis & Bakker, 2010). After analysing large datasets containing people's self-described specific behavioural dispositions (e.g. keeping an orderly office), five groups of highly intercorrelated specific self-descriptors arose – openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism – suggesting these were general behavioural dispositions. Furthermore, these measures were reliable across a person's life, suggesting the existence of general and stable dispositions, which were then shown to correlate with different sorts of life outcomes. Another relatively successful attempt to better define character traits was made using the CAPS framework. It redefines character traits as clusters of interrelated dispositions that are activated in subjectively similar but objectively different situations, i.e. they are patterns of dispositions that arise in different contexts that are subjectively experienced as similar (Mischel & Shoda, 1995).

Gilbert Harman, one of the leading critics of virtue ethics' lack of empirical adequacy, has countered that such discoveries do not reflect actual dispositional patterns but only patterns in the way we talk about dispositions (Harman, 2009). He judges personality

[86] Only not finding *any*, after substantial research, would offer support to the thesis they do not exist or that we are not close to finding them.

psychology to be merely the study of folk personality theory. But the reliability, validity and explanatory power of these traits is a strong reason to believe they are real.[87] Moreover, they do correlate with life outcomes. Nonetheless, the criticism that the big five framework excessively rely on linguistic descriptors is valid and will count against it when I compare it with other available frameworks whose primary focus is behaviour.

A virtue framework would not only increase safety, but orient the project of moral enhancement towards a richer moral psychology, which can arguably provide the basis for intervening in aspects of morality that many critics accuse the project of ignoring. Some might even argue that the virtue framework has too many burdensome requirements for moral enhancement. The enhancement of complex features such as moral traits is an audacious task, likely to be technically unfeasible for the foreseeable future. Claiming those traits must fulfil the criteria for being virtues will increase the difficulties of already complicated project. The neurological machinery responsible for human morality is already a complex higher-order phenomenon and virtues might be seen as building upon that complexity. It could be argued that it would be hopeless to enhance virtue by merely using a simple biochemical substance. For instance, Jotterand (2011) has claimed that the moral psychology underlying the idea of moral enhancement is simply too crude to change either moral reasoning or moral content. However, that is misguided. Firstly, there is nothing inherent in the idea of moral enhancement that requires it to be done by just one simple neurochemical manipulation. It might take, for instance, a series of groups of manipulations done in conjunction with other traditional means. Moral content and moral reasoning also arise inside the human brain and as already mentioned, some have already proposed

[87] That linguistic patterns reflect real patterns is the central hypothesis of the big five framework, called Lexical Hypothesis (de Raad & Mlacic, 2015). To the extent that this framework has enjoyed empirical success, the hypothesis has been confirmed.

reasonable candidate areas responsible for different styles of moral reasoning and areas responsible for storing our concept of moral value,[88] as well as a crude understanding of the interconnections between these areas and sensory and planning areas of the brain. It is not unreasonable to assume that an intervention in those areas could change even one of the most abstract aspects of morality such as our conception of the good itself. What is more, moral education and the acquisition of virtue rely on interacting with the outside environment and learning from it; it would be odd to require that virtue enhancement does not. For instance, acquiring the proper measure of courage might take a combination of emotional, affective and mood interventions, the exposure to the proper environment fostering exploration, and cognitive and motivational interventions. One may lack courage due to a combination of being in a sad emotional state, having an insecure affective style, being in an anxious mood, never being exposed to a moderately risky but rich environment in which courage pays off, and having insufficient cognitive abilities to plan and execute a courageous course of action. It might take the administration of SSRIs, neurohormones, anxiolytics, exposure to medium-risk/resource-rich environment and dopamine reuptake inhibitors, in the right order and in the right amount.

3.3 Correct balance

Since its beginnings in Aristotle to its modern revival, one constant idea behind most virtue theories is that a disposition toward certain behaviours is a virtue only in so far as it is a disposition towards certain behaviours to the right extent.[89] For instance, the disposition towards facing one's fears is only the virtue of courage when it does not come as an excess

[88] The anterior temporal lobe and fronto-mesolimbic region have been identified as reasonable candidates for storing moral values (Moll, de Oliveira-Souza & Zahn, 2008).

[89] Following my earlier definitions, I classify Aristotelian virtue ethics as a virtue theory that further claims to be the correct moral theory.

leading one to take unnecessary risks (or in lack, becoming cowardice). Moreover, being virtuous is also a result of having all relevant dispositions in the correct extent balancing each other.[90] The proper amount of courage may vary depending on one's level of ambition; perhaps slightly excessive courage may not be as detrimental when paired with slightly insufficient ambition as it would be when paired with slightly excessive ambition. How far a person is from being fully virtuous is not just a result of how far each of his relevant dispositions is from being at the right level but also which dispositions are unbalanced and in which direction. For instance, in Chapter 2 I have noted how increasing propensity towards co-operation between individuals might lead to a decrease in the propensity towards co-operation between groups. One of the reasons such modification fails to improve a person's overall virtuousness could be due to failing to account for how increasing one desirable trait (e.g. individual co-operation) leads to an imbalance by also increasing undesirable ones (e.g. in-group favouritism).

Two central risk factors for deep moral enhancement discussed in Chapter 2 are already acknowledged in most conceptions of virtue. Firstly, that overly increasing a good trait – or decreasing a bad one – may not lead to moral improvement of the agent; and secondly, that the best settings for each trait cannot be easily reduced to a single value because their correct values vary with each other, hence acknowledging that human moral traits have a high degree of complexity. More specifically, I have also pointed to the risk that small increases in one moral trait, when done in isolation, can have a cascading effect of producing further compounding small increases, which could rapidly result in drastic and unwanted changes. But if we instead focus on improving several traits in conjunction while

[90] From section 6, Book II of Aristotle's Nicomachean Ethics: "Virtue, then, is a state of character concerned with choice, lying in a mean, i.e. the mean relative to us, this being determined by a rational principle, and by that principle by which the man of practical wisdom would determine it." (Aristotle, Brown & Ross, 2009, p.31)

recognising the fact that their correct settings have a complex correlation, we would be significantly lowering the risk of producing drastic self-reinforcing changes. This is because an increase in a currently desirable moral trait would constantly be evaluated against a wider background of other traits and contexts in order to be considered a true moral enhancement. Under a virtue theory framework, we would be more likely to be sensitive to possible derailments due to either overly increasing a currently desirable trait to undesirable levels or successive chains of improvements to a single trait in disregard of other relevant traits. In fact, a good application of such a framework would prevent self-reinforcing runaway modifications because it would involve a constant awareness of how other traits are affected.

3.4 Preservation of identity

3.4.1 The problem of identity loss

In Chapter 3, I have also argued that any kind of drastic enhancements, even when impersonally desirable, may undermine certain individual interests due to being detrimental to psychological continuity. A person's psychological continuity is preserved during a period – i.e. the person survives during the period – whenever all neighbouring pairs of time-slices of that person during that period preserve a certain minimal amount of psychological connectedness between themselves, thus meaning all time-slices of that period preserve psychological interconnectedness. The level of psychological connectedness a person's time-slices hold with each other is a function of how many psychological relations exist between them, and how strong those psychological relations are. Preserving the same memories is one of the most obvious and important of such relations, but preserving certain core intentions or values might be as well. As discussed in Chapter 3, a person may even increase her level of psychological connectedness by enhancing herself. For instance, suppose someone had a higher genetic disposition towards aggressiveness and was brought

up in a high-stress environment promoting even higher aggressiveness. Instead of letting her aggressiveness remain at a level set by her environment and the past evolutionary pressures that selected her high-aggression genes, she may deliberate on this disposition, decide a lower level of it is more consistent with her interests and undertake moral enhancement by decreasing her aggressiveness. This will increase the psychological connectedness between her two time-slices. If she remained unenhanced, her aggression levels at the later time-slice would bear less relation to her interests, intentions and values at a former time-slice than if she set those levels according to her will. Notwithstanding this possibility, I highlighted how the enhancement of moral status via deep moral enhancement did not necessarily mean those relations would be strengthened. For instance, one may single out one desirable trait such as co-operativeness and increase it so radically as to produce someone who may be morally better but who has a personality sharply different from his initial unenhanced version.

3.4.2 Virtue and identity

I contend that one way of guaranteeing the preservation or even strengthening of those psychological relations, and hence of psychological continuity, is to focus on enhancing virtue because virtue and personal identity are often deeply connected. The relationship between virtue and personal identity is evidenced by the fact that many concepts of virtue are intimately related to personal identity.

At the outset, empirical evidence indicates that a strong connection between virtue and identity is intuitive. Data from a series of experiments evaluating how people attribute values, happiness, weakness of will, blame or praise to another agent seem to be best explained by the hypothesis that participants assume virtuous behaviour is authentic, emanating from someone's "true self", and vicious behaviour inauthentic. For instance, someone wasn't judged to be authentically happy if vicious behaviour accompanied that happiness. Participants seemed to connect virtuous behaviour with someone's identity

intuitively (Newman, De Freitas, & Knobe, 2015). A second group of researchers found that increasing someone's virtuousness is seen as less damaging to personal identity than decreasing it by a similar amount (Molouki & Bartels, 2017). Others have found that someone addicted to a drug that led to moral improvement is judged to undergone less identity change that someone addicted to a drug leading to moral decay (Earp, Skorburg, Everett, & Savulescu, 2018). A recent study went so far as to conclude that a sense of morality is the most central aspect of one's identity (Strohminger & Nichols, 2014).

Moreover, van Hooft's (2014) notes that identity and self-realisation lie at the centre of many virtue theories such as those of Korsgaard, Lovibond or Levinas. He contends that functioning as a unified agent is a necessary condition of being a moral agent and that "Virtue is an existential quest for self-realization that provides the internal link between reason and motivation so as to drive my practical engagement with the world" (p. 161). Radically enhancing one's capacity to conform to universal moral rules or to produce sentient pleasure may eventually lead to a complete destruction of what constitutes one's identity. But if virtue is intimately tied to self-realisation, even radically enhancing virtue seems unlikely to lead to abrupt disruption to psychological continuity.

Korsgaard (2009) argues moral reasons cannot be exclusively external as they arise from internal reasons for the agent that develop from a struggle for integrity. She seems to argue further that they are exclusively internal reasons; however, regardless of whether this claim is endorsed, her argumentation is relevant for solving the identity loss problem as it would draw attention back to individual interests instead of merely attempting to maximise impersonal interests. Van Hooft notes that to build her case, Korsgaard makes an exposition of the Platonic virtues and concludes that they rely on a sense of unity of the self and its inner order; that is, she concludes that being able to act in a single unified nature forms the basis of acting virtuously. She argues that Plato's analogy between a person and a State being

governed by a unifying constitution reveals that self-governance as a unit of agency is what forms the basis of virtue. For instance, she cites a passage where the virtue of courage is linked to resolution in pursuing a personal course of action despite external dangers, and another linking temperance to the unifying rule of reasoning over appetites. If this is the case, it seems unlikely that one can enhance virtue while harming personal identity. One cannot become more virtuous at the expense of sacrificing psychological continuity because the continuity of the same person across time as a unity is intrinsically linked to the exercise of virtue.

Furthermore, Mela (2011) has argued that MacIntyre's (1981) definition of personal identity seems to be best characterised as one's attempt to structure a *narrative self* that responds to the question of "what is a good life?", the same question that virtue ethics attempts to resolve. Exercising the virtues by attempting to build one's self as a virtuous agent seems to form a basis of his conception of personal identity. It would then be impossible to become more virtuous at the expense of losing psychological connectedness. Enhancing virtue would be unlikely to bring about the sorts of abrupt discontinuities described in Chapter 3 given that these discontinuities would lead to a loss of structure in the virtuous self's narrative and the exercise of virtue. Interestingly, one of the critiques of the moral enhancement project uses MacIntyre's abstract conception of virtue to argue that the moral psychology underlying moral enhancement could not possibly affect the higher-order faculties, such as moral reasoning and moral content, that are essential to virtue (Jotterand, 2011).

3.4.3 Moral uncertainty and identity

The practical robustness to moral uncertainty will also help prevent identity loss. Given that a virtue framework is likely to include a multiplicity of empirically grounded psychological faculties, forms of moral enhancement that are guided by it are less likely to

produce abrupt psychological or generational discontinuities. Because they will consider more psychological properties in general, they will be less likely to dismiss the psychological properties relevant for psychological continuity that could be dismissed by relying on an incomplete or wrong moral theory. Therefore, they will be less likely to dismiss what matters either in individual survival or the continuity of human generations. For instance, Jim might possess a strong sense of justice paired with his strong desire for co-operation. When evaluating whether he should take the co-operative pill, a utilitarian framework will be more likely to prescribe levels of co-operative enhancement that would harm his sense of justice by allowing him to prefer outcomes of extremely unequal distribution but with large total benefits; the same could happen if he enhances his sense of justice alone, harming co-operation. A virtue framework would take into account the impact of increasing co-operation on his egalitarian trait. At the generational level, a virtue framework would avoid the sole utility-maximisers possibility of section 4.2 in Chapter 3 because it would not generate a strong commitment to ever-increasing just one preferred human trait in society. It could consider, for instance, that artistic expression without immediate utility is likely to be an essential part of humanity. It should be noted that with long-term deep moral enhancement it would be unlikely we would preserve all relevant psychological connections that might impact personal identity. Any attempt to modify fundamental human traits in order to improve ourselves is likely to come at some cost to personal and human identity. Perhaps Jim ought to sacrifice some of his egalitarian preferences in order to become more co-operative. If Jim sees more value in overall aggregative good in the world than in the way it is distributed and willingly and knowingly enhances in the co-operative direction at the cost of decreasing his egalitarian disposition, the overall result would entail an increase in his psychological connectedness since his later self will be more connected to his past plans and values than otherwise. In MacIntyre's words, his narrative virtuous self will be able to

exercise his virtue more by being less influenced by innate dispositions that he may not value as highly as co-operation. However, there might be cases where individuals would have to sacrifice their identity in order to become morally better people. For instance, suppose we invent a new drug, which by enabling neuronal growth in areas responsible for moral sensitivity and moral reasoning would be able to treat individuals with extreme levels of psychopathic traits. There might be no way in which psychological connectedness would not abruptly decrease after gaining a fully functioning moral capacity. Most people would see the intervention as clearly positive if effective.[91] The proposed virtue framework would then have to consider how gaining fully functioning moral capacity would entail a significant enhancement of virtue. It seems that person would increase his virtuousness while losing personal identity. On the other hand, empirical data indicates that psychopaths are usually incapable of pursuing consistent long-term projects due to externalising behaviour, suggesting they lack a properly integrated self to begin with (Krueger, Markon, Patrick, Benning, & Kramer, 2007). Therefore, in so far as conferring them with fully functioning normal moral capacity would enable the pursuit of such projects, it would enhance their personal identities (Molouki & Bartels, 2017). Nonetheless, regardless of if we were to concede identity loss, what the framework aims to solve is the relative insensibility to losses of personal identity of a simple application of more abstract moral theories, not to place identity above other values.

[91] I will leave the possible harmful effects of introducing the practice of such drastic interventions aside, as well as considerations of deterrence and punishment. With regards to a single individual instance, conferring a known psychopath with moral capacities would be morally desirable. In the case of a convicted psychopath, one needs to consider deterrence and the criminal law. Moreover, the introduction of such practices by institutions should also bring considerations of balancing individual and collective interests into play. For a more extensive discussion of such cases see Pugh & Douglas (2016).

3.4.4 Preservation of identity – conclusion

One of the points that the possibilities explored in Chapter 3 seem to reveal is that it is not wise to attempt moral status enhancement while being exclusively guided by the question of what sort of value we want more of in the future, because we risk forgetting that what sort of evaluators we want in the future can also be relevant. I contend it would not only be relevant, but safer, to ask the latter question. Sufficient moral enhancement with an immediate eye on creating whatever dispositions realise what we think has value risks extinguishing the evaluators, and they might be the only ones able to tell if those values were mistaken. It will be safer, not to mention more pragmatic and intuitive, to centre the question around what sort of people we want to become using our current moral traits as a starting point. It seems clear that an application of virtue enhancement as a means to producing moral status enhancement in the cases explored in Chapter 3 would decrease the risk that individual interests dependent on psychological continuity would face significant unnecessary harm. Virtue enhancement would entail a richer and more careful modification, anchored by the traits we already value or traits we wish to have, instead of being primarily oriented towards the realisation of one abstract value that might well be incompatible with the existence of our continuants. In all three cases, it would help in guiding a more gradual and safe replacement of persons by supra-persons, either in individual or generational cases. For instance, the risk of losing too much psychological continuity in the *enhancing persons* case would be decreased. Given that their higher moral status would be achieved through a more balanced and richer form of enhancement across several traits, supra-persons would be more likely to be sufficiently psychologically related to persons and, therefore, more likely to fully realise the individual interests of persons. Current persons would have no reason to ascribe less relevance to the future benefit to supra-persons from their gain in moral status than to the future harm to persons from their loss in moral status.

3.5 Practical considerations

The modification of moral traits might avoid self-reinforcing chains, preserve identity, and account for individual balance between traits, but still fall short of producing moral enhancement if it ignores large-scale societal effects of introducing the possibility (or obligation) of such modification. When considering whether to develop or use a moral enhancement, one should consider the societal and long-term effects of changing individuals' moral traits. Once more, a virtue theory framework would help take such risks into account given that another of its central tenets is a proper consideration of context and practical rationality.

For instance, simply increasing altruistic dispositions could decrease overall co-operation in society by allowing highly individualistic and short-term oriented individuals to exploit others. Not only would virtue theory prescribe altruism or compassion to the right extent, but also in the right context. It would focus on increasing altruism while also preserving our sensibility to unfair situations and the disposition to take aggressive actions to punish cheaters.[92] As mentioned in Chapter 2, most plausible models for the evolution of human co-operation rely on the development of aggression either towards cheaters or towards out-groups. Given that most problems leading to the extreme global risks that moral enhancement is proposed to solve have to do with lack of large-scale co-operation, one disposition likely to be decreased would be aggression towards outsiders. Therefore, a future society would be even more dependent on aggression towards cheaters to preserve the stability of co-operation.

Practical considerations can reveal that even enhancements not targeting moral traits, such as those decreasing extreme suffering, could impact moral traits. Some inspired by a

[92] One important critique of moral enhancement was exactly that it could lead to unrestricted altruistic disposition even in cases where aggression to correct injustice was necessary (Harris, 2013a, 2013b).

short-term hedonist framework have proposed the elimination of extreme suffering with the use of biochemical manipulations. A standard objection is that this would have detrimental effects on motivation. In line with this objection, a utilitarian population-level approach seems to point to catastrophic results if this manipulation allows individuals who produce very low levels of intersubjective individual value to experience life without low levels of subjective well-being. The initial work I have done with Anders Sandberg of modelling the evolution of social preferences in society once they can be technologically manipulated reveals that one of the few conditions in which co-operation can become unstable is produced by allowing agents to have very low individual payoffs without any disincentives.[93] In this idealised model, co-operation is a stable strategy only in so far as there are strong incentives against agents having very low individual payoffs. Reducing extreme subjective suffering without introducing some other incentive against extreme low levels of individual value production could eventually lead to the disappearance of co-operative dispositions. Such utilitarian considerations must be balanced against our ethical intuitions aimed at reducing extreme suffering and inequality.[94] A virtue framework is relatively more open to balance these competing considerations. The initial development of such models, which focus on simplified moral traits, seems to be fruitful grounds for reflection on the long-term consequences of deep moral enhancement and offers at least one instance in which shifting the attention to traits and their population dynamics can reveal unexpected consequences along the way.

[93] See Sandberg & Fabiano (2017). This work will be explored in more detail in section 4.3.2 of this chapter

[94] Which seems to mean one should model institutions aimed at reducing extreme inequality, a research direction I aim to develop in the future with Sandberg.

4. Some possible frameworks

I will now introduce five candidates for a virtue framework and analyse how well they comply with the five desiderata. I will argue the one that has already been suggested for moral enhancement is deeply insufficient. In the next and final section, I will directly compare the five of them regarding desiderata compliance.

4.1 Big five, character strengths and their problems

The big five and character strengths are two prominent theories of character traits that would seem to offer an obvious basis for a virtue framework. Indeed, James Hughes discusses both in his paper "Moral Enhancement Requires Multiple Virtues", where he also advocates the use of virtue theory for moral enhancement due to its empirical adequacy and emphasis on the balance of dispositions (Hughes, 2015). His main argument is that moral enhancement should not be concerned with enhancing single traits but should instead use a richer model of moral character in which moral dispositions are interconnected. He proposes using one specific virtue-theoretic framework to guide the project of moral enhancement, called character strengths. I will argue that although he correctly concludes virtue theory has the right features of a guiding framework, his specific character strengths proposal actually fails on both empirical adequacy and balance, besides not even mentioning practical robustness to moral uncertainty and preservation of identity.

The big five is a model of human personality traits based on patterns in the use of common language self-descriptors. For instance, research demonstrated that certain self-descriptors were highly intercorrelated; someone who describes him/herself has having an organised desk is likely to also claim to be always prepared and to carry out tasks until their execution. Using a large dataset of self-assessed questionnaire responses about behavioural dispositions, dozens of specific self-descriptors were grouped into a handful of broad traits

containing self-descriptors that were empirically determined to be highly inter-correlated. Once these self-descriptors were grouped based in their correlations, five main groups emerged: conscientiousness, extraversion, agreeableness, openness and neuroticism. For instance, conscientiousness includes behavioural tendencies ranging from having an organised office, to driving with a seatbelt, to good performance in goal-oriented tasks. One's scores in each of these groups were found to be stable over the years and consistent with one's external evaluations by friends and family. They are also correlated with certain life outcomes. Conscientiousness seems to reflect a general tendency towards self-discipline and organisation and has a strong correlation with career success. Extensive research has been made on the big five traits, showing they are valid and reliable measures. However, there is no obvious correspondence between them and the virtues, and therefore they are not a good framework for moral enhancement. One of the most common critiques of the big five as providing a basis of moral character traits and thus virtue is precisely that they do not obviously correspond to virtuous traits. Firstly, it is not evident that having any of those traits is morally desirable or undesirable; e.g. high extraversion is not an obvious virtue or vice. Secondly, they come in degrees and are present in every individual; everyone has all the five traits to some degree and virtues are supposed to be less widespread and be a threshold trait.[95] In response to these objections, researchers built the character strengths model with the goal of finding traits that more directly correspond to the virtues; traits that if sufficiently present or sufficiently lacking would, respectively, correspond to virtues or vices. Instead of conducting a pure statistical analysis of the self-descriptors and letting traits emerge, the character strengths model tries to group self-descriptors into traits that correspond to commonly accepted virtues, while also being highly correlated. Those are wisdom, courage,

[95] I believe one can easily respond to this last objection by building the virtues as having each trait within the proper degree given certain life contexts.

humanity, justice, temperance and transcendence. However, both frameworks are based on the same method of statistical analysis that groups highly correlated self-descriptors together. This methodology comes with several problems for providing a virtue framework for moral enhancement.

Firstly, despite the fact that virtue theory is relatively less concerned with action than other moral theories, being able to anticipate the actual behavioural changes of any type of modification is essential for evaluating any potential form of moral enhancement. However, both frameworks are exclusively measured through self-assessed questionnaires. These are only linguistic representations of how people see themselves and might reflect the structure behind our linguistic assessments of personalities instead of a structure behind actual behavioural dispositions. Extensive research that successfully connects these traits to consistent real-life behavioural patterns is still lacking. There are, however, results showing a correlation between big five traits and long-term life outcomes such as professional success, income, health and well-being.[96] These general life outcomes perhaps represent more morally relevant dimensions than simple behavioural dispositions. However, they would only be relevant if we could target the enhancement of the traits leading to these outcomes; if, as I have argued, these prove to be too abstract for proper technological intervention, then we ought to look at other traits.

Secondly, the underlying goal of grouping several specific dispositions into broad traits is that these dispositions should be highly correlated inside the broad traits and scarcely correlated between traits. This means that there is, by design, very little appreciation of how different traits might affect each other and overall behaviour.[97] It follows that there cannot

[96] For instance, high conscientiousness predicts good work performance and academic achievement (Neal, Yeo, Koy, & Xiao, 2012). Agreeableness is associated with higher relationship quality in married couples (Holland & Roisman, 2008).

[97] There are some results on the interplay between several character strengths (Peterson & Park, 2012), but they are tentative and do not suggest how to produce a correct balance.

be an appreciation of the proper balance between several traits. In fact, despite his initial observation that we need to focus on the correct balance between moral dispositions, Hughes makes only a very brief mention of how, for instance, having too much of the character strength niceness may be overcome by having a higher cognitive capacity or how having higher self-control may overcome insufficient niceness.

It is symptomatic of this framework that Hughes spends a substantial portion of his paper praising the desirability of increasing niceness. This ignores how the altruistic aspect of niceness is likely to be linked to other undesirable emotions such as parochialism, as discussed in Chapter 2, or how the highly overlapping trait of agreeableness seems to mediate compliance with unjust orders, as explored in Milgram-like (1974) experiments (Begue et al., 2015).[98] Compliance with injustice and parochialism seem to be strongly connected with niceness and agreeableness, which have been argued to be important factors contributing to the emergence of totalitarian regimes that were perhaps the closest that modern human society has ever been to the global catastrophes moral enhancement is proposed to help us avoid. Aversion to causing harm to others, a component of niceness, seems to be increased by increasing serotonin levels, but this also has been shown to lead to a decrease in the willingness to punish cheaters (Crockett, Clark, Hauser, & Robbins, 2010) – a disposition that, as mentioned, might be indispensable to human co-operation. Whenever one of the main suggestions of a framework is the enhancement of a trait that has enabled past global catastrophes, it seems fair to say it failed as a proper framework for moral enhancement.[99]

---

[98] Although Milgram's results have found weaker replication (Blass, 1999; Dolinski et al., 2017), it remains a fact that the perceived desirability of agreeableness and niceness has been heavily called into question, particularly by psychologists specialising on the emergence of totalitarianism.

[99] Some moral philosophers have also argued against an empathy-based morality, an emotional subdimension of niceness or agreeableness. Jesse Prinz (2011) has argued such emotion is prone to several biases, parochialism being one of them. The disposition towards co-operating between individuals, against which I have argued, seems more emotionally mediated by empathy than the more morally relevant disposition of between-group co-operation.

Finally, because they arise from high-level linguistic descriptors, there is little to no empirical research on manipulating either the big five or the character strength traits with the use of potential moral enhancers. Not only that, but there is not even a well-established link between these traits and brain structures or neurochemistry. For instance, conscientiousness would be the most obviously desirable trait to enhance given its correlation to desirable life outcomes but, unfortunately, its neurological correlates seem remarkably elusive.[100] A sensible connection of virtues to big five traits to anything that might serve as the target of moral enhancement is lacking every step of the way. Moreover, although the character strengths traits have a correspondence with common virtues, their actual behavioural changes or neurochemical correlates are unknown. Most personality psychologists, including the pioneer of the big five (de Raad & Mlacic, 2016), believe that these traits might reflect linguistic patterns in our evaluation of our personality, and although they might have emerged under the selective pressure of actually modelling our personality with some accuracy, they might alternatively reveal only emergent personality traits that could be too far removed from traits that can be manipulated with the use of technological modification.

4.2 Social good CAPS

Nancy Snow (2010) presents an empirically grounded theory of the virtues that avoids a couple of the problems above. Later, I will argue another already developed framework fits Snow's theory but avoids most of these problems. Snow defines virtues as types of cognitive-affective processing systems (CAPS), systems which she, in turn, conceptualises as a group of beliefs, desires, feelings or values that produce consistent behaviour across objectively different but subjectively similar situations. CAPS are an well-

[100] There is a growing research into the neural basis of personality traits but no clear picture has emerged (Allen & DeYoung, 2016).

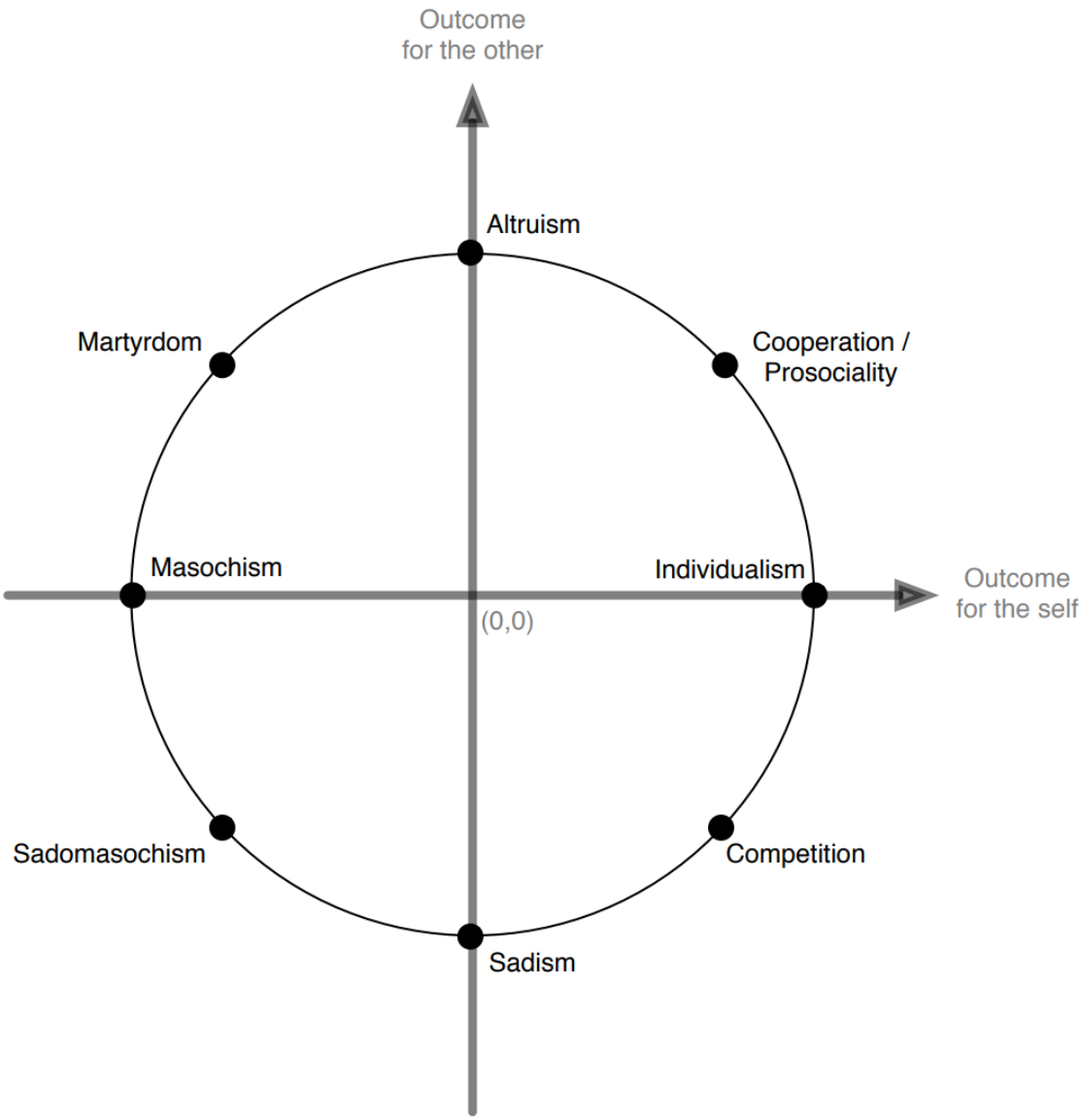
known model in psychology developed by situationists psychologists critical of the idea of certain stable character traits (Mischel & Shoda, 1995). They are psychological, social and physiological features that enable people to respond to their environment in a stable fashion; they include, for instance, mental states that are commonly activated together. These features, however, are only activated in relevant contexts and depend on environmental input; that is, they are contingent on context. For instance, someone's response upon seeing a spider might tend to include the same subjective feeling of fear and physical responses of avoiding contact with the spider. These responses will tend to come together upon contact with spiders. Meanwhile, another person might have a different set of responses such as curiosity instead of fear. This framework helps explain the absence of a simple correlation between past and present honesty, as it would claim a proposed "honesty CAPS" (if there is any) would manifest itself across situations perceived as similar in the relevant sense for that disposition. Virtues are then defined as CAPS that aim at fulfilling good social goals. Virtues deal with the successful pursuit of good social goals (e.g. comforting a friend in distress) whereas vices deal with the successful pursuit of bad social goals (e.g. taking advantage of a friend in distress). This framework is defined according to actual behavioural outcomes and is inherently social, but it similarly lacks neuroscientific research into its neurological basis and does not necessarily provide a way of evaluating the correct balance between different CAPS. Nonetheless, the conceptual development of virtue by Snow seems to fulfil all other desiderata, can eventually be empirically verified and would otherwise be a good framework for moral enhancement. Therefore, I will look at another framework that seems to fit most of Snow's conceptual development but whose neural correlates were better researched and which provides a precise way of balancing competing dispositions.

4.3. Social value orientations

4.3.1 Introduction

There is a third trait model, called social value orientations (SVO), which fits Snow's definition of desires leading to consistent behaviour across a variety of situations, is inherently social, well-established empirically, has a way of measuring the balance between dispositions and whose neurochemical manipulation has already been researched. The SVO framework was created from the analysis of social interactions and the observation that individuals presented consistent preferences for a specific distribution of benefits (or harms) between themselves and others (Van Lange, Joireman, Parks, & Van Dijk, 2013). In a wide variety of social contexts, if individuals are faced with deciding between a defined set of actions each resulting in a specific distribution of benefits between the involved parties, then empirical research demonstrates that these individuals will consistently act according to preferences for specific distributions of benefits; distributions that are contrary to pure rational choices (Murphy, Ackermann, & Handgraaf, 2011). Rather than solely aiming at maximising their own benefit, individuals will choose specific distributions of benefits between themselves and others. For instance, some individuals consistently prefer outcomes where the overall sum of everyone's gains is maximised (pro-social preferences); some consistently prefer outcomes with the minimal amount of inequality (egalitarian preferences); some, outcomes causing the greatest amount of harm to others (aggressive preferences). As first proposed by Griesinger & Livingston (1973), these specific and consistent preferences for various sorts of distributions can be mapped onto a Cartesian plane measuring how much importance is given to the benefits to oneself versus benefits to others (Ackermann, 2012):[101]

[101] Reprinted courtesy of the Copyright Holder under a Creative Commons License CC BY-SA 3.0. Retrieved from https://upload.wikimedia.org/wikipedia/commons/7/72/SVO_Ring.pdf



An individual's overall SVO is consistent across a variety of social interactions. Hence it produces consistent behaviour across objectively different situations. It relies on the individual's values and desires regarding the distribution of benefits. It also depends on correctly conceptualising social interactions, the possible course of actions for one and others and their consequences. Hence, it depends on social intelligence. Moreover, there is research indicating how the manipulation of various neurochemicals can alter these preferences. Increasing serotonin levels can lead to more pro-social preferences, higher levels of testosterone can lead to more individualistic preferences, and oxytocin can mediate pro-social preferences towards in-groups and aggressive or competitive preferences towards out-

groups. Finally, given that the all possible preferences in the SVO framework are the result of different weightings given to the same set of fundamental measures (the benefit to oneself versus others), the interaction between different levels of each of these preferences is well understood. The SVO framework seems to avoid most of difficulties facing the previous frameworks.

4.3.2 Applying the SVO framework

The fact the SVO framework is based on quantifiable variables with a well-defined mathematical relationship allows the construction of computer models that can shed light on how the possibility of altering social preferences can change a population's overall orientation.

In a previous publication with Anders Sandberg, we simulated a model of moral enhancement where agents play games with each other and can enhance their orientations based on maximising personal satisfaction (Sandberg & Fabiano, 2017). If agents who have consistently low scores (i.e. individual payoffs) on these games are removed from the population, agents tend to all converge to a pro-social orientation. However, the balance between pro-sociality and individual benefit-maximization is affected by different factors. A generally pro-social population is stable, but we concluded one must consider three potential sources for risks. First, the level of variation inside the population can be affected by various factors − e.g. availability of modifications and how/when agents are removed − and it is not clear how this would affect overall society. The level of selection against low scores can affect the stability of a pro-social population, and there is a great deal of uncertainty about the real-world strength of this selection. Finally, there are many other factors not included in our current model, such as inter-group conflict and vulnerability to cheaters, that could lead to paradoxical effects even if individual agents become more pro-social.

4.3.3 Limitations and future directions

The SVO framework avoids important problems that most others do not, but it is still fairly limited for several reasons. It focuses on just one relatively rational behaviour in cases of complete information, and it lacks a proper integrative model of how various emotions, sensory information and social norms tap into these types of choices. For instance, in the case of moral enhancement it is fundamental to look at the relations between individual preferences and group membership, such as the case of costly aggression targeted at out-groups as a means of manifesting altruistic preferences towards in-groups. Luckily, there are already some empirical results on that relationship and some have proposed adding a third dimension to the SVO ring to capture group-membership and how it affects social preferences (Aksoy, 2015). Furthermore, although the SVO is a good model of certain individual dispositions in social games, it lacks any normative information. In doing so it fails the final criterion for being a virtuous disposition: it is not targeted towards the good (or anything proposed to resemble the good). One still needs to add in the moral facts to enable a justification of any sort of intervention in to produce whatever change of social orientations.

Nevertheless, there seems to be a strong intuition that certain orientations are not conductive to the good. Exclusively aggressive, masochistic or sadistic orientations seem obviously morally undesirable. In building models that help understand which kind of settings could lead to these orientations to dominate society, we can help prevent the creation of catastrophes from seemingly innocent changes (e.g. changes in the strength of selection against low individual payoffs). There is also a strong intuition that preserving and enhancing the structure of co-operation in society is highly desirable. Given that goal, individualistic, altruistic and co-operative orientations are highly desirable if aggression can still be expressed in the right context. Our model reveals that a population of exclusively co-

operative agents does not seem stable; rather it tends to shift towards a mix between pro-social and individualistic agents as the norm.[102] Other models show the presence of altruistic punishment is also necessary for the stability of co-operation(Boyd, Gintis, Bowles, & Richerson, 2003). Although the SVO framework presents no straightforward mapping into virtuous dispositions, lack of inherent moral desirability can be solved by evaluating the specific outcomes of changing SVO, given the type of desirable societal change we would like to produce. Instead of ascribing intrinsic desirability to any particular angle, we ought to evaluate its societal effects first. Our and others' results suggest it would be a mistake to attribute inherent moral value to the co-operative orientation given it would not produce stability and would be likely to increase undesirable traits such as parochialism. Therefore, that there is no obvious case for certain orientations being desirable in the absence of practical considerations seems to be an instance where we are better off with no immediate prescription over risking having the wrong one.

Moreover, there needs to be a proper evaluation of the consequences of changing SVO to social institutions. Large-scale incentive structures of modern capitalist societies are based on only modest levels of pro-sociality paired with modest levels of individualism. If individuals put too high a value on co-operation, the incentives might no longer work – for instance, most markets depend on competition and lack of co-ordination between participants. Furthermore, since co-operative preferences are likely to be connected to more complex cognitive processes of empathy, sympathy or even fairness, one might find it harder to establish the social incentives enabling the construction of large-scale institutions.

[102] Individualistic orientations lead to co-operative behaviour either in the presence of external long-term selection forces or an internal individual orientation towards the future.

4.4. Moral development

Another possible virtue framework for guiding moral enhancement, at least initially, can be built by using the literature on moral development from childhood into adulthood to project at least some modest level of enhancement. If we can generally agree certain moral traits are better developed or only present in healthy adults, we can identify a natural and already existing progression towards higher moral development. Given that even adults possess moral traits developed to varying degrees, it would be possible to exert the same natural maturing neurochemical processes to those traits in order to produce modest moral enhancement. Perhaps that would mean creating the same advance as years of therapy or other intensive but traditional interventions with just a specific set of neurochemicals, or even advances unfeasible by traditional means but which would reach levels present in the human population. We are likely to be already tentatively doing so with the use of depression or ADHD medication.[103] For instance, the ability to notice one is over-harvesting resources from a common pool seems to depend on normal levels of serotonin in the brain, levels which SSRIs are supposed to restore in depressed individuals. Even interventions not intended to affect moral traits might have already substantially altered them. Oestradiol levels across the menstrual cycle in women seem to have a significant correlation to SVO, with higher levels leading to a more individualist orientation (Anderl, Hahn, Klotz, & Rutter, 2015). It is possible that certain forms of birth control could have a small but now widespread effect over the SVO of a significant percentage of the population leading to emergent societal effects of which we are unaware.[104] Intentional and more precise modifications done in conjunction might lead to a more significant change, especially if also widespread in society.

[103] In the US population aged over 12, over one in ten people takes antidepressants (Pratt, Brody, & Gu, 2011) and almost one in twenty takes ADHD medication (Express Scripts, 2014).

[104] Although, plausibly, the higher control over reproduction is likely to be the biggest effector, changes in social strategies would be expected following the administration of a fundamental reproductive hormone known to act as a neuromodulator such as oestradiol (Cutter, Norbury, & Murphy, 2003).

For example, Kohlberg's model of moral development has six stages of which most normal adults reach only the first four or five stages. The stages are characterised, respectively, by obedience and avoidance of punishment, instrumental self-interest, conformity to norms and approval-seeking, maintaining authority and social order, following social contracts and universal principles. Empirical studies seem to confirm these stages are sequential, distinct and increase with age; that is, people follow the stages without skipping any, they are in only one stage at a time and the likelihood of being in later stages increases as people age (Kohlberg & Hersh, 1977; Rest, Narvaez, Thoma, & Bebeau, 2000; Snarey, 1985; Walker, 1982). People at higher Kohlberg stages are less likely to be criminals or delinquents (Chandler & Moran, 1990). Using this framework would be extremely likely to preserve personal identity given the enhancement would be mirroring natural stages of moral development that would have perhaps already taken place in ideal conditions. Moreover, by studying the natural progression of hormones, neurotransmitters and brain development one could hopefully track down the biological correlates of each stage making the use of this framework technically feasible.

5. Exploratory applications

The proper focus on other-directed traits, an appreciation of their measured empirical effects in social interaction, as well as attention to the interplay between different types of traits can help us make tentative explorations concerning the application of different moral enhancements in different groups at the population level.

5.1 Liberty enhancement

Deep moral enhancement targeted at virtues can assist traditional institutional moral progress. As explored in section 4.2 of Chapter 4, Buchanan & Powell (2018) provide important political reflections for the project of moral enhancement. Moral progress implies

the progress of moral norms and social institutions. In Chapter 2, related reflections surfaced when I argued that we must consider the group-level effects of increasing co-operation between individuals. In Chapter 3, they surfaced as the consideration that generational moral enhancement needs to preserve the continuity of human generations in some sense. The full realisation of human interests in the future requires the existence of human continuants. In this Chapter, a focus on virtue has been proposed to help address these kinds of concerns. Technological interventions targeted at the individual must take into account that human beings are inescapably social. This critique can be stated in stronger and simpler terms by using the words of Aristotle: humans are a political animal,[105] moral enhancement needs to deal with this brute fact. Buchanan & Powell's critique, however, seems to focus exclusively on the group level via institutions. Politics takes place in the minds and interactions of beings who are not just political, but also animals with biological limitations that can be, at times, more easily overcome via technology targeted at moral traits than via political processes.

Inspired by the Enlightenment tradition, Buchanan & Powell (2017) argue that the removal of invalid moral norms is a form of moral progress.[106] They observe that removing a prohibition against sex between unmarried individuals can be seen as the removal of norms that were extremely costly to enforce but that do not produce sufficient benefits to human welfare to offset such costs. These came to be interpreted as improper moral norms and their removal has produced increases in general welfare. Similarly, they mention how removing a prohibition against homosexuality eventually led to an expansion of basic rights to

[105] Aristotle states in Politics, Book I (1235a), "From these things therefore it is clear that the city-state is a natural growth, and that man is by nature a political animal, and a man that is by nature and not merely by fortune citiless is either low in the scale of humanity or above it." (Aristotle & Rackham, 1932, p. 9)
To clarify, the meaning of political (πολιτικὸν in the original) is likely to be closer to "member of a organised society (polis)", or "civic" and "social" than to "involved with politics" in a contemporary understanding.

[106] One might choose a different terminology here. If one wishes to reserve the term moral rules to mean correct rules, "invalid moral norms" can be perhaps be seen as "immoral rules" or "wrong rules". I take Buchanan & Powell's (2017) definition of moral to be what can also be referred to as "descriptive morality" or "social norms".

homosexual individuals who can now freely exercise their identities. These are paradigmatic cases of moral progress. Technological interventions targeted at moral traits can lower the costs of enforcing such norms and run the risk of enabling the enforcement of improper norms. The idea that moral enhancement can be used to reinforce previous improper moral norms seems currently unlikely given the liberal and progressive tradition in which the project is inserted.[107] Nonetheless, lowering the costs of enforcing norms does entail an increased likelihood of enforcing improper norms in general. Technologies that target moral traits can be misused by institutions to enforce improper moral norms in general, either ancient, current or new improper moral norms. However, I will argue that this is not a cogent objection if we make use of a virtue framework. Firstly, it should be briefly noted that attempts to re-enforce improper moral norms would fail to pass the requirements of a virtue framework. Moreover, moral enhancement targeted at virtues that lead to a rejection of improper moral norms can actually assist with avoiding improper moralisation. I argue this can be easily seen in the following practical proposal: enhancing liberty as a virtue. Deep moral enhancement, if properly orchestrated alongside traditional moral education, can promote the virtue of liberty. Liberty can be framed as a moral trait. Liberty is a general and stable disposition that fosters the unconstrained pursuit of individual interests. There is at least one study that proposes liberty is one out of six foundations of human morality (Iyer, Koleva, Graham, Ditto, & Haidt, 2012); the study found this trait correlates with a preponderance of cognitive over emotional reasoning and lower interdependence. Studies on its neurological basis are still lacking, but it is plausible that orchestrated interventions that increase tolerance for outsiders and their ideas, increase cognitive reasoning and decrease emotional negative responses for out-groups can enhance liberty.

[107] For a discussion of the risk of using technology to enforce sexual orientation see Earp, Sandberg, & Savulescu (2014).

Deep moral enhancement can help increase tolerance and the recognition of basic human rights for members of all groups and identities, promoting free competition of ideas, avoiding improper moralisation of new or foreign ideas while guaranteeing that this is done in a safe environment that does not lead to infringing or violating basic human rights. Using technological interventions to produce some form of enforced inclusivism would certainly be questionable because it would run the risk of violating proper local norms and erasing group identities that are useful for producing co-operation at smaller scales. For instance, a staunch inclusivist institution might decide to use interventions targeted at moral traits to erase absolutely all racial preferences whatsoever, which would be likely to lead to some level of social unrest. Many people would, erroneously or not, want to retain some level of partiality towards their race.[108] On the other hand, promoting values that are conducive to a proper exercise of liberty and tolerance is not easy even in the context of liberal democracies. Buchanan & Powell (2018) recognise themselves that Jonathan Haidt and others have drawn attention to the fact that liberal democracies are a rare and recent institution in human history, an institution that requires fine-tuning of moral traits. Haidt (2017) states:

> "Here is the fine-tuned liberal democracy hypothesis: as tribal primates, human beings are unsuited for life in large, diverse secular democracies, unless you get certain settings finely adjusted to make possible the development of stable political life" (para. 6).

Being able to entertain foreign ideas freely is not trivial even inside a highly cohesive group such as a single democratic country. Inclusivist moralities seem to be rare enough that extreme levels of polarisation and dehumanisation arise inside a single unified country. What's more, these political divergences often arise in part due to different innate personalities (Graham, Haidt, & Nosek, 2009) and are being amplified by new technologies

[108] Historical developments in the last few years such as Brexit and the election of nationalist leaders indicate that even traditional political attempts to promote inclusivism can be met with a significant parochial response. Technological methods are likely to face stronger opposition.

such as social media that promote moral outrage (Crockett, 2017). In this context, dispensing with technological interventions altogether seems unwise. Conditions of abundance do not necessarily produce more inclusivism. Social media technologies seem to foster new possibilities of exclusivism even if their goal is simply to produce social interaction.[109]

The United States of America and the European Union might both be said to be successful cases of people from different ethnicities, religions and ideas able on the whole to peacefully live together inside an unprecedentedly large co-operative structure following the value of liberty. It seems unlikely these structures will crumble, but neither seems to be in a position to outright refuse technological interventions that could help alleviate out-group conflict – in particular, problems that arise due to new technologies themselves.

5.2 Developing countries, individualism and trust

Suppose we have the goal of increasing the speed of modernisation within developing countries. Research demonstrates that developed countries have higher rates of individualism (Shahrier, Kotani, & Kakinaka, 2016) but also higher trust between individuals (Inglehart et al., 2014), and that among developed countries those with higher levels of wealth have even higher rates of individualism (Inglehart et al., 2014). As mentioned before, it might be that our modern social institutions are better adapted to work in populations with only modest levels of pro-sociality and high individualism; their incentive structure seems to depend on self-interest. They might fail to work properly in populations in which special relationships matter more than the rule of law and where trust in the rule of law is lower. If the goal is to modernise societies in the direction of developed countries, it might make sense to alter the levels of trust and individualism in developing

[109] As pointed by Crockett (2017), data from a previous study (Hofmann, Wisneski, Brandt, & Skitka, 2014) shows that moral outrage at diverging opinions is higher online than offline. Dehumanization, one of the major mechanisms of exclusivism explored by Powell & Buchanan (2018), is likely to be more prevalent in online conversations than face to face.

countries. Interventions aimed at increasing both individualism and trust would enable faster rates of modernisation. This framework might also help decide whether those modern social institutions should be the main goal of developing countries given their underlying moral traits.

5.3 Stagnation and temporal discount

Another competing modern problem with increasing the rate of modernisation of poorer countries is how to deal with the secular stagnation of rich countries.[110] Growth rates in many developed countries have been decreasing, and some have had negligible growth over the last two decades (World Bank Dataset, 2018). Both in these relatively stagnated countries and those developed countries that still enjoy a fair amount of growth, it seems the main engine of growth (increased productivity) has stagnated – it may soon be demonstrated that those developed countries that have experienced growth may be doing so for accidental reasons, if that growth soon stops. There are various economic models attempting to explain the true origins of this stagnation, but most of them state that slow productivity growth, low interest rates, low investment rates and higher savings play a role (Summers, 2016). If we look at the traits leading to these sorts of behaviour one could propose attacking this problem at the neurochemical basis of those traits. Higher savings, low interest rates and low investment in productivity all seem to be the result of a relatively flat temporal discount curve in relation to money.[111] We still lack a comprehensive model of how different drugs can alter temporal discount, but we do have a fair idea that the dopaminergic system and sex

[110] Secular stagnation is a long-term condition of negligible economic growth in advanced economies. It happens when high per capita income leads to an increase in savings over long-term investments necessary to sustain economic growth. That is, capital stands still in savings instead of being invested, halting economic growth. The degree to which this phenomenon is relevant is debated, but many economists have come to take it seriously. For a compilation of experts' analyses of secular stagnation see Teulings & Baldwin (2014).

[111] This translates to behaviours that produce lower future value. Financial flat discount curves lead to higher savings, lower investment and lower productivity growth.

hormones can, respectively, affect the levels of investment in the present versus future in humans (Joutsa et al., 2015) and exploratory behaviour in animal models (Krsková & Talaroviová, 2005). There are longitudinal data indicating that levels of these sex hormones have been declining in developed countries (Perheentupa et al., 2013; Travison, Araujo, O'Donnell, Kupelian, & McKinlay, 2007). Whether this is partly the true origin or partly a mediator of the flatter temporal discount curve and lack of increased productivity, it would seem interventions to reverse that decline would help to correct the discount curve and increase exploratory behaviour into ways of increasing productivity. It might as well be that we find other solutions to this problem, but if stagnation proves to be a strong attractor for economic development, we will have to target this problem from all possible angles if we wish to continue the path of human development.

5.4 ISIS and similar cases

In cases where the use of lethal aggression is justified, or the status quo solution, against a group of outside aggressors, it seems deep moral enhancement would be either justifiable or preferred over the status quo. The main disposition which seems to drive an individual towards ISIS over other more pacific forms of social change is extreme levels of out-group aggression.[112] If one could target social preferences in order to decrease parochialism and aggression, the propensity to commit extreme acts of violence against out-groups would be substantially decreased and other forms of social change might be preferred. One could argue this solution would amount to an extreme form of coercion over a group by changing their deepest moral orientations without consent. However, their annihilation by

[112] A neurobiological approach to fostering virtue in such groups by targeting out-group aggression was proposed in Casebeer (2009).

lethal aggression seems far more coercive and comes at a much higher cost than the spread of a moral enhancement in a terrorist network.

6. Conclusion

6.1 Comparison of frameworks' desiderata compliance

The explorations in section 4 can be summarised by making explicit how each of the five proposed frameworks comply with the list of desiderata using the table below, which I will assess in more detail here.

Desiderata/ Framework	**Big five**	**Character strength**	**Social good CAPS**	**SVO**	**Moral development**
Virtue mapping	Unlikely (-1)	Built-in (+2)	Built-in (+2)	Possibly (0)	Built-in (+2)
Practical robustness to moral uncertainty (2x)	Built-in (+2)	Built-in (+2)	Possibly (0)	Built-in (+2)	Possibly (0)
Empirical adequacy: realistic traits	Likely (+1)	Likely (+1)	Likely (+1)	Built-in (+2)	Built-in (+2)
Empirical adequacy: technological feasibility	Unlikely (-1)	Unlikely (-1)	Possibly (0)	Likely (+1)	Likely (+1)
Preservation of identity (2x)	Likely (+1)	Likely (+1)	Possibly (0)	Possibly (0)	Likely (+1)

Correct balance (2x)	Unlikely (-1)	Unlikely (-1)	Possibly (0)	Built-in (+2)	Possibly (0)
Practical considerations (2x)	Unlikely (-1)	Unlikely (-1)	Likely (+1)	Likely (+1)	Likely (+1)
Overall compliance	**1**	**4**	**5**	**13**	**9**

I have given the name *virtue mapping* to the unlisted tacit desiderata that the framework contains traits which, if not virtuous themselves, have good prospects of being strongly correlated with virtues given some intuitive assumptions. Recapitulating the other desiderata, *practical robustness to moral uncertainty* refers to the framework being unlikely to generate catastrophic prescriptions even if the guiding moral theory is wrong or incomplete. I have divided *empirical adequacy* into two sub-desiderata, one relating to the framework containing general and stable dispositions of human moral psychology (*realistic traits*) and the other to these dispositions being feasibly targeted by technological intervention in the foreseeable future (*technological feasibility*). Being sensitive to potential losses of psychological continuity is filed under *preservation of identity*. *Correct balance* and *practical considerations* denote, respectively, having some common measure for evaluating and balancing competing traits, and being able to produce a practical analysis sensitive to context – addressing emergent societal effects being a relevant instance of the latter. Furthermore, I have created four levels of desiderata compliance. If a framework has an inherent feature that clearly satisfies a desideratum, it has that requirement *built-in*. In cases in which the framework seems prone to fulfil a requirement given some natural development of its research programme, the desideratum is *likely* to be complied with; whereas if it would need extra background assumptions to do so, the desideratum is *possibly*

complied with. Finally, if the framework not only has no immediate feature complying with a desideratum, but cannot naturally develop a solution to the requirement, nor be easily added with background assumptions to generate compliance, then it is deemed *unlikely* to satisfy that requirement. This last level will also include cases where it may be impossible for a framework to comply, but since proving full incompatibility would require extensive and careful analysis outside the scope of this chapter, I have grouped both levels together. A rough measure of overall desiderata compliance can be produced using the following system. Each desideratum is ascribed a relative weight. Given that identity, uncertainty, correct balance and practical considerations were the issues argued to lead to direct catastrophes – respectively, by creating a future without human continuants, overly enhancing a single trait, radically enhancing in the wrong direction and ignoring undesirable societal effects – they have double weight. All the others affect feasibility more than risks, therefore they have single weight.[113] Each of the four levels is paired with integers ranging from -1 to 2, the reasoning being that *unlikely* has negative value, *possibly* is neutral and *likely* and *built-in* have increasingly positive values.

Big five traits are not obviously morally desirable or undesirable, come in degrees, all but one is other-directed, everyone has them to some degree, and they might only be linguistic descriptors. Although with some further adaptations a second framework with virtuous traits could be built on top of the big five, they seem unlikely to be mapped onto virtuous traits themselves. There is no reliance on any moral-theoretic assumptions; hence, they are robust to moral uncertainty. Pure statistical analysis of large data produced them, so they seem to be realistic. However, it is still unclear whether they reveal a linguistic structure with no basis in a dispositional structure. It seems reasonable to assume it would be

[113] This means the frameworks are being analysed with an emphasis on safety. If the framework is highly safe but unlikely to be feasible, it will be harmless but useless.

evolutionary and culturally deleterious to have linguistic descriptors of dispositions without reasonable correlation to actual dispositions; therefore, they seem likely to be realistic traits. Research on their neurological basis has failed to find useful correlates; therefore, technological interventions in the foreseeable future seem unlikely. Self-assessed trait descriptions are arguably heavily tied to someone's identity; it seems likely they are relevant properties for personal identity. Identity would be likely to be preserved if someone knowingly and willingly enhanced them. Given the big five were conceptualised to have little correlation with each other and no good research on their interaction exists, it is unlikely that a correct balance or common measure could arise from them. Despite research correlating some of them to overall life outcomes, they are based on self-assessments instead of actual behaviour, most are not other-directed, and therefore the societal consequences of manipulating them are hard to predict. Proper practical considerations seem unlikely to arise from them at the current stage of research. As mentioned, the only main difference and advantage of character strengths over the big five is the fact they seem to be virtuous traits themselves; therefore, virtue mapping is built-in. This is the only evaluation of them I have changed.

Social Good CAPS are intentionally conceptualised as virtues; hence they have this feature built-in. Given that they need to rely on a definition of social good that might or might not depend on moral theory, they are possibly robust to moral uncertainty. They are empirically falsifiable, and as some early research has confirmed the existence of a few of them, they are likely to be realistic. However, there is no research into their neurochemical basis or manipulation, though it is possible such research will be conducted in the foreseeable future. It is unclear how related they are to personal identity, but it seems reasonable to expect they could be. Once a proper definition of social good is added, it is likely that it will be possible to measure and predict their practical effects on society.

The SVO framework has no built-in feature allowing it to be mapped onto virtues. However, given a proper evaluation of these orientations' societal effects and the provision of desirable societal outcomes, it is possible to ascribe moral desirability to them. There is no reliance on moral theory; consequently, SVO are certainly robust to moral uncertainty. Empirical research has demonstrated that SVO are reliable, stable and predictive in subjectively similar but objectively different situations (Balliet, Parks, & Joireman, 2009; de Dreu & van Lange, 1995; McClintock & Allison, 1989; Van Lange, Bekkers, Schuyt, & Van Vugt, 2007). One's identity would be likely tied to one's social orientations, but it is unlikely that social orientations would be anything more than a partial picture. To preserve identity, other relevant psychological properties would have to be considered, so SVO may be identity-preserving. The correct balance between traits is SVO's major strength given it provides a clear mathematical relationship between all dispositions and a common measure. The work that has already been done on social simulations and real-life social dilemmas shows practical considerations can be easily derived given proper research. The initial research on inserting a third dimension into the SVO ring tracking group-membership seems promising in that regard.

The stages of moral development show a progression towards higher morality by the development of increasingly sophisticated moral cognition. Moreover, there are studies showing their correlation with morally desirable or undesirable behaviour. They are uncontroversially mapped onto virtues. Such a framework has been commonly criticised as depending on normative assumptions when it evaluates the proposed later developmental stages as morally superior to previous ones. Arguably, one can build such a developmental framework without an inherent moral ranking; hence, I deem them possibly robust to moral uncertainty. There is extensive empirical evidence showing they are sequential, distinct, stable and reliable inside each stage; therefore, they are realistic. Given the extensive and

growing research on the neurobiological processes underlying human maturation, it seems likely we will be able to provide the proper neurological basis of each stage and induce them with technological intervention. In so far as they mirror natural development, they are likely to preserve personal identity. It is unclear how dispositions of previous stages are to be balanced or simply superseded, but such balance seems at least possible inside the framework. More extensive research is still lacking, but there are some studies correlating each stage with socially desirable or undesirable behaviour. On the other hand, it is still not clear how to predict or measure emergent societal effects; consequently, I have classified them as possibly producing proper practical considerations.

The overall compliance measure indicates the SVO framework, despite its limitations, is significantly superior to the other frameworks being considered. It reasonably complies with most desiderata, and, more importantly it has two double-weighted requirements built-in. The moral development and social good CAPS framework are, respectively, the second and third best alternatives, but seem to depend much more on future research to comply fully with several important requirements. It seems clear that despite being arguably the most reliable and validated trait theory, with regards to being a framework for moral enhancement, the big five have decidedly the lowest safety of the five alternatives. The big five-derived character strengths framework suffers from most of the same problems and is the second worst.

6.2 Final considerations

In order to avoid its own extreme risks, deep moral enhancement needs a safety framework incorporating the five plausible desiderata: practical robustness to moral uncertainty, empirical adequacy, correct balance, preservation of identity, and sensitivity to practical considerations. I examined in detail how these desiderata avoid the risks I discussed in Chapters 2 and 3. The relationship is evident in the following cases. Paradoxical effects

from increasing traits such as co-operation can be averted by aiming at a correct balance between connected traits such as individual cooperation and in-group favouritism. Moreover, empirical adequacy and sensitivity to practical considerations will require a careful and objective analysis of these effects. Self-reinforcing and irreversible effects on motivation can be avoided by aiming at a correct balance between motivations, and being robust to moral uncertainty prevents being overconfident about the desirability of any specific motivation. Empirical adequacy brings the focus to the epistemic complexity of moral traits. Preserving identity avoids undermining individual interests which are sensitive to abrupt losses to psychological connectedness.

A virtue theory framework can meet all of these desiderata. It does not make strong theoretical commitments, and so it is robust to moral uncertainty. Empirical research into moral traits indicates that a virtue theory fits better with the nature of moral psychology than competing moral theories, and virtue theories have traditionally focused on moral traits. The correct balance between dispositions is a central idea to most conceptions of virtue, which are also intimately related to personal identity and practical considerations. Safe deep moral enhancement is likely to mean virtue enhancement; it will require a guiding virtue framework to avoid extreme risks. A careful analysis of these desiderata, which also considers important objections to virtue theory, can objectively evaluate how different frameworks comply with safety requirements. All common proposals fail to comply entirely – especially the ones being proposed in the moral enhancement literature – while some stand a significant amount of scrutiny. For instance, the SVO model offers a promising direction for future investigation. Further research in moral psychology holds the promise of developing new improved models, but lies outside the scope of this thesis. One such model has been explored in parallel with this thesis and research efforts are still in progress. I also presented several important practical problems that lack a current satisfactory solution but

that could be addressed with a virtue framework, from terrorism to fostering liberal democracies. Regardless of how well any framework complies with the safety desiderata, it would still need to be implemented in order to work and this might dependent on certain institutional decisions.

Conclusion

Regardless of one's views on the many open problems of moral philosophy, it seems clear that we often do not behave in the way we think we should. A technological intervention that would enable us to substantially bridge this gap would be extremely desirable. It is equally clear that our moral dispositions are both vital to our identity and unusually complex when compared to other possible targets of human enhancement, such as physical endurance or short-term memory. Moral enhancement is both extremely desirable and unusually problematic. This thesis has attempted to solve this tension by first trying to pinpoint the extreme risks of moral enhancement, then by noting several current extreme risks that can be addressed with the use of safe moral enhancement, and finally by presenting a safety framework for its development.

Our present moral dispositions are unable to provide the level of large-scale co-operation necessary to deal with risks such as nuclear proliferation, drastic climate change and pandemics. In order to survive in an environment with powerful and easily available technologies, we need to improve our fundamental moral dispositions with moral enhancement.

However, moral enhancement would be in itself a powerful new technology and thus becomes extremely risky if managed improperly, perhaps to a degree comparable to nuclear or chemical weapons. I explored some of the reasons for such extreme risks, and I argued that deep moral enhancement is relatively prone to unexpected consequences, facing an unusually high degree of complexity. Moreover, deep moral enhancement might be detrimental to personal identity, thus undermining interests whose realisation depends on the existence of the same individuals, or type of individuals, who held those interests. If we use some form of virtue theory as a guiding framework for deep moral enhancement we likely to avoid these risks and to address other concerns present in the literature. If we apply such

a framework, deep moral enhancement is extremely desirable because it significantly decreases a wide range of extreme risks from our moral failings (e.g. nuclear war from lack of global co-operation). I will now summarise these substantive conclusions.

There is a large body of academic papers on moral enhancement, which has been proven to be fruitful grounds for practical ethical reflections. A review of the literature reveals that the common critiques of the project of moral enhancement can be placed into five groups. Moral enhancement is unfeasible, fails to improve morality, threatens freedom, has flawed reasons or produces undesirable social effects. However, there is a lack of unified conceptual framework and many of the critiques do not withstand scrutiny.

It is hard to oppose moral enhancement in the light of the seminal arguments in its favour presented in Persson & Savulescu and Douglas. Strong and rigorous critiques can still be made. If we take the bold aims of moral enhancement seriously (Persson & Savulescu's goal to decrease extreme risks for humanity in particular), moral enhancement could be catastrophic if done improperly.

Moral enhancement, when targeted at solving extreme risks, is likely to be deep moral enhancement – that is, it will be targeted at fundamental traits expected to lead to better moral behaviour or motives. Deep moral enhancement, in turn, is likely to lead to extreme risks in the absence of safety considerations. A focus on intuitively desirable traits can lead to long-term negative consequences.

The complexity of social interactions can lead to paradoxical emergent effects at the group level. Two exemplary cases of this are increasing co-operation and agreeableness, intuitively desirable traits that can lead to undesirable emergent effects, such as in-group favouritism, and compliance with unjust authority, respectively. I examined increasing co-operation in more detail and argued that increases in co-operation between individuals could aggravate extreme risks. It has been proposed that agreeableness might reinforce injustices.

As with other big five traits, it might be a linguistic artefact not reflecting a general dispositional pattern. Moreover, increasing morally desirable motivations seems uncontroversially good, but can lead to irreversible and self-reinforcing chains of modifications producing permanent and extreme motivational increases that are undesirable.

Many of our interests can be better realised by vastly better versions of ourselves, which provides a strong reason for performing deep moral enhancement. We might enhance moral traits so as to produce beings who are so drastically better than us that they come to enjoy a higher moral status, i.e. supra-persons, making us susceptible to being harmed for their sake. The realisation of some of our interests is sensitive to abrupt losses of psychological connections. The degree to which supra-persons can realise these interests is proportional to the degree of psychological connections between us and them. But increasing moral status does not imply preserving psychological connections. After supra-persons come into existence, it is unlikely we will be in a position to address this issue.

We have strong reasons for attempting deep moral enhancement. Many defensible positions in moral philosophy support the belief that risks in the distant future are of extreme moral relevance. Deep moral enhancement can help solve a large range of theses risks whose solution may not come through traditional means. But the arguments in Chapters 2 and 3 reveal that deep moral enhancement can also have negative long-term consequences, and thus there are also strong reasons against it.

This conflict can be resolved by providing a safety framework for deep moral enhancement. In order to solve the risks I have found, the framework should meet five desiderata: practical robustness to moral uncertainty, empirical adequacy, correct balance, preservation of identity, and sensitivity to practical considerations. A virtue theory will meet all of them by, respectively, having modest theoretical commitments, fitting well with empirical findings, and being conceptually bound to the ideas of correct balance, personal

identity, and practical reasoning. Possible frameworks incorporate these desiderata to variable degrees. Frameworks discussed in the moral enhancement literature fare poorly. The SVO framework is one of the most promising candidates. I derive exploratory applications for a virtue framework, ranging from economic stagnation to terrorism.

I did not offer a final framework in this thesis. Space precluded surveying all the possible extreme risks from deep moral enhancement. It is not reasonable to expect such an exhaustive survey to be forthcoming soon; that is a major project. I believe the safe and proper path for deep moral enhancement is a continuous probing of extreme risks which, when found, should then be accompanied by an appropriate update of the proposed safety framework. Accordingly, any proper virtue framework should be, by design, open-ended to philosophical reflection and ongoing empirical revision. Moreover, I have not explored solutions to the potential political obstacles of implementing a virtue framework. Initially, I believe that philosophical groundwork into a safety framework can help guide scientific research into potential moral enhancements. However, when and if any safety framework is implemented once moral enhancers are deployed in society will be a political decision that I have not attempted to foresee or analyse here.

The project of moral enhancement can be made coherent and safe if one follows a safety framework. It is imprudent to see it as an innocuous intervention without extreme long-term risks, or to dismiss it as an option and exclusively favour traditional moral progress. New models need to be developed with the aim of safely and feasibly performing deep moral enhancement. Current models are not sufficient, but indicate potentially fruitful directions. Attempting to close philosophical investigation of the project now as a result of overconfidence or dismissiveness towards its prospects would be catastrophic or imprudent, respectively. In so far as this thesis illustrates that surveying and safeguarding against risks can be accomplished, it makes the case for the moral desirability of deep moral enhancement

stronger, by revealing that it can withstand novel and challenging opposition. The thesis has presented a new and stronger case for moral enhancement by conceding its risks, analysing them in detail and providing a promising solution.

Bibliography

Aaldering, H. (2014). *Parochial and Universal Cooperation in Intergroup Conflicts*. Universiteit van Amsterdam. Retrieved from http://hdl.handle.net/11245/1.432600

Aaldering, H., Greer, L. L., Van Kleef, G. A., & De Dreu, C. K. W. (2013). Interest (mis)alignments in representative negotiations: Do pro-social agents fuel or reduce inter-group conflict? *Organizational Behavior and Human Decision Processes*, *120*(2), 240–250. http://doi.org/10.1016/j.obhdp.2012.06.001

Ackermann, K. (2012). SVO Ring. Wikimedia Commons. Reprinted courtesy of the Copyright Holder under a Creative Commons License CC BY-SA 3.0. Retrieved from https://upload.wikimedia.org/wikipedia/commons/7/72/SVO_Ring.pdf

Agar, N. (2010). *Humanity's end: why we should reject radical enhancement*. MIT Press.

Agar, N. (2013a). Moral bioenhancement is dangerous. *Journal of Medical Ethics*, 1–4. http://doi.org/10.1136/medethics-2013-101325

Agar, N. (2013b). Why is it possible to enhance moral status and why doing so is wrong? *Journal of Medical Ethics*, *39*(2), 67–74. http://doi.org/10.1136/medethics-2012-100597

Agar, N. (2014a). *Truly Human Enhancement: A Philosophical Defense of Limits*. MIT Press.

Agar, N. (2014b). A question about defining moral bioenhancement. *Journal of Medical Ethics*, *40*(6), 369–70. http://doi.org/10.1136/medethics-2012-101153

Agar, N. (2015). Moral bioenhancement and the utilitarian catastrophe. *Cambridge Quarterly of Healthcare Ethics*, *24*(1), 37–47. http://doi.org/10.1017/S0963180114000280

Aksoy, O. (2015). Social identity and social value orientations. *Unpublished Manuscript. Nuffield Centre for Experimental Social Sciences, Oxford, UK.*, 1–8. Retrieved from https://cess-web.nuff.ox.ac.uk/files/pdfs/working_papers/CESS_DP2014_002.pdf

Alexander, J. M. (2009). *The Structural Evolution of Morality*. *Economics and Philosophy* (Vol. 25). http://doi.org/Doi 10.1017/S0266267108002320

Allen, T. A., & DeYoung, C. G. (2016). Personality Neuroscience and the Five Factor Model. In T. Widiger (Ed.), *The Oxford Handbook of the Five Factor Model* (Vol. 1, pp. 1–59). Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199352487.013.26

Amadeo, K. (2018). The Ongoing Costs of the Iraq War. Retrieved July 24, 2018, from https://www.thebalance.com/cost-of-iraq-war-timeline-economic-impact-3306301

Anderl, C., Hahn, T., Klotz, C., & Rutter, B. (2015). Cooperative preferences fluctuate across the menstrual cycle. *Judgment and Decision Making*, *10*(5), 400–406.

Andrews, T., & Burke, F. (2007). What Does It Mean to Think Historically? *Perspectives on History | American Historical Association*, *15*(2).

Annas, J. (2007). Virtue Ethics. In D. Copp (Ed.), *The Oxford Handbook of Ethical Theory* (pp. 1–24). Oxford University Press. http://doi.org/10.1093/oxfordhb/9780195325911.003.0019

Archer, A. (2016). Moral Enhancement and Those Left Behind. *Bioethics*, *30*(7), 500–510. http://doi.org/10.1111/bioe.12251

Aristotle, Brown, L., & Ross, D. (2009). *The Nicomachean Ethics*. Oxford University Press.

Aristotle, & Rackham, H. (transl. . (1932). *Politics*. Harvard University Press.

Arrhenius, G. (2003). The person-affecting restriction, comparativism, and the moral status of potential people. *Ethical Perspectives*, *10*(3–4), 185–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16206457

Baertschi, B. (2014). Neuromodulation in the service of moral enhancement. *Brain Topography*, *27*(1), 63–71. http://doi.org/10.1007/s10548-012-0273-7

Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes and Intergroup Relations*, *12*(4), 533–547. http://doi.org/10.1177/1368430209105040

Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup Favoritism in Cooperation: A Meta-Analysis. *Psychological Bulletin*, *140*(6), 1556–1581. http://doi.org/10.1037/a0037737

Barilan, Y. M. (2015). Moral enhancement, gnosticism, and some philosophical paradoxes. *Cambridge Quarterly of Healthcare Ethics*, *24*(1), 75–85. http://doi.org/10.1017/S0963180114000322

Baumgartner, M. (2009). Inferring Causal Complexity. *Sociological Methods & Research*. http://doi.org/10.1177/0049124109339369

Beck, B. (2014). Conceptual and Practical Problems of Moral Enhancement. *Bioethics*, *9702*(3), 130–131. http://doi.org/10.1111/bioe.12090

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future*. Rutgers University.

Begue, L., Beauvois, J. L., Courbet, D., Oberle, D., Lepage, J., & Duke, A. A. (2015). Personality Predicts Obedience in a Milgram Paradigm. *Journal of Personality*, *83*(3), 299–306. http://doi.org/10.1111/jopy.12104

Bennett, A., & Elman, C. (2006). Complex Causal Relations and Case Study Methods: The Example of Path Dependence. *Political Analysis*, *14*, 250–267. http://doi.org/Doi 10.1093/Pan/Mpj020

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912–915. http://doi.org/10.1038/nature04981

Blass, T. (1999). The milgram paradigm after 35 years: Some things we now know about obedience to authority. *Journal of Applied Social Psychology*, *29*(5), 955–978. http://doi.org/10.1111/j.1559-1816.1999.tb00134.x

Bornstein, G. (2003). Intergroup Conflict: Individual, Group, and Collective Interests. *Personality and Social Psychology Review*, *7*(2), 129–145. http://doi.org/10.1207/S15327957PSPR0702_129-145

Bornstein, G., Gneezy, U., & Nagel, R. (2002). The effect of intergroup competition on group coordination: An experimental study. *Games and Economic Behavior*, *41*(1), 1–25. http://doi.org/10.1016/S0899-8256(02)00012-X

Bornstein, G., Winter, E., & Goren, H. (1996). Experimental study of repeated team-games. *European Journal of Political Economy*, *12*(4), 629–639. http://doi.org/10.1016/S0176-2680(96)00020-1

Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*.

Bostrom, N. (2003). Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*, *15*(3), 308–314.

Bostrom, N. (2004). The Future of Human Evolution. In *Death and anti-death: Two hundred years after Kant, fifty years after Turing* (pp. 339–371). Ria University Press. Retrieved from https://nickbostrom.com/fut/evolution.pdf

Bostrom, N. (2007). Technological Revolutions: Ethics and Policy in the Dark. In *Nanoscale: Issues and Perspectives for the Nano Centur* (Vol. 1, pp. 1–26).

Bostrom, N. (2009). The Future of Humanity. In J. Olsen, E. Selinger, & S. Riis (Eds.), *New Waves in Philosophy of Technology* (pp. 1–29). Palgrave McMillan.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2), 71–85. http://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31. http://doi.org/10.1111/1758-5899.12002

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bostrom, N., & Cirkovic, M. (Eds.). (2008). *Global catastrophic risks*. Oxford University Press.

Bostrom, N., Douglas, T., & Sandberg, A. (2016). The Unilateralist's Curse and the Case for a Principle of Conformity. *Social Epistemology*, *1728*, 1–22. http://doi.org/10.1080/02691728.2015.1108373

Bostrom, N., & Sandberg, A. (2008). The Wisdom of Nature : An Evolutionary Heuristic for Human Enhancement. In N. Bostrom & J. Savulescu (Eds.), *Human Enhancement* (pp. 375–416). Oxford University Press.

Bostrom, N., & Savulescu, J. (2008). Human Enhancement Ethics : The State of the Debate. In N. Bostrom & J. Savulescu (Eds.), *Human Enhancement* (pp. 1–22). Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Cambridge University Press.

Bowles, S., & Gintis, H. (2013). The Coevolution of Institutions and Behaviors. In *A cooperative species: Human Reciprocity and Its Evolution* (pp. 119–146). Princeton University Press.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, *100*(6), 3531–3535. http://doi.org/10.1073/pnas.0630443100

Brambilla, M., Hewstone, M., & Colucci, F. P. (2013). Enhancing moral virtues: Increased perceived outgroup morality as a mediator of intergroup contact effects. *Group Processes & Intergroup Relations*, *16*, 648–657. http://doi.org/10.1177/1368430212471737

Braumoeller, B. (2013). Is War Disappearing? In *APSA Chicago 2013 Meeting* (pp. 1–28). Retrieved from https://ssrn.com/abstract=2317269

Braumoeller, B. (2018). Systemic Trends In War And Peace. In O. Njolstad (Ed.), *The Causes of Peace: What We Now Know—Nobel Symposium 161*. The Norwegian Nobel Institute.

Brink, D. O. (2014). Principles and Intuitions in Ethics: Historical and Contemporary Perspectives. *Ethics*, *124*(4), 665–694. http://doi.org/10.1086/675878

Brooks, A. S., Yellen, J. E., Potts, R., Behrensmeyer, A. K., Deino, A. L., Leslie, D. E., … Clark, J. B. (2018). Long-distance stone transport and pigment use in the earliest Middle Stone Age. *Science*, *360*(6384), 90–94. http://doi.org/10.1126/science.aao2646

Buchanan, A. (2009). Moral Status and Human Enhancement. *Philosophy & Public Affairs*, *37*(4), 3–35.

Buchanan, A. (2011). *Beyond humanity?* Oxford University Press.

Buchanan, A., & Powell, R. (2015). The Limits of Evolutionary Explanations of Morality and Their Implications for Moral Progress. *Ethics*, *126*(1), 37–67.

Buchanan, A., & Powell, R. (2017). De-moralization as emancipation: Liberty, progress, and the evolution of invalid moral norms. *Social Philosophy and Policy*, *34*(2), 108–135. http://doi.org/10.1017/S0265052517000231

Buchanan, A., & Powell, R. (2018). *The evolution of moral progress: a biocultural theory*. Oxford University Press. http://doi.org/10.1093/oso/9780190868413.001.0001

Burton-Chellew, M. N., Ross-Gillespie, A., & West, S. a. (2010). Cooperation in humans: competition between groups and proximate emotions. *Evolution and Human Behavior*, *31*(2), 104–108. http://doi.org/10.1016/j.evolhumbehav.2009.07.005

Cardenas, J. C., & Mantilla, C. (2015). Between-group competition, intra-group cooperation and relative performance. *Frontiers in Behavioral Neuroscience*, *9*(February), 1–9. http://doi.org/10.3389/fnbeh.2015.00033

Carson, R. T., & Mitchell, R. C. (1993). The Issue of Scope in Contingent Valuation Studies. *American Journal of Agricultural Economics*, *75*(5), 1263. http://doi.org/10.2307/1243469

Carter, A. (2011). Some groundwork for a multidimensional axiology. *Philosophical Studies*, *154*(3), 389–408. http://doi.org/10.1007/s11098-010-9557-5

Carter, J. A., & Gordon, E. C. (2013). On Cognitive and Moral Enhancement: A Reply to Savulescu and Persson. *Bioethics*, *9702*(4). http://doi.org/10.1111/bioe.12076

Casal, P. (2015). On not taking men as they are: reflections on moral bioenhancement. *Journal of Medical Ethics*, *41*(4), 340–342. http://doi.org/10.1136/medethics-2013-101327

Casebeer, W. D. (2009). The Neurobiology of Virtue: Evolution, Cognition, and Human Flourishing. Retrieved November 20, 2017, from http://www.nourfoundation.com/media-gallery/videos/Toward-A-Common-Morality/The-Neurobiology-of-Virtue/William-Casebeer-PhD.html

Casebeer, W. D., & Churchland, P. S. (2003). The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and philosophy*, *18*(1), 169-194.

Castel, A. (1992). *Decision in the West: the Atlanta Campaign of 1864*. University Press of Kansas.

Chan, S., & Harris, J. (2011). Moral enhancement and pro-social behaviour. *Journal of Medical Ethics*, *37*(3), 130–1. http://doi.org/10.1136/jme.2010.041434

Chandler, M., & Moran, T. (1990). Psychopathy and moral development: A comparative study of delinquent and nondelinquent youth. *Development and Psychopathology*, *2*(3), 227–246. http://doi.org/10.1017/S0954579400000742

Cheikbossian, G. (2012). The collective action problem: Within-group cooperation and between-group competition in a repeated rent-seeking game. *Games and Economic Behavior*, *74*, 68–82. http://doi.org/10.1016/j.geb.2011.05.003

Chew, C., Douglas, T., & Faber, N. (2018). Biological Interventions for Crime Prevention. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice* (pp. 1–40). Oxford University Press.

Churchland, P. (2005). *Moral decision-making and the brain*. (J. Illes, Ed.)*Neuroethics: Defining the issues in theory, practice, and policy*. Oxford University Press. Retrieved from

http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198567219.001.0001/acprof-9780198567219-chapter-1

Cirillo, P., & Taleb, N. N. (2016). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and Its Applications*, *452*, 29–45. http://doi.org/10.1016/j.physa.2016.01.050

Cirkovic, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic Shadow: Observation Selection Effects and Human Extinction Risks. *Risk Analysis*, *30*(10), 1495–1506. http://doi.org/10.1111/j.1539-6924.2010.01460.x

Courtiol, A., Pettay, J. E., Jokela, M., Rotkirch, A., & Lummaa, V. (2012). Natural and sexual selection in a monogamous historical human population. *Proceedings of the National Academy of Sciences*, *109*(21), 8044–8049. http://doi.org/10.1073/pnas.1118174109

Cox, M., Arnold, G., & Villamayor, S. (2010). A Review of Design Principles for Community-based Natural Resource Management. *Ecology and Society*, *15*(4), 28. http://doi.org/38

Credit Suisse. (2017). *Global Wealth Databook 2017*. Retrieved from http://publications.credit-suisse.com/tasks/render/file/index.cfm?fileid=A8BD95FB-A213-1EE7-59CC7F2F001A11AF

Crisp, R. (1998). *How Should One Live?* (R. Crisp, Ed.). Oxford: Oxford University Press. http://doi.org/10.1093/0198752342.001.0001

Crittenden, P. (2002). On Virtue Ethics. *Australasian Journal of Philosophy*, *80*(1), 114–116. http://doi.org/10.1093/0199247994.001.0001

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. http://doi.org/10.1038/s41562-017-0213-3

Crockett, M. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. http://doi.org/10.1016/j.tics.2013.06.005

Crockett, M. (2014). Moral bioenhancement: a neuroscientific perspective. *Journal of Medical Ethics*, *40*(6), 370–1. http://doi.org/10.1136/medethics-2012-101096

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, *107*(40), 17433–17438. http://doi.org/10.1073/pnas.1009396107

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(48). http://doi.org/10.1073/pnas.1408988111

Cutter, W., Norbury, R., & Murphy, D. (2003). Oestrogen, brain function, and neuropsychiatric disorders. *Journal of Neurology, Neurosurgery, and Psychiatry*, *74*(7), 837–40. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1738534&tool=pmcentrez&rendertype=abstract

Dahlsgaard, K., Peterson, C., & Seligman, M. E. P. (2005). Shared virtue: The convergence of valued human strengths across culture and history. *Review of General Psychology*, *9*(3), 203–213. http://doi.org/10.1037/1089-2680.9.3.203

Dancy, J. (2004). *Ethics Without Principles*. Oxford: Oxford University Press. http://doi.org/10.1093/0199270023.001.0001

De Dreu, C. K. W., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. J. (2011). Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences*, *108*(4), 1262–1266. http://doi.org/10.1073/pnas.1015316108

De Dreu, C. K. W. (2012). Oxytocin modulates cooperation within and competition between groups: An integrative review and research agenda. *Hormones and Behavior*, *61*(3), 419–428. http://doi.org/10.1016/j.yhbeh.2011.12.009

De Dreu, C. K. W. (2013). Social conflict. In *Current Sociology* (Vol. 61, pp. 696–713). http://doi.org/10.1177/0011392113499487

De Dreu, C. K. W., Dussel, D. B., & Ten Velden, F. S. (2015). In intergroup conflict, self-sacrifice is stronger among pro-social individuals, and parochial altruism emerges especially among cognitively taxed individuals. *Frontiers in Psychology*, *6*(MAY), 1–9. http://doi.org/10.3389/fpsyg.2015.00572

De Dreu, C. K. W., Shalvi, S., Greer, L. L., van Kleef, G. a., & Handgraaf, M. J. J. (2012). Oxytocin Motivates Non-Cooperation in Intergroup Conflict to Protect Vulnerable In-Group Members. *PLoS ONE*, *7*(11). http://doi.org/10.1371/journal.pone.0046751

De Dreu, C. K. W., Balliet, D., & Halevy, N. (2014). Parochial Cooperation in Humans: Forms and Functions of Self- Sacrifice in Intergroup Conflict. In *Advances in Motivation Science* (pp. 1–47).

De Dreu, C. K. W., Greer, L. L., Handgraaf, M. J. J., Shalvi, S., Kleef, G. A. Van, Baas, M., … Feith, S. W. W. (2010). *The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflic Among Humans*. *Science* (Vol. 328).

De Dreu, C. K. W., & van Lange, P. A. M. (1995). The Impact of Social Value Orientations on Negotiator Cognition and Behavior. *Personality and Social Psychology Bulletin*, *21*(11), 1178–1188. http://doi.org/10.1177/01461672952111006

De Dreu, C. K. W., & Kret, M. E. (2015). Oxytocin Conditions Intergroup Relations Through Upregulated In-Group Empathy, Cooperation, Conformity, and Defense. *Biological Psychiatry*, (April 2016), 1–9. http://doi.org/10.1016/j.biopsych.2015.03.020

de Lazari-Radek, K., & Singer, P. (2012). The Objectivity of Ethics and the Unity of Practical Reason. *Ethics*, *123*(1), 9–31. http://doi.org/10.1086/667837

de Melo-Martin, I., & Salles, A. (2014). Moral Bioenhancement: Much Ado About Nothing? *Bioethics*, *9702*, 124–131. http://doi.org/10.1111/bioe.12100

de Raad, B., & Mlacic, B. (2016). The Lexical Foundation of the Big Five Factor Model. In *The Oxford Handbook of the Five Factor Model*. Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199352487.013.12

De Vos, J. M., Joppa, L. N., Gittleman, J. L., Stephens, P. R., & Pimm, S. L. (2015). Estimating the normal background rate of species extinction. *Conservation Biology*, *29*(2), 452–462. http://doi.org/10.1111/cobi.12380

Dees, R. H. (2008). Better Brains, Better Selves? The Ethics of Neuroenhancements. *Kennedy Institute of Ethics Journal*, *17*(4), 371–395. http://doi.org/10.1353/ken.2008.0001

DeGrazia, D. (2008). Moral status as a matter of degree? *Southern Journal of Philosophy*, *46*(2), 181–198. http://doi.org/10.1111/j.2041-6962.2008.tb00075.x

DeGrazia, D. (2014). On the Moral Status of Infants and the Cognitively Disabled: A Reply to Jaworska and Tannenbaum. *Ethics*, *124*(3), 543–556. http://doi.org/10.1086/675077

DeGrazia, D. (2014). Moral enhancement, freedom, and what we (should) value in moral behaviour. *Journal of Medical Ethics*, *40*(6), 361–8. http://doi.org/10.1136/medethics-2012-101157

Deigh, J. (2013). *Ethics in the Analytic Tradition*. Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199545971.013.0028

del Val, E., Rebollo, M., & Botti, V. (2013). Strategies for cooperation emergence in distributed service discovery. *Communications in Computer and Information Science*, *365*, 251–262. http://doi.org/10.1007/978-3-642-38061-7_25

Dolinski, D., Grzyb, T., Folwarczny, M., Grzybała, P., Krzyszycha, K., Martynowska, K., & Trojanowski, J. (2017). Would You Deliver an Electric Shock in 2015? Obedience in the Experimental Paradigm Developed by Stanley Milgram in the 50 Years Following the Original Studies. *Social Psychological and Personality Science*, *8*(8), 927–933. http://doi.org/10.1177/1948550617693060

Dorrough, A. R., Gl??ckner, A., Hellmann, D. M., & Ebert, I. (2015). The development of ingroup favoritism in repeated social dilemmas. *Frontiers in Psychology*, *6*(APR). http://doi.org/10.3389/fpsyg.2015.00476

Dorsey, D. (2015). The Significance of a Life's Shape. *Ethics*, *125*(2), 303–330.

Douglas, T. (2008). Moral Enhancement. *Journal of Applied Philosophy*, *25*(3), 228–245. http://doi.org/10.1111/j.1468-5930.2008.00412.x

Douglas, T. (2010). Intertemporal Disagreement and Empirical Slippery Slope Arguments. *Utilitas*. http://doi.org/10.1017/S0953820810000087

Douglas, T. (2013a). Human enhancement and supra-personal moral status. *Philosophical Studies*, *162*(3), 473–497. http://doi.org/10.1007/s11098-011-9778-2

Douglas, T. (2013b). Moral enhancement via direct emotion modulation: a reply to John Harris. *Bioethics*, *27*(3), 160–8. http://doi.org/10.1111/j.1467-8519.2011.01919.x

Douglas, T. (2014a). The Relationship Between Effort and Moral Worth: Three Amendments to Sorensen's Model. *Ethical Theory and Moral Practice*, *17*(2), 325–334. http://doi.org/10.1007/s10677-013-9441-4

Douglas, T. (2014b). The Morality of Moral Neuroenhancement. In J. Clausen & N. Levy (Eds.), *Handbook of Neuroethics*. Springer.

Douglas, T. (2014c). Enhancing Moral Conformity and Enhancing Moral Worth. *Neuroethics*, *7*, 75–91. http://doi.org/10.1007/s12152-013-9183-y

Douglas, T. (2015). The harms of enhancement and the conclusive reasons view. *Cambridge Quarterly of Healthcare Ethics : CQ : The International Journal of Healthcare Ethics Committees*, *24*(1), 23–36. http://doi.org/10.1017/S0963180114000218

Dovidio, J. F., Saguy, T., & Shnabel, N. (2009). Cooperation and conflict within groups: Bridging intragroup and intergroup processes. *Journal of Social Issues*, *65*(2), 429–449. http://doi.org/10.1111/j.1540-4560.2009.01607.x

Driver, J. (1998). The Virtues and Human Nature. In R. Crisp (Ed.), *How Should One Live?* Oxford: Oxford University Press. http://doi.org/10.1093/0198752342.003.0007

Duriez, B., Meeus, J., & Vansteenkiste, M. (2012). Why are some people more susceptible to ingroup threat than others? The importance of a relative extrinsic to intrinsic value orientation. *Journal of Research in Personality*, *46*(2), 164–172. http://doi.org/10.1016/j.jrp.2012.01.003

Earp, B. D., Sandberg, A., & Savulescu, J. (2014). Brave New Love: The Threat of High-Tech "Conversion" Therapy and the Bio-Oppression of Sexual Minorities. *AJOB Neuroscience*, *5*(1), 4–12. http://doi.org/10.1080/21507740.2013.863242

Earp, B. D., & Savulescu, J. (2017). Love drugs : Why scientists should study the effects of pharmaceuticals on human romantic relationships. *Technology in Society*, (February).

Earp, B. D., Skorburg, J. A., Everett, J. A. C., & Savulescu, J. (2018). Addiction, identity, morality. *PsyArXiv (Pre-Print)*, (June), 1–35. http://doi.org/10.17605/OSF.IO/EVM84

Earp, B. D., & Vierra, A. (2019). Sexual Orientation Minority Rights and High-Tech Conversion Therapy. In D. Booning (Ed.), *Handbook on Philosophy and Public Policy*. Palgrave Macmillan. Retrieved from https://www.academia.edu/36145698/Sexual_orientation_minority_rights_and_high-tech_conversion_therapy

Everett, J. A. C., Faber, N. S., Crockett, M. J., & De Dreu, C. K. W. (2015). Economic games and social neuroscience methods can help elucidate the psychology of parochial altruism. *Frontiers in Psychology*, *6*(July), 1–4. http://doi.org/10.3389/fpsyg.2015.00861

Everett, J. a. C., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in Behavioral Neuroscience*, *9*(February), 1–21. http://doi.org/10.3389/fnbeh.2015.00015

Everett, J. A. C., Faber, N. S., & Crockett, M. J. (2015). The influence of social preferences and reputational concerns on intergroup prosocial behaviour in gains and losses contexts. *Royal Society Open Science*, *2*(12), 150546. http://doi.org/10.1098/rsos.150546

Express Scripts. (2014). *Turning Attention To ADHD: US Medication Trends for Attention Deficint Hyperactivity Disorder*. Retrieved from http://lab.express-

scripts.com/insights/industry-updates/~/media/89fb0aba100743b5956ad0b5ab286110.ashx

Fabiano, J. (2014). *Human enhancement: an evolutionary heuristic and existential risks*. University of Sao Paulo.

Faust, H. S. (2008). Should we select for genetic moral enhancement? A thought experiment using the MoralKinder (MK+) haplotype. *Theoretical Medicine and Bioethics*, *29*(6), 397–416. http://doi.org/10.1007/s11017-008-9089-6

Fellner, G., & Lünser, G. K. (2014). Cooperation in local and global groups. *Journal of Economic Behavior and Organization*, (July). http://doi.org/10.1016/j.jebo.2014.02.007

Fenton, E. (2010). The perils of failing to enhance: a response to Persson and Savulescu. *Journal of Medical Ethics*, *36*(3), 148–51. http://doi.org/10.1136/jme.2009.033597

Gell-mann, M. (1995). What is complexity? *Complexity*, *1*(1).

Gell-mann, M., Lloyd, S., & Gell-mann, M. (2003). Effective Complexity Effective Complexity. *Santa Fe Institute Working Papers*.

Gilleen, J., Michalopoulou, P. G., Reichenberg, A., Drake, R., Wykes, T., Lewis, S. W., & Kapur, S. (2014). Modafinil combined with cognitive training is associated with improved learning in healthy volunteers – A randomised controlled trial. *European Neuropsychopharmacology*, *24*(4), 529–539. http://doi.org/10.1016/j.euroneuro.2014.01.001

Goette, L., Huffman, D., & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement. *American Economic Review*, *96*(2), 212–216. http://doi.org/10.1257/000282806777211658

Gour, N., & Lajoie, S. (2016). Epithelial Cell Regulation of Allergic Diseases. *Current Allergy and Asthma Reports*, *16*(9), 1–9. http://doi.org/10.1007/s11882-016-0640-7

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. http://doi.org/10.1037/a0015141

Grant, A. M., & Schwartz, B. (2011). Too much of a good thing: The challenge and opportunity of the inverted U. *Perspectives on Psychological Science*, *6*(1), 61–76. http://doi.org/10.1177/1745691610393523

Greene, J. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin Press.

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, *124*(4), 695–726. http://doi.org/10.1515/lehr-2015-0011

Griesinger, D. W., & Livingston, J. W. (1973). Toward a model of interpersonal motivation in experimental games. *Behavioral Science*, *18*(3), 173–188. http://doi.org/10.1002/bs.3830180305

Gunson, D., & McLachlan, H. (2013). Risk, Russian-roulette and lotteries: Persson and Savulescu on moral enhancement. *Medicine, Health Care, and Philosophy*, *16*(4), 877–84. http://doi.org/10.1007/s11019-012-9461-1

Guo, J., Wu, Y., Zhu, Z., Zheng, Z., Trzaskowski, M., Zeng, J., … Yang, J. (2018). Global genetic differentiation of complex traits shaped by natural selection in humans. *Nature Communications*, *9*(1), 1–9. http://doi.org/10.1038/s41467-018-04191-y

Gyngell, C., & Easteal, S. (2015). Cognitive Diversity and Moral Enhancement. *Cambridge Quarterly of Healthcare Ethics*, *24*(01), 66–74. http://doi.org/10.1017/S0963180114000310

Hagopian, A., Flaxman, A. D., Takaro, T. K., Esa Al Shatari, S. A., Rajaratnam, J., Becker, S., … Burnham, G. (2013). Mortality in Iraq Associated with the 2003-2011 War and Occupation: Findings from a National Cluster Sample Survey by the University Collaborative Iraq Mortality Study. *PLoS Medicine*, *10*(10). http://doi.org/10.1371/journal.pmed.1001533

Haidt, J. (2017). The Age of Outrage. Retrieved July 4, 2018, from https://www.city-journal.org/html/age-outrage-15608.html

Halevy, N., Chou, E. Y., Cohen, T. R., & Bornstein, G. (2010). Relative deprivation and intergroup competition. *Group Processes & Intergroup Relations*, *13*(6), 685–700. http://doi.org/10.1177/1368430210371639

Halevy, N., Bornstein, G., & Sagiv, L. (2008). "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm: Research article. *Psychological Science*, *19*(4), 405–411. http://doi.org/10.1111/j.1467-9280.2008.02100.x

Harman, G. (1999). Moral Philosophy Meets Social Psychology : Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society, New Series*, *99*(1999), 315–331. Retrieved from http://www.jstor.org/stable/4545312%0D

Harman, G. (2009). Skepticism about character traits. *Journal of Ethics*, *13*(2–3), 235–242. http://doi.org/10.1007/s10892-009-9050-6

Harpending, H., & Cochran, G. (2015). Genetics and Social Behavior. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, 1–15.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics*, *25*(2), 102–11. http://doi.org/10.1111/j.1467-8519.2010.01854.x

Harris, J. (2013a). “Ethics is for bad guys!” Putting the “moral” into moral enhancement. *Bioethics*, *27*(3), 169–73. http://doi.org/10.1111/j.1467-8519.2011.01946.x

Harris, J. (2013b). Moral progress and moral enhancement. *Bioethics*, *27*(5), 285–90. http://doi.org/10.1111/j.1467-8519.2012.01965.x

Harris, J., & Savulescu, J. (2015). A Debate about Moral Enhancement. *Cambridge Quarterly of Healthcare Ethics : CQ : The International Journal of Healthcare Ethics Committees*, *24*(1), 8–22. http://doi.org/10.1017/S0963180114000279

Hawks, J., Wang, E. T., Cochran, G. M., Harpending, H. C., & Moyzis, R. K. (2007). Recent acceleration of human adaptive evolution. *Proceedings of the National Academy of Sciences*, *104*(52), 20753–20758. http://doi.org/10.1073/pnas.0707650104

Helgason, A., Einarsson, A. W., Gumundsdóttir, V. B., Sigursson, Á., Gunnarsdóttir, E. D., Jagadeesan, A., … Stefánsson, K. (2015). The Y-chromosome point mutation rate in humans. *Nature Genetics*, *47*(5), 453–457. http://doi.org/10.1038/ng.3171

Henning, T. (2011). Moral Realism and Two-Dimensional Semantics. *Ethics*, *121*(4), 717–748. http://doi.org/10.1086/660695

Hertz, S. G., & Krettenauer, T. (2016). Does moral identity effectively predict moral behavior?: A meta-analysis. *Review of General Psychology*, *20*(2), 129–140. http://doi.org/10.1037/gpr0000062

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup Bias. *Annual Review of Psychology*, *53*(1), 575–604. http://doi.org/10.1146/annurev.psych.53.100901.135109

Hoffman, S. J., & Behdinan, A. (2015). Towards an International Treaty on Antimicrobial Resistance. *Ottawa L. Rev.*, *47*, 503–534.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. http://doi.org/10.1126/science.1251560

Holland, A. S., & Roisman, G. I. (2008). Big Five personality traits and relationship quality: Self-reported, observational, and physiological evidence. *Journal of Social and Personal Relationships*, *25*(5), 811–829. http://doi.org/10.1177/0265407508096697

Hughes, J. (2015). Moral Enhancement Requires Multiple Virtues. *Cambridge Quarterly of Healthcare Ethics*, *24*(01), 86–95. http://doi.org/10.1017/S0963180114000334

Human Security Research Group. (2013). *Human Security Report 2013 The Decline In Global Violence: Evidence, Explanation And Contestation*. Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/HSRP_Report_2013_140226_Web.pdf

Hursthouse, R. (2001). *On Virtue Ethics*. Oxford: Oxford University Press. http://doi.org/10.1093/0199247994.001.0001

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., … Puranen, B. (2014). World Values Survey: All Rounds—Country-Pooled Datafile Version. JD Systems Institute. Retrieved from http://www.worldvaluessurvey.org/WVSDocumentationWVL.js

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, *21*(3), 384–388. http://doi.org/10.1037/h0032317

Ito, T., Friedman, N., Bartholow, B., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a Comprehensive Understanding of Executive Cognitive Function in Implicit Racial Bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE*, *7*(8). http://doi.org/10.1371/journal.pone.0042366

Jaworska, A., & Tannenbaum, J. (2014). Person-Rearing Relationships as a Key to Higher Moral Status. *Ethics*, *124*(2), 242–271. http://doi.org/10.1086/673431

Jebari, K. (2014). What to Enhance: Behaviour, Emotion or Disposition? *Neuroethics*. http://doi.org/10.1007/s12152-014-9204-5

Joli, T. (2012). Climate change and human moral enhancement. Retrieved from https://bib.irb.hr/datoteka/701073.Jolic-Climate_change_and_human_moral_enhancement.pdf

Jotterand, F. (2011). "Virtue Engineering" and Moral Agency: Will Post-Humans Still Need the Virtues? *AJOB Neuroscience*, *2*(4), 3–9. http://doi.org/10.1080/21507740.2011.611124

Jotterand, F. (2014). Questioning the Moral Enhancement Project. *The American Journal of Bioethics*, *14*(4), 1–3. http://doi.org/10.1080/15265161.2014.905031

Joutsa, J., Voon, V., Johansson, J., Niemelä, S., Bergman, J., & Kaasinen, V. (2015). Dopaminergic function and intertemporal choice. *Translational Psychiatry*, *5*(1), e491–e491. http://doi.org/10.1038/tp.2014.133

Joyce, R. (2013). Unfit for the Future: The Need for Moral Enhancement * By INGMAR PERSSON AND JULIAN SAVULESCU. *Analysis*, *73*(3), 587–589. http://doi.org/10.1093/analys/ant021

Kafri, R., Levy, M., & Pilpel, Y. (2006). The regulatory utilization of genetic redundancy through responsive backup circuits. *Proceedings of the National Academy of Sciences*, *103*(31), 11653–11658. http://doi.org/10.1073/pnas.0604883103

Kahane, G., & Savulescu, J. (2012). The Concept of Harm and the Significance of Normality. *Journal of Applied Philosophy*, *29*(4), 318–332. http://doi.org/10.1111/j.1468-5930.2012.00574.x

Kay, D., & Nelkin, D. K. (2015). Psychopaths , Incorrigible Racists , and the Faces of Responsibility. *Ethics*, *125*(2), 357–390.

Kelso, A. (1994). The enigma of cytokine redundancy. *Immunology and Cell Biology*, *72*, 97–101.

Koc, A., Gasch, A. P., Rutherford, J. C., Kim, H.-Y., & Gladyshev, V. N. (2004). Methionine sulfoxide reductase regulation of yeast lifespan reveals reactive oxygen species-dependent and -independent components of aging. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(21), 7999–8004. http://doi.org/10.1073/pnas.0307929101

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory Into Practice*, *16*(2), 53–59. http://doi.org/10.1080/00405847709542675

Kollock, P. (1998). Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, *24*(1), 183–214. http://doi.org/10.1146/annurev.soc.24.1.183

Komi, D., Sharma, L., & Dela Cruz, C. S. (2017). Chitin and Its Effects on Inflammatory and Immune Responses. *Clinical Reviews in Allergy and Immunology*, *54*(2), 1–11. http://doi.org/10.1007/s12016-017-8600-0

Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford University Press.

Korsgaard, C. M. (2018). *Fellow Creatures: Our Obligations to the Other Animals*. *Uehiro Series in Practical Ethics*. Oxford: Oxford University Press. http://doi.org/10.1093/oso/9780198753858.001.0001

Krebs, D. (2011). *The Origins of Morality*. Oxford University Press.

Krebs, D. (2015). The Evolution of Morality. In *The handbook of evolutionary psychology.* (pp. 747–771). John Wiley & Sons.

Krsková, L., & Talaroviová, A. (2005). Influence of maternal testosterone on the strategies in the open field behaviour of rats. *Neuro Endocrinology Letters*, *26*(2), 121–4. http://doi.org/NEL260205A03 [pii]

Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: An integrative quantitative model of the adult externalizing spectrum. *Journal of Abnormal Psychology*, *116*(4), 645.

Kupferschmidt, K. (2017). How Canadian researchers reconstituted an extinct poxvirus for $100,000 using mail-order DNA. *Science*, *9*(2), 8. http://doi.org/10.1126/science.aan7069

Lechner, S. (2014). Why Moral Bioenhancement Is a Bad Idea and Why Egalitarianism Would Make It Worse. *The American Journal of Bioethics*, *14*(4), 31–32. http://doi.org/10.1080/15265161.2014.889252

Levy, N., Douglas, T., Kahane, G., Terbeck, S., Cowen, P. J., Hewstone, M., & Savulescu, J. (2014). Are You Morally Modified?: The Moral Effects of Widely Used Pharmaceuticals. *Philosophy, Psychiatry, & Psychology*, *21*(2), 111–125. http://doi.org/10.1353/ppp.2014.0023

Lewis, D. (1983). Survival and Identity. In *Philosophical Papers Volume I*. New York: Oxford University Press. http://doi.org/10.1093/0195032047.003.0005

Lewis, D. (1986). Causal Explanation. *Philosophical Papers, Volume II*. Oxford University Press.

Marichal, T., Starkl, P., Reber, L. L., Kalesnikoff, J., Oettgen, H. C., Tsai, M., … Galli, S. J. (2013). A beneficial role for immunoglobulin E in host defense against honeybee venom. *Immunity*, *39*(5), 963–975. http://doi.org/10.1016/j.immuni.2013.10.005

Marshall, F. (2014). SSRIs as Moral Enhancers: Conceptual Clarification Needed in Defining Moral Enhancement. *AJOB Neuroscience*, *5*(3), 31–32. http://doi.org/10.1080/21507740.2014.911218

Marshall, F. (2014). Would Moral Bioenhancement Lead to an Inegalitarian Society? *The American Journal of Bioethics*, *14*(4), 29–30. http://doi.org/10.1080/15265161.2014.889253

Martin, B. (1982). Critique of Nuclear Extinction. *Journal of Peace Research*, *19*(4), 287–300. http://doi.org/10.1177/002234338201900401

Mcallister, J. W. (2003). Effective Complexity as a Measure of Information Content. *Philosophy of Science*, *70*(2), 302–307.

McClintock, C. G., & Allison, S. T. (1989). Social value orientation and helping behavior. *Journal of Applied Social Psychology*, *19*, 353–362. http://doi.org/10.1111/j.1559-1816.1989.tb00060.x

McCrae, R. R., & Terracciano, A. (2005). Personality profiles of cultures: aggregate personality traits. *Journal of Personality and Social Psychology*, *89*(3), 407–425. http://doi.org/10.1037/0022-3514.89.3.407

McIntyre, A. (1981). *After Virtue: A Study in Moral Theory*. University of Notre Dame Press.

McMahan, J. (2002). *The Ethics of Killing*. Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199396078.001.0001

McMahan, J. (2005). Our fellow creatures. *Journal of Ethics*, *9*(3–4), 353–380. http://doi.org/10.1007/s10892-005-3512-2

McMahan, J. (2009). *Killing in War*. Oxford: Claredon Press.

McMahan, J. (2010). Cognitive Disability and Cognitive Enhancement. *Cognitive Disability and Its Challenge to Moral Philosophy*, 345–367. http://doi.org/10.1002/9781444322781.ch20

McMahan, J. (1981). Problems of Population Theory. *Ethics*, *92*(1), 96–127. Retrieved from http://www.jstor.org/stable/2380707

Mela, L. (2011). MacIntyre on Personal Identity. *Public Reason*, *3*(1), 103–113. Retrieved from http://www.publicreason.ro/pdf/6#page=105

Mikhail, J. (2014). Any Animal Whatever ? Harmful Battery and Its Elements as Building Blocks of Moral Cognition. *Ethics*, *124*(4), 750–786. http://doi.org/10.1086/675906

Milgram, S. (1974). *Obedience to Authoruty: An Experimental View*. Harper & Row. http://doi.org/10.1177/0094306118779813a

Miller, C. B. (2017). Empirical Approaches to Moral Character. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2017/entries/moral-character-empirical

Mischel, W., & Shoda, Y. (1995). A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure. *Psychological Review*, *102*(2), 246–268. http://doi.org/10.1037/0033-295X.102.2.246

Moll, J., De Oliveira-Souza, R., Basilio, R., Bramati, I. E., Gordon, B., Rodríguez-Nieto, G., … Grafman, J. (2018). Altruistic decisions following penetrating traumatic brain injury. *Brain*, *141*(5), 1558–1569. http://doi.org/10.1093/brain/awy064

Moll, J., De Oliveira-Souza, R., & Zahn, R. (2008). The neural basis of moral cognition: Sentiments, concepts, and values. *Annals of the New York Academy of Sciences*, *1124*, 161–180. http://doi.org/10.1196/annals.1440.005

Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, *93*, 1–17. http://doi.org/10.1016/j.cogpsych.2016.11.006

Moon, S., Sridhar, D., Pate, M. A., Jha, A. K., Clinton, C., Delaunay, S., … Piot, P. (2015). Will Ebola change the game? Ten essential reforms before the next pandemic. the report of the Harvard-LSHTM Independent Panel on the Global Response to Ebola. *The Lancet*, *386*(10009), 2204–2221. http://doi.org/10.1016/S0140-6736(15)00946-0

Morioka, M. (2014). Some Remarks on Moral Bioenhancement. In *The Future of Bioethics: International Dialogues*. http://doi.org/10.1093/acprof

Murphy, R. O., Ackermann, K., & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, *6*(8), 771–781. Retrieved from http://dx.doi.org/10.2139/ssrn.1804189

Murphy, T. F. (2014). Preventing Ultimate Harm as the Justification for Biomoral Modification. *Bioethics*, *9702*. http://doi.org/10.1111/bioe.12108

Neal, A., Yeo, G., Koy, A., & Xiao, T. (2012). Predicting the form and direction of work role performance from the Big 5 model of personality traits. *Journal of Organizational Behavior*, *33*(2), 175–192. http://doi.org/10.1002/job.742

Nebel, J. M. (2015). Status Quo Bias , Rationality , and Conservatism about Value. *Ethics*, *125*(2), 449–476.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96–125. http://doi.org/10.1111/cogs.12134

Nichols, S. (2014). Process Debunking and Ethics. *Ethics*, *124*(4), 727–749. http://doi.org/10.1086/675877

Nussbaum, M. (2001). *The Fragility of Goodness*. Cambridge University Press.

O'Connor, B. P. (2002). A Quantitative Review of the Comprehensiveness of the Five-Factor Model in Relation to Popular Personality Inventories. *Assessment*, *9*(2), 188–203. http://doi.org/10.1177/10791102009002010

Omohundro, S. M. (2008). The Basic AI Drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence* (pp. 483–492). Amsterdam, The Netherlands, The Netherlands: IOS Press. Retrieved from http://dl.acm.org/citation.cfm?id=1566174.1566226

Parfit, D. (1986). *Reasons and Persons*. Oxford: Oxford University Press. http://doi.org/10.1093/019824908X.001.0001

Paulo, N., & Bublitz, J. C. (2017). How (not) to Argue For Moral Enhancement: Reflections on a Decade of Debate. *Topoi*, *0*(0), 1–15. http://doi.org/10.1007/s11245-017-9492-6

Perheentupa, A., Makinen, J., Laatikainen, T., Vierula, M., Skakkebaek, N. E., Andersson, A.-M., & Toppari, J. (2013). A cohort effect on serum testosterone levels in Finnish men. *European Journal of Endocrinology*, *168*(2), 227–233. http://doi.org/10.1530/EJE-12-0288

Persson, I. (2001). Equality, Priority and Person-Affecting Value. *Ethical Theory and Moral Practice*, *4*(1), 23–39.

Persson, I. (2012). Could it be permissible to prevent the existence of morally enhanced people? *Journal of Medical Ethics*, *38*(11), 692–693. http://doi.org/10.1136/medethics-2012-100831

Persson, I., & Savulescu, J. (2008). The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity. *Journal of Applied Philosophy*, *25*(3), 162–177. http://doi.org/10.1111/j.1468-5930.2008.00410.x

Persson, I., & Savulescu, J. (2010). Moral Transhumanism. *Journal of Medicine and Philosophy*, *35*(6), 656–669. http://doi.org/10.1093/jmp/jhq052

Persson, I., & Savulescu, J. (2011). The turn for ultimate harm: a reply to Fenton. *Journal of Medical Ethics*, *37*(7), 441–444. http://doi.org/10.1136/jme.2010.036962

Persson, I., & Savulescu, J. (2012). *Unfit for the Future*. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199653645.001.0001

Persson, I., & Savulescu, J. (2013). Getting moral enhancement right: the desirability of moral bioenhancement. *Bioethics*, *27*(3), 124–31. http://doi.org/10.1111/j.1467-8519.2011.01907.x

Persson, I., & Savulescu, J. (2014a). Should moral bioenhancement be compulsory? Reply to Vojin Rakic. *Journal of Medical Ethics*, *40*(4), 251–252. http://doi.org/10.1136/medethics-2013-101423

Persson, I., & Savulescu, J. (2014b). Against Fetishism About Egalitarianism and in Defense of Cautious Moral Bioenhancement. *The American Journal of Bioethics*, *14*(4), 39–42. http://doi.org/10.1080/15265161.2014.889248

Persson, I., & Savulescu, J. (2015). Summary of Unfit for the Future. *Journal of Medical Ethics*, *41*(4), 338–339. http://doi.org/10.1136/medethics-2013-101323

Persson, I., & Savulescu, J. (2017). Moral hard-wiring and moral enhancement. *Bioethics*, *31*(4), 286–295. http://doi.org/10.1111/bioe.12314

Persson, I., & Savulescu, J. (2018). The Moral Importance of Reflective Empathy. *Neuroethics*, *11*(2), 183–193. http://doi.org/10.1007/s12152-017-9350-7

Peterson, C., & Park, N. (2012). Classifying and Measuring Strengths of Character. In S. Lopez & C. R. Snyder (Eds.), *The Oxford Handbook of Positive Psychology, (2 Ed.)*. Oxford University Press. http://doi.org/10.1093/oxfordhb/9780195187243.013.0004

Portenoy, R. K., Jarden, J. O., Sidtis, J. J., Lipton, R. B., Foley, K. M., & Rottenberg, D. A. (1986). Compulsive thalamic self-stimulation: A case with metabolic, electrophysiologic and behavioral correlates. *Pain*, *27*(3), 277–290. http://doi.org/10.1016/0304-3959(86)90155-7

Powell, R. (2013). The biomedical enhancement of moral status. *Journal of Medical Ethics*, *39*(2), 65–66. http://doi.org/10.1136/medethics-2012-101312

Powell, S. K. (2014). SSRIs as a Component of, Rather Than Exclusive Means to, Moral Enhancement. *AJOB Neuroscience*, *5*(3), 33–34. http://doi.org/10.1080/21507740.2014.911215

Pratt, L. A., Brody, D. J., & Gu, Q. (2011). Antidepressant use in persons aged 12 and over: United States, 2005-2008. *NCHS Data Brief*, *127*(76), 1–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22617183

Prinz, J. (2011). Against empathy. *Southern Journal of Philosophy*, *49*(SUPPL. 1), 214–233. http://doi.org/10.1111/j.2041-6962.2011.00069.x

Pugh, J., & Douglas, T. (2016). Justifications for Non-Consensual Medical Intervention: From Infectious Disease Control to Criminal Rehabilitation. *Criminal Justice Ethics*, *35*(3), 205–229. http://doi.org/10.1080/0731129X.2016.1247519

Puurtinen, M., & Mappes, T. (2009). Between-group competition and human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1655), 355–360. http://doi.org/10.1098/rspb.2008.1060

Railton, P. (2014). The Affective Dog and Its Rational Tale: Intuition and Attunement. *Ethics*, *124*(4), 813–859. http://doi.org/10.1086/675876

Rakić, V. (2015). We Must Create Beings with Moral Standing Superior to Our Own. *Cambridge Quarterly of Healthcare Ethics*, *24*(01), 58–65. http://doi.org/10.1017/S0963180114000309

Rakić, V., & Hughes, J. (2015). Guest Editorial. *Cambridge Quarterly of Healthcare Ethics*, *24*(01), 3–6. http://doi.org/10.1017/S0963180114000267

Ram-Tiktin, E. (2014). The Possible Effects of Moral Bioenhancement on Political Privileges and Fair Equality of Opportunity. *The American Journal of Bioethics*, *14*(4), 43–44. http://doi.org/10.1080/15265161.2014.889246

Rapoport, A., & Bornstein, G. (1987). Intergroup competition for the provision of binary public goods. *Psychological Review*, *94*(3), 291–299. http://doi.org/10.1037/0033-295X.94.3.291

Raus, K., Focquaert, F., Schermer, M., Specker, J., & Sterckx, S. (2014). On Defining Moral Enhancement: A Clarificatory Taxonomy. *Neuroethics*, *7*(3), 263–273. http://doi.org/10.1007/s12152-014-9205-4

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (2000). A Neo-Kohlbergian Approach to Morality Research. *Journal of Moral Education*, *29*(4), 381–395. http://doi.org/10.1080/713679390

Richards, R. A. (2016). Sexual Selection : Its Possible Contribution to Recent Human Evolution. *ELS. John Wiley & Son*. http://doi.org/10.1002/9780470015902.a0021788.pub2

Richardson, H. S. (2015). Introduction. *Ethics*, *124*(4), 659–664.

Roberts, M. A. (2011). The nonidentity Problem. In *Stanford Encyclopedia of Philosophy* (pp. 1–11). The Metaphysics Research Lab.

Robichaud, P. (2014). Moral Capacity Enhancement Does Not Entail Moral Worth Enhancement. *The American Journal of Bioethics*, *14*(4), 33–34. http://doi.org/10.1080/15265161.2014.889251

Roncarati, M., Bridges, P., Brassey, A., Creaby, C., Holder, G., & Unwin, A. (2006). Does the use of Inter-Group Competition Enhance Performance in Short-term Summative Testing? *Teaching Business & Economics*. Retrieved from http://eprints.ioe.ac.uk/251/

Ross, D. (1930). *The Right and the Good*. (P. Stratton-Lake, Ed.) (2002nd ed.). Oxford University Press.

Rubin, M., Badea, C., & Jetten, J. (2014). Low status groups show in-group favoritism to compensate for their low status and to compete for higher status. *Group Processes and Intergroup Relations*, *17*(5), 563–576. http://doi.org/10.1177/1368430213514122

Sahakian, B. J., & Morein-Zamir, S. (2015). Pharmacological cognitive enhancement: treatment of neuropsychiatric disorders and lifestyle use by healthy people. *The Lancet Psychiatry*, *2*(4), 357–362. http://doi.org/10.1016/S2215-0366(15)00004-8

Sandberg, A., & Fabiano, J. (2017). Modeling the social dynamics of moral enhancement: Social strategies sold over the counter and the stability of society. In *Cambridge*

Quarterly of Healthcare Ethics (Vol. 26, pp. 431–445). http://doi.org/10.1017/S0963180116001109

Savelkoul, M., Hewstone, M., Scheepers, P., & Stolle, D. (2015). Does relative out-group size in neighborhoods drive down associational life of Whites in the U.S.? Testing constrict, conflict and contact theories. *Social Science Research*, *52*, 236–252. http://doi.org/http://dx.doi.org/10.1016/j.ssresearch.2015.01.013

Savulescu, J. (2009). Moral Status of Enhanced Beings: What Do We Owe the Gods? In J. Savulescu & N. Bostrom (Eds.), *Human Enhancement*. Oxford University Press.

Savulescu, J., & Persson, I. (2012). Moral Enhancement, Freedom, and the God Machine. *The Monist*, *95*(3), 399–421.

Schaefer, G. O., & Savulescu, J. (2016). Procedural Moral Enhancement. *Neuroethics*, 1–12. http://doi.org/10.1007/s12152-016-9258-7

Scheffler, S. (2018). *Why Worry About Future Generations? Uehiro Series in Practical Ethics* (Vol. 1). Oxford: Oxford University Press. http://doi.org/10.1093/oso/9780198798989.001.0001

Schwartz, S. H. (1992). Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In *Advances in Experimental Social Psychology* (Vol. 25, pp. 1–65). http://doi.org/10.1016/S0065-2601(08)60281-6

Selgelid, M. J. (2014). Freedom and moral enhancement. *Journal of Medical Ethics*, *40*(4), 215–216. http://doi.org/10.1136/medethics-2014-102111

Shahrier, S., Kotani, K., & Kakinaka, M. (2016). Social Value Orientation and Capitalism in Societies. *PLOS ONE*, *11*(10), e0165067. http://doi.org/10.1371/journal.pone.0165067

Sherman, N. (2013). *Moral Psychology and Virtue*. Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199545971.013.0035

Sherman, P. W., Holland, E., & Sherman, J. S. (2008). Allergies: Their Role in Cancer Prevention. *The Quarterly Review of Biology*, *83*(4), 339–362. http://doi.org/10.1086/592850

Shi, Z. (2014). *Are Groups More Pro-Self Than Individuals ? Individual-Group Comparisons on Social Value Orientation and Ethical Decision Making*. Loyola University Chicago.

Shoemaker, D. (2016). Personal Identity and Ethics. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2016/entries/identity-ethics/

Shook, J. R. (2012). Neuroethics and the Possible Types of Moral Enhancement. *AJOB Neuroscience*, *3*(4), 3–14. http://doi.org/10.1080/21507740.2012.712602

Sidanius, J., & Veniegas, R. C. (2000). Gender and race discrimination: The interactive nature of disadvantage. In S. Oskamp (Ed.), *Reducing Prejudice and Discrimination The Claremont Symposium on Applied Social Psychology* (pp. 47–69). Lawrence Erlbaum Associates.

Smith, M., Lewis, D., & Johnston, M. (1989). Dispositional Theories of Value. In *Proceedings of the Aristotelian Society, Supplementary Volumes* (Vol. 63, pp. 89–174). Retrieved from https://www.jstor.org/stable/4106918%0D

Snarey, J. R. (1985). Cross-cultural universality of social-moral development: A critical review of Kohlbergian research. *Psychological Bulletin*, *97*(2), 202–232. http://doi.org/10.1037//0033-2909.97.2.202

Snow, N. (2010). *Virtue as Social Intelligence: An Empirically Grounded Theory*. Routledge.

Sorensen, K. (2014). Moral Enhancement and Self-Subversion Objections. *Neuroethics*, *7*(3), 275–286. http://doi.org/10.1007/s12152-014-9208-1

Sparrow, R. (2014). Unfit for the Future: The Need for Moral Enhancement, by Persson, Ingmar, and Julian Savulescu. *Australasian Journal of Philosophy*, *92*(2), 404–407. http://doi.org/10.1080/00048402.2013.860180

Sparrow, R. (2014). Better Living Through Chemistry? A Reply to Savulescu and Persson on 'Moral Enhancement.' *Journal of Applied Philosophy*, *31*(1), 23–32. http://doi.org/10.1111/japp.12038

Specker, J., Focquaert, F., Raus, K., Sterckx, S., & Schermer, M. (2014). The ethical desirability of moral bioenhancement : a review of reasons. *BMC Medical Ethics*, *15*(4).

Starmans, C., & Bloom, P. (2018). Nothing Personal: What Psychologists Get Wrong about Identity. *Trends in Cognitive Sciences*, 85–87. http://doi.org/10.1016/j.tics.2018.04.002

Streiffer, R. (2005). At the Edge of Humanity: Human Stem Cells, Chimeras, and Moral Status. *Kennedy Institute of Ethics Journal*, *15*(4), 347–370. http://doi.org/10.1353/ken.2005.0030

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. http://doi.org/10.1016/j.cognition.2013.12.005

Suderow, B. A. (1997). The Battle of the Crater: The Civil War's Worst Massacre. *Civil War History*, *43*(3), 219–224.

Summers, L. (2016, March). The Age of Secular Stagnation. *Foreign Affairs*. Retrieved from https://www.foreignaffairs.com/articles/united-states/2016-02-15/age-secular-stagnation

Temkin, L. S. (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. *Oxford Ethics Series*. Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199759446.001.0001

Teulings, C., & Baldwin, R. (2014). *Secular Stagnation: Facts, Causes and Cures. VoxEU.org eBook*. Retrieved from http://www.voxeu.org/sites/default/files/Vox_secular_stagnation.pdf

Thomas, W. G. (2011). *The Iron Way: Railroads, the Civil War, and the Making of Modern America*. Yale University Press.

Travison, T. G., Araujo, A. B., O'Donnell, A. B., Kupelian, V., & McKinlay, J. B. (2007). A Population-Level Decline in Serum Testosterone Levels in American Men. *The Journal of Clinical Endocrinology & Metabolism*, *92*(1), 196–202. http://doi.org/10.1210/jc.2006-1375

Turnbull, S. (2010). *The Mongol Invasions of Japan, 1274 and 1281*. Osprey Publishing.

Vallentyne, P. (1988). Teleology, consequentialism, and the past. *The Journal of Value Inquiry*, *22*(2), 89–101. http://doi.org/10.1007/BF00135455

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327. http://doi.org/10.1016/j.jrp.2010.03.003

Van Hooft, S. (2014). Virtue and identity. In S. Van Hooft (Ed.), *The Handbook of Virtue Ethics* (pp. 153–162). Acumen Publishing. Retrieved from

https://www.cambridge.org/core/books/handbook-of-virtue-ethics/virtue-and-identity/A48F35337D52FD3A29919D50F5E388A6

Van Lange, P. A. M., Bekkers, R., Schuyt, T. N. M., & Van Vugt, M. (2007). From games to giving: Social value orientation predicts donations to noble causes. *Basic and Applied Social Psychology*, *29*(4), 375–384. http://doi.org/10.1080/01973530701665223

Van Lange, P. A. M., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*(2), 125–141. http://doi.org/10.1016/j.obhdp.2012.11.003

Velleman, J. D. (1991). Well-Being and Time. *Pacific Philosophical Quarterly*, *72*(1), 48–77. http://doi.org/10.1111/j.1468-0114.1991.tb00410.x

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*(3), 0446–0458. http://doi.org/10.1371/journal.pbio.0040072

Volokh, E. (2003). The Mechanisms of the Slippery Slope. *Harvard Law Review*, *116*(4), 1026. http://doi.org/10.2307/1342743

Walker, L. J. (1982). The Sequentiality of Kohlberg's Stages of Moral Development. *Child Development*, *53*(5), 1330. http://doi.org/10.2307/1129023

Warren, M. A. (2000). The Concept of Moral Status. In *Moral Status*. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780198250401.003.0001

Wasserman, D. (2013). Devoured by our own children: the possibility and peril of moral status enhancement. *Journal of Medical Ethics*, *39*(2), 78–9. http://doi.org/10.1136/medethics-2012-100843

Wasserman, D. (2014). When bad people do good things: will moral enhancement make the world a better place? *Journal of Medical Ethics*, *40*(6), 374–375. http://doi.org/10.1136/medethics-2012-101094

Weisel, O. (2015). Negative and positive externalities in intergroup conflict: Exposure to the opportunity to help the outgroup reduces the inclination to harm it. *Frontiers in Psychology*, *6*(OCT). http://doi.org/10.3389/fpsyg.2015.01594

Wikler, D. (1979). Paternalism and the Mildly Retarded. *Philosophy and Public Affairs*, *8*(4), 377–392. Retrieved from http://www.jstor.org/stable/2265070

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. a, & Schopler, J. (2003). Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect. *Psychological Bulletin*, *129*(5), 698–722. http://doi.org/10.1037/0033-2909.129.5.698

Williams, B. (1985). Morality, the Peculiar Institution. In *Ethics and the Limits of Philosophy*. Routledge.

Williams, B. (2008). The Human Prejudice. In B. Williams (Ed.), *Philosophy as a Humanistic Discipline* (pp. 135–152). Princeton University Press. Retrieved from http://www.jstor.org/stable/j.ctt7rx9w.17

Wills, C. (2008). Evolution theory and the future of humanity. In N. Bostrom & M. Cirkovic (Eds.), *Global Catastrophic Risks*. Oxford University Press.

Wilson, A. T. (2014). Egalitarianism and Successful Moral Bioenhancement. *The American Journal of Bioethics*, *14*(4), 35–36. http://doi.org/10.1080/15265161.2014.889250

Wilson, D. S., & Wilson, E. O. (2008). Evolution "for the Good of the Group." *American Scientist*, *96*(5), 380. http://doi.org/10.1511/2008.74.1

Wiseman, H. (2014). SSRIs as Moral Enhancement Interventions: A Practical Dead End. *AJOB Neuroscience*, *5*(3), 21–30. http://doi.org/10.1080/21507740.2014.911214

Wiseman, H. (2014). SSRIs and Moral Enhancement: Looking Deeper. *AJOB Neuroscience*, *5*(4), W1–W7. http://doi.org/10.1080/21507740.2014.957116

Wiseman, H. (2014). Moral Enhancement—"Hard" and "Soft" Forms. *The American Journal of Bioethics*, *14*(4), 48–49. http://doi.org/10.1080/15265161.2014.889247

Wong, L., Huang, C. H., & Lee, B. W. (2016). Shellfish and house dust mite allergies: Is the link tropomyosin? *Allergy, Asthma and Immunology Research*, *8*(2), 101–106. http://doi.org/10.4168/aair.2016.8.2.101

World Bank Dataset. (2018). GDP per capita growth (annual %) for Japan, Germany, France and UK. Google Public Data Explorer. Retrieved from https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&ctype=l&strail=false&bcs=d&nselm=h&met_y=ny_gdp_pcap_kd_zg&scale_y=lin&ind_y=false&rdim=region&idim=country:GBR:DEU:FRA:JPN&ifdim=region&hl=en_US&dl=en_US&ind=false

World Health Organization. (2017). Plague. Retrieved from http://www.who.int/en/news-room/fact-sheets/detail/plague

Yamagishi, Toshio; Mifune, N. (2009). Social exchange and solidarity: in-group love or out-group hate? *Evolution and Human Behavior*, *30(4)*, 229–237.

Youssef, F. F., Bachew, R., Bissessar, S., Crockett, M. J., & Faber, N. S. (2018). Sex differences in the effects of acute stress on behavior in the ultimatum game. *Psychoneuroendocrinology*, *96*, 126–131. http://doi.org/10.1016/J.PSYNEUEN.2018.06.012

Zahn, R., de Oliveira-Souza, R., & Moll, J. (2011). The Neuroanatomical Basis of Moral Cognition and Emotion. In *From DNA to Social Cognition* (pp. 123–138). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9781118101803.ch8

Zhang, J. (2012). Genetic Redundancies and Their Evolutionary Maintenance. In O. S. Soyer (Ed.), *Evolutionary Systems Biology* (Vol. 751, pp. 279–300). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4614-3567-9_13