The fragility of moral traits to technological interventions[1]

Joao Fabiano[2]

Abstract: I will argue that deep moral enhancement is relatively prone to unexpected consequences. 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 insofar as deep moral enhancement targets higher-order desires (desires to desire something), it is prone to be self-reinforcing and irreversible. Fourthly, I argue that the complex causal history of moral traits, with its relatively high frequency of contingencies, indicates their fragility. Finally, I conclude 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 likely see a paradoxical decrease in co-operation between groups and a self-reinforcing increase in the disposition to engage in further modifications – both of which are potential problems.

Keywords: moral enhancement, fragility, moral traits

[1] This article has been accepted for publication in Neuroethics in 2020 following peer review, and the Version of Record can be accessed online at https://doi.org/10.1007/s12152-020-09452-6

[2] Department of Philosophy, University of Sao Paulo (jlafabiano@gmail.com). Research supported by grant #2019/22383-5 from the Sao Paulo Research Foundation (FAPESP).

1. Introduction

1.1 Deep moral enhancement

Advances in moral psychology and neuroscience indicate that human morality could soon be enhanced with the use of technology (Molly J Crockett, 2014). These technologies could improve moral reasoning, increase co-operation, refine our empathy and enhance other traits implied in human morality; when doing so, they would be hard to oppose on moral grounds (Douglas, 2008). Moreover, currently insurmountable global problems, such as nuclear proliferation, global warming and deadly pandemics, could be addressed once we become morally better people by enhancing co-operation. In fact, Ingmar Persson and Julian Savulescu claim these problems are so severe, risking our extinction, that we have a moral imperative to technologically enhance our co-operation for their prevention (Persson & Savulescu, 2008). Douglas (2008) coined and defined *moral enhancement* as any intervention that is expected to lead to morally better behaviours or motives.[3] Here I will be concerned with a specific kind of moral enhancement, that is, significant technological interventions directly targeted at human traits primarily expected to lead to morally better behaviour or motives, which I will call *deep moral enhancement* and expound shortly.

A simple argument for deep 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 could significantly strengthen those traits, then we should implement the intervention. Moreover, if these interventions could

[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 et al. (2014)). As a matter of usage, a search for the term on academic databases will reveal that there seems to be no publication on traditional moral education or moral progress using the term moral enhancement (Google Scholar, 2018). 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 will aim to produce a precise definition shortly.

help mitigate the risk of global catastrophes, we ought to implement them. However, I will argue that due to the fragility of moral traits, increases in currently morally desirable traits often are not themselves desirable and lead to unexpected effects.

Improving co-operation through traditional means of moral development such as education is mostly desirable and safe. Nonetheless, by using deep moral enhancement we could bring about changes in our levels of 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.[4] I will argue such deep changes could actually bring about moral decay. The limitations of traditional moral education create a safe boundary, which deep moral enhancement will be likely to breach. In this paper, I will argue that our moral traits – the target of deep moral enhancement – are fragile to technological intervention and that attempts to enhance them have a relatively high amount of risk.

As an example outside moral enhancement, 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, 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

[4] It is a tacit premise of the moral enhancement project that it will be more powerful than *most* forms of moral education in some way. It would be hard to propose a technology that might or might not be feasible, and might or might not be risky, just to achieve worse results than what we already have with moral education. It seems we have plenty of previous examples of pharmacological interventions that efficiently treat conditions that would be otherwise untreatable. I file this premise under the feasibility assumption to be mentioned shortly, which I will not directly address it. For a partial comparison between moral enhancement and moral education see Author (2020).

else (Portenoy et al., 1986). While pleasure might be 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.[5] The fragility of deeply enhancing morally significant things seems to indicate that partial enhancements can lead to not only partial but also undesirable consequences.

I will not defend the idea that every form of moral enhancement is undermined by fragility. 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, aggression or extraversion. 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 would not be a deep moral enhancement (Douglas, 2013). Although this would be a significant change leading to morally better behaviour, it would not significantly alter a human trait due to its narrow

[5] According to this view, defended by Alan Carter, 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 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 (Carter, 2011).

scope.[6] 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. 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. If a significant portion of the human population uses this drug, then the range of normal human variation would have been breached. 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 inhibit certain specific behaviours but it is unlikely to permanently change stable personality traits such as agreeableness or conscientiousness (Cobb-Clark & Schurer, 2012). 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 to be considered deep moral enhancement.[7]

The type of moral enhancement that Persson & Savulescu (2008) advocate to solve global problems 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. Douglas has discussed shallow and indirect forms of moral enhancement, but also mentions improving character traits directly (Douglas, 2014a). Although the distinction between interventions done by oneself or a third-party is relevant – especially when addressing risks to

[6] Assuming there would be no other significant consequences of this change. Such a drug could work by merely preventing his brain from processing someone's race during a trial while leaving his other traits and overall propensity to discriminate in other contexts unchanged. However, upcoming arguments about fragility will indicate that preventing cascading effects is not trivial.

[7] It might be the case that radical moral education or persistent environmental changes will cause significant changes in traits primarily expected to lead to morally better behaviour or motives, but for the sake of simplicity I will mainly focus my arguments on cases of technological innovations directly targeted at traits primarily expected to lead to morally better behaviour or motives. Also, per assumption, I presume that most current forms of moral education are not as effective as deep moral enhancement (see footnote 2).

freedom and assessing the overall desirability of specific interventions – I will not address it here. It is outside the scope of an investigation into fragility. Switching my upcoming examples from self to third-party interventions and vice-versa will have relevant consequences but not with regards to fragility. Lastly, the definition is based on expected results. This is to avoid defining deep moral enhancement in a way wherein it is almost tautologically safe because it is defined as producing actual instead of expected improvements.

Many opponents of moral enhancement doubt that it is even feasible (Agar, 2013b; Harris, 2013; Sparrow, 2014). I will not directly address this scepticism. In order to motivate research into moral enhancement's possible risks, it is sufficient to assume we will at least attempt to develop these technologies. Failed attempts to morally enhance can be dangerous, as even those who think of it as unfeasible point out (e.g. Agar 2013b). Such attempts will plausibly produce some effect, even if it falls short of moral improvement. Moral behaviour is mediated by neurochemistry, therefore changes in neurochemistry must produce some change of moral behaviour (Persson & Savulescu, 2014). There is nothing in the concept of moral enhancement that violates any known nature law, making it at least physically possible (Douglas, 2014b). Moreover, many studies indicate our morality is indeed subject to pharmacological manipulation. Increasing participants' serotonin made them more likely to have deontological judgements in moral dilemmas (Crockett et al. 2010) and more likely to co-operate in repeated prisoner's dilemma (Tse & Bond, 2002). Decreasing participants' serotonin made them more likely to overharvest a common resource pool (Bilderbeck et al., 2014), less likely to reject unfair plays in a game (Crockett et al. 2008), and less likely to co-operate in the repeated prisoner's dilemma (Wood et al., 2006). These findings are tentative, but promising. Although they might only be weak evidence that successful moral enhancement is feasible, they present stronger evidence that attempting it will alter moral traits. Development of technologies that attempt to produce moral improvement will likely happen even if we lack

scientific evidence that we can reliably do so (Shook, 2012). In fact, technological interventions that we already make wide use of, such as anti-depressants and ADHD medication, have a modest but measurable impact on our moral traits (Levy et al., 2014). Despite pessimism about empirical feasibility being a defensible position, it can lead to ignoring potential risks of attempting it or if the project turns out to be feasible (which is certainly an open question).

1.2 The fragility of moral traits thesis: background

The *fragility of moral traits thesis* simply means that deep moral enhancement could have major unexpected consequences, risking severe catastrophes. *Moral traits* are 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.[8] 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. Perhaps a seemingly simple moral enhancement would be to increase one's moral disposition towards raising the pleasure of others. Nonetheless, this might make one seek artificial means to induce pleasure, such as the aforementioned direct brain stimulation, disregarding other potentially important dimensions such as truthfulness and the variety of experiences. The disposition towards increasing others' pleasure is likely to be a worthwhile human trait; however, excessively and fundamentally enhancing this disposition will lead to results that are not only incomplete but also detrimental.

[8] 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.

2. The fragility of moral traits thesis

2.1 Introduction

The fragility thesis 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 further 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 thesis is: the actual outcome of performing a modification is further from the reasonably expected outcome when performed on moral traits than when performed on most physical or cognitive traits.[9] The particular setting of moral traits is such that 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. Fragility is also not a measure of uncertainty, as one can still have high uncertainty about consequences even when all possible outcomes deviate very little from the expected if the outcomes are many and their probability unknown. Moreover, the scope of what I mean by outcomes in my definition is very broad. Outcomes include changes in behaviours, motives, reasoning and society at large. It includes the consequences of the actions arising from a successfully modified trait, its associated changes in motives, and its

[9] I do not claim moral traits are the most fragile. There might be traits or other features more or equally fragile (examples range from consciousness to cancers, which all seem more or equally hard to improve upon).

societal impacts.[10] But it also includes outcomes from a failed modification that backfires because the intervention did not change the trait in the expected way (likewise, deep moral enhancement was defined in a way that includes failed attempts).

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 would be for a stuffed animal; thus, the former is more fragile than the latter. 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 and decreased fragility. Therefore, I am not claiming that this fragility is immutable, as these are unexpected but not permanently unexpectable (or unknowable) consequences. On the other hand, merely being aware of fragility will not, by itself, reduce it. Expecting the unexpected is not the same as expecting each specific outcome in the large set of unexpected consequences of deep moral enhancement. We need to know how things will go wrong to prevent it and not just that they will.

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 deep moral enhancement, individual co-operational enhancement, could go wrong because it would have big, 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

[10] Therefore, although I focus on consequences, I do not restrict myself to any specific form of consequentialism here. If anything, my analysis is pluralist regarding moral theory. If the modified individual behaves in the expected way but for the wrong motives, I count it as an unexpected outcome. Unexpected changes due to reduced deliberation also count by themselves.

unexpected results. Thirdly, I will contend that the intricate causal history of moral traits and its relatively high frequency of contingencies indicates they are fragile.

2.2 Fragility of human co-operation

2.2.1 Parochialism

Scientific research on understanding and increasing co-operation has largely focused on an individual level, though it is the group level that is problematic and in need of enhancement (Ostrom, 1990). Although one might 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. In order to reduce global risks, co-operation between countries is more crucial than co-operation between citizens, and the same is valid for ethnicities, political orientations and cultures.[11] Unfortunately, not only increases in moral dispositions conducive to co-operation *between individuals* are not guaranteed to promote co-operation *between groups*, but sometimes will cause increased competition.

Parochialism is the scientific term for the known tendency to prefer to co-operate with members of your group over out-groups, sometimes even if this comes at the expense of harming out-groups (Balliet et al., 2014). Groups that are highly co-operative internally will tend to be the least co-operative with other groups (de Dreu, 2014). Competition between groups leads to increased contribution to the public good within-group and to increased group effectiveness (Cardenas & Mantilla, 2015). In an overview of the literature on solidarity mechanisms, the psychologist Gary (Bornstein, 2003) 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 […]. Whereas the

[11] Some might argue that aggregating individual co-operation should produce group co-operation. However, as I will explain in the next paragraphs, individual co-operation is often restricted to one's group and sometimes to the detriment of outsides. Aggregating it will not remove these effects.

> 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"

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 coevolved with aggression towards out-groups. For example, Bowles & Gintis (2013) attempted to model the evolution of co-operation using our best estimates regarding group-size and food-sharing during the Palaeolithic. Their results show that parochialism and co-operation could only have evolved together.[12] It would be evolutionary disadvantageous to express co-operation without parochialism, therefore the activation of brain networks responsible for only co-operation seems unlikely. 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, 2016), but it is also known to produce in-group favouritism, leading to ethnocentrism and parochialism (De Dreu, 2012).

2.2.2. Problems with private altruism

It is plausible that the case of human co-operation is just one example of several paradoxical emergent effects 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, parochialism is so intrinsically connected with co-

[12] 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.

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; non-parochial co-operators are rare (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 (Author, 2020).

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

According to some views, certain dispositions consist of higher-order desires – the desire to desire *x* – producing motivations that generate behaviour to pursue *x*. Under ideal conditions, *x* will correspond to our moral values (Smith et al., 1989); hence these dispositions are moral traits. Therefore, enhancements targeting higher-order dispositions will be moral enhancements according to my definition. Increasing co-operation, as proposed by Persson & Savulescu, targets a higher-order desire. Co-operativeness as a trait is a social value orientation, it is a preference for a certain distribution of benefits between self and others affecting a wide range of situations and behaviours, it is not a desire for any specific outcome (Van Lange et al., 2013).[13] Perhaps some enhancements targeting first-order desires will also be moral enhancements, but those will not be immediately affected by my arguments here. I contend that

[13] Moreover, if the trait of being co-operative were a first-order desire, it would have to correspond to the entire set of desires for each co-operative outcome, including outcomes that are yet to happen. It is more plausible that co-operativeness is a desire to desire co-operative outcomes, which gives birth to specific first-order desires in different situations.

insofar as moral enhancement targets higher-order desires, 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 thesis – assuming some small mistakes will be made.

2.3.1 Self-reinforcing

Steve considers all forms of violence wrong.[14] 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 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 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

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

him to consider life on Earth as not worthwhile. But morally enhanced Steve, eventually, would.[15]

Deep moral enhancement is likely to produce *self-reinforcing chains of modifications*. Increasing one's moral dispositions to desire to desire X will increase the perceived value of increasing one's moral dispositions to desire to desire X even more 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 article), then the result of one small enhancement interaction will probably 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 higher-order desires, 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 likely be more prone to want to develop and take new utilitarian enhancing drugs, which in turn would make them desire those values and have those goals to an ever-greater extent. Eventually, many 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. The enhancement operation will likely 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

[15] 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 stop with just one single modification.

embark on the first enhancement iteration 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. 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 argues that slippery slope arguments can be self-defeating when they 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 (Douglas, 2010). 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 upon 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. It might well be that for future persons, extreme utilitarianism is morally desirable, but the fact that 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.

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.

2.3.2 Irreversibility

Iterative deep moral enhancement was bad for Steve. Let us say we fixed that problem by simply committing to not continuously enhancing. Steve morally enhanced to become less violent. But moral traits are interconnected. 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 higher-order desires towards X will decrease the perceived value of decreasing one's higher-order desires towards X – that is, of reversing the increase.[16] Presumably, prior to the enhancement, the perceived value of decreasing one's higher-order desires 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 the higher-order desires of a motivational structure in a certain direction, this plausibly creates a chain of motivations that would function to maintain those higher-order desires, 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 towards value 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, then it becomes more probable that we will not want to change our goals and reverse this change.

[16] To clarify, by irreversible I do not mean it would be technically unfeasible to revert, but merely one would be unwilling to revert. One could still be coerced into reverting.

2.3.3 Additional remarks

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 effects occur in deep moral enhancement. Any possible mistake from performing moral enhancement is either amplified by self-reinforcement or made unfixable by irreversibility.

Deep moral enhancement might be performed without directly changing higher-order desires; in that case, the two effects explored here would be less likely. On the other hand, shallow forms of moral enhancement might modify higher-order desires. Nevertheless, insofar as higher-order desires are moral traits (as suggested by dispositional theories of value such as Smith et al. (1989)), (1) deep moral enhancement is likely to change these higher-order desires, and (2) changes primarily expected to alter these higher-order desires will fall under the definition of deep moral enhancement. Finally, even if higher-order desires were not seen as an important aspect of human morality but deep moral enhancement affected them, 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 higher-order desires in morality.

In my example, one could argue that Steve could decrease only his desire to be aggressive, not his desire to desire being aggressive. But aggressiveness, just like co-operativeness, is a trait, thus it is a higher-order desire. Traits are defined as a general and stable behavioural pattern. A consistent behavioural pattern across distinct situations without a higher-order desire motivating such a pattern would be an odd coincidence; more so if this pattern is morally laden. Likewise, a technological intervention precisely targeting Steve's specific desires to take aggressive actions without affecting his higher-order desire towards aggression seems implausible. Even if possible, such intervention would not fall into my

definition of deep moral enhancement as it would not be targeting any human trait but only the set of behaviours associated with such trait.[17]

Some might concede that deep moral enhancement will affect higher-order desires but argue that even higher-order desires would be unaffected, and these desires would prevent self-reinforcing effects. For instance, one might enhance the second-order desire to be co-operative but not its associated third-order desire. This third-order desire would keep further enhancements in check. I defined higher-order desires as the desire to desire *x*. But could *x* be itself a desire? The short answer is likely no. Dispositional theories of value stop at second-order desires. Second-order desires are enough to account for what we value in such theories. It is hard to even conceive what it means to desire to desire to desire something.[18]

2.4 Aetiological complexity increases fragility

Our moral traits are the product of many contingent and accidental events with random processes involved. If we assume our moral traits are (at least partially) shaped by natural history and human history, then it is easy to see how those two histories were populated with contingencies that happened for no good reason and that have shaped our moral traits. Firstly, one of the major processes influencing natural history is natural selection. Natural selection 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

[17] To see how these differ consider a perfectly secure prison for serial killers. Few would argue its inmates are no longer aggressive.

[18] It is even harder to conceive desiring to desire something but desiring not to desire to desire it. One may have conflicting second-order desires, but how can one desire to desire orange juice, but desire not to desire to desire orange juice? If they exist, third-order desires would likely only mirror second-order desires. Of course, properly settling this question lies outside of the scope of this article, but it is enough that these desires are unlikely to play a role here.

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.[19] Another example, one of the explanations for our apparent 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).

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 a particular hominid population was hit by a hurricane will dictate whether a particular brain structure, with a particular way of reasoning about morality, will continue to exist. It should be noted that these kinds of extinction

[19] 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.

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 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.

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. 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. For instance, the ability to 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.

3. Objections and responses

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 fragile. 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.

It could be argued that our current understanding of moral traits is limited, thus we are bound to see the matter as extremely complex. This was true of several other fields before we found a very simple law or principle.[20] However, I have given arguments that indicate that

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

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 the simplest possible theory is produced, such a theory will still be very complex. Secondly, 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 perform deep moral enhancement after a full account of moral traits is provided. Breaking something requires less knowledge than fixing it, therefore we will likely have the power to significantly alter our moral traits before we have enough knowledge to determine how to safely improve them.

Some might believe that a complex causal history would have no influence on the final complexity (and thus fragility) of moral traits. However, this fails to account for the fact that a causal history is constructed as the simplest explanation of an event (Lewis, 1986). If an event is simple enough that it could have been brought about by a very simple process, then its causal history will also be simple. The fact we need to evoke complex processes and contingencies to explain moral traits indicates they have high complexity.[21] Every time a contingent effect occurs, 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 in order to explain moral traits is higher than those evoked to explain vertebrate body template.

Finally, a significant line of counter-argument against my fragility thesis can be found in Allen Buchanan's chapter "Conservatism and Enhancement" in his 2011 book *Beyond*

[21] 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.

Humanity (Buchanan, 2011). Buchanan argues against the idea that the human organism is akin to a house of cards wherein only one small apparent improvement could bring the entire system down. According to him, the evolutionary processes that resulted in our current traits tend to produce very robust end products. This robustness means they are resistant to changes that would catastrophically alter them, greatly decreasing the range of unexpected consequences and thus making them less fragile according to my definition. 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 undergo 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. 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.[22] 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 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 et al., 2015). 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

[22] 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.

modularity argument does not hold for moral enhancement, although it might hold for most other targets of human enhancement.

4. Conclusion: fragility leads to increased risks

Any substantial technological modification of moral traits 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, 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.

Deep moral enhancement is likely to be hard to develop safely, but not necessarily be impossible or 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 traits are well suited to dealing with global problems. On the contrary, there are certainly reasons to expect that there are better traits that could be brought about by enhancement technologies. However, I believe my arguments indicate there are also much worse, more socially disruptive, traits accessible through technological intervention.

References

Aaldering, H. (2014). *Parochial and Universal Cooperation in Intergroup Conflicts*. Universiteit van Amsterdam.

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

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

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

Author. (2020). _____. *Journal of Medical Ethics*.

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

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

Bilderbeck, A. C., Brown, G. D. A., Read, J., Woolrich, M., Cowen, P. J., Behrens, T. E. J., & Rogers, R. D. (2014). Serotonin and Social Norms. *Psychological Science*, *25*(7), 1303–1313. https://doi.org/10.1177/0956797614527830

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

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.

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

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

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

Cobb-Clark, D. A., & Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, *115*(1), 11–15. https://doi.org/10.1016/j.econlet.2011.11.015

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin Modulates Behavioral Reactions to Unfairness. *Science*, *320*(5884), 1739–1739. https://doi.org/10.1126/science.1155577

Crockett, Molly J. (2014). Moral bioenhancement: A neuroscientific perspective. *Journal of Medical Ethics*, *40*(6), 370–371. https://doi.org/10.1136/medethics-2012-101096

Crockett, Molly 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. https://doi.org/10.1073/pnas.1009396107

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. https://doi.org/10.1016/j.yhbeh.2011.12.009

de Dreu, C. K. W. (Ed.). (2014). *Social conflict within and between groups.* Psychology Press.

De Dreu, C. K. W., & Kret, M. E. (2016). Oxytocin Conditions Intergroup Relations Through Upregulated In-Group Empathy, Cooperation, Conformity, and Defense. *Biological Psychiatry*, *79*(3), 165–173. https://doi.org/10.1016/j.biopsych.2015.03.020

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. https://doi.org/10.1111/cobi.12380

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

Douglas, T. (2010). Intertemporal Disagreement and Empirical Slippery Slope Arguments. *Utilitas*, *22*(2), 184–197. https://doi.org/10.1017/S0953820810000087

Douglas, T. (2013). Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics*, *27*(3), 160–168. https://doi.org/10.1111/j.1467-8519.2011.01919.x

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

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

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

Google Scholar. (2018). *Moral Enhancement*. https://scholar.google.com/scholar?q=%22moral enhancement%22

Harris, J. (2013). 'Ethics is for bad guys!' Putting the 'moral' into moral enhancement. *Bioethics*, *27*(3), 169–173. https://doi.org/10.1111/j.1467-8519.2011.01946.x

Krebs, D. (2015). The Evolution of Morality. In D. Buss (Ed.), *The Handbook of Evolutionary Psychology* (pp. 747–771). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470939376.ch26

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. https://doi.org/10.1353/ppp.2014.0023

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

Ostrom, E. (1990). *Governing The Commons The Evolution of Institutions for Collective Action Cooperation Commons*. Cambridge University Press.

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. https://doi.org/10.1111/j.1468-5930.2008.00410.x

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

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. https://doi.org/10.1016/0304-3959(86)90155-7

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

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

Smith, M., Lewis, D., & Johnston, M. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *63*, 89–174.

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. https://doi.org/10.1080/00048402.2013.860180

Tse, W., & Bond, A. (2002). Serotonergic intervention affects both social dominance and affiliative behaviour. *Psychopharmacology*, *161*(3), 324–330. https://doi.org/10.1007/s00213-002-1049-7

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. https://doi.org/10.1016/j.obhdp.2012.11.003

Wood, R. M., Rilling, J. K., Sanfey, A. G., Bhagwagar, Z., & Rogers, R. D. (2006). Effects of Tryptophan Depletion on the Performance of an Iterated Prisoner's Dilemma Game in Healthy Adults. *Neuropsychopharmacology*, *31*(5), 1075–1084. https://doi.org/10.1038/sj.npp.1300932

Zhang, J. (2012). Genetic Redundancies and Their Evolutionary Maintenance. In O. S. Soyer (Ed.), *Evolutionary Systems Biology* (Vol. 751, pp. 279–300). Springer New York. https://doi.org/10.1007/978-1-4614-3567-9_13