# Moral Enhancement and Moral Disagreement

G. Owen Schaefer

**Thesis submitted for the degree of D.Phil. in Philosophy**

Lincoln College
Faculty of Philosophy

Trinity Term 2014

***For Poh Lin***

# Abstract

At first glance, the project of moral enhancement (making people more moral) may appear uncontroversial and obviously worth supporting; surely it is a good idea to make people better. However, as the recent literature on moral enhancement demonstrates, the situation is not so simple – there is significant disagreement over the content of moral norms as well as appropriate means by which to manipulate them. This disagreement seriously threatens many proposals to improve society via moral enhancement. In my dissertation, I develop an understanding of how, exactly, disagreement poses problems for moral enhancement. However, I also argue that there is a way forward. It is possible to bring about moral improvement without commitment to particular and controversial moral norms, but instead relying on relatively uncontroversial ideas concerning morally reliable processes. The upshot is that, while attempting to directly manipulate people's moral ideas is objectionable, it is relatively unproblematic to focus on helping people reason better and avoid akrasia, with the justified expectation that this will generally lead to moral improvement.

We should, therefore, focus not on how to bring people in line with what we take to be the right ideas, motives or behaviors. Rather, we should look to helping people determine for themselves what being moral consists in, as well as help ensure that they act on those judgments. Traditional, non-moral education, it turns out, is actually one of the best moral enhancers we have. In fact, the tools of philosophy (which is, in many aspects, concerned with proper reasoning) are central to the project of indirect moral enhancement. Ultimately, one of the best ways to make people morally better may well be to make them better philosophers.

# Table of Contents

# Acknowledgments

This dissertation is the culmination of my graduate career at Oxford, and it would most certainly not have been possible without the help of a number of people, groups and organizations who must be acknowledged. First and foremost, I would like to thank my supervisor Roger Crisp for his careful and invaluable oversight of this dissertation, as well as the B.Phil. thesis that preceded it. A work is only as good as its corrections, and the numerous revisions prompted by Roger's incisive comments undoubtedly led to a stronger and more compelling work. I very much appreciate Roger's unflagging support for my research and his patience as I progressed through the degree.

I would also like to thank those individuals who have seen and provided comments on various excerpts from my dissertation. These include my D.Phil. examiners Guy Kahane and Ingmar Persson, as well as Tom Douglas (whose own dissertation and publications were ground-breaking in the area of my research, moral enhancement),Julian Savulescu, Jordan Ridgewell, Wenwen Fan, my anonymous B.Phil. examiners, members of the Oxford Applied Ethics Discussion Group, attendees at the 2012 Central States Philosophical Association Conference, attendees at the Seventh International Conference on Applied ethics, attendees at the 2012 Dutch Conference on Practical Philosophy, and attendees at the 2013 International Conference on Enhancement: Cognitive, Mood and Moral.

In addition, there are number of institutions that have provided various forms of support for this research. I thank the Clarendon Fund for providing generous financial support for my research; the Uehiro Centre for Practical Ethics for the office space, collegial atmosphere and casual work opportunities that helped make ends meet; the Faculty of Philosophy for accepting me in the first place and continuing to be an incredibly rich source of education and enlightenment; Duke University for being so welcoming during my time spent abroad; Lincoln College for being a wonderful source of well-being and supporting various conference activities during my D.Phil.; and St. Cross College for being a welcoming and open institution when I first came to Oxford on the B.Phil.

Finally, I would be remiss if I didn't express my appreciation for my friends and family, who provided social, spiritual and extracurricular support during the arduous process of completing my dissertation. The most important thing, I have found, is to stay sane, have a level head and keep everything in perspective.

# Part I: Preliminaries

The main purpose of this work is to provide arguments against one (direct) form of moral enhancement, due to concerns over disagreement, and in favor of another (indirect) form that avoids those problems. But before I can delve into those arguments, the nature and scope of moral enhancement must be clarified. What exactly is meant by 'moral enhancement' is discussed in Chapter 1. In particular, direct moral enhancement is defined as an intervention designed to bring about what the enhancer takes to be the morally correct ideas, motives or behaviors; indirect moral enhancement, by contrast, is designed to bring about moral improvement without commitment to what the morally correct ideas, motives or behaviors are. I later go into more detail concerning the particular forms of traditional (Chapter 2) and biological (Chapter 3) moral enhancement. This helps explain how the debate has impacted different areas to date and foreshadows several concerns with different forms of moral enhancement that will be addressed later. I will also lay the groundwork for later discussion of the implications of my position in various domains.

# Chapter 1: Introduction

Throughout history, determining the nature of morality and promoting philosophical ideals of moral behavior have been central concerns in human societies. These concerns have involved not only trying to determine what is good and right but also trying to determine how to ensure people will in fact be good and do the right thing. While the former has received a great amount of philosophical attention, the latter has – until recently – been somewhat overlooked. The attention that has been given to the question of how to inculcate values has focused on traditional methods such as education, (dis)incentives and social pressure. In addition, recent scientific developments open up the prospect of influencing individuals' moral dispositions and behavior through biological interventions, particularly in the form of chemical, neurological or, more speculatively, genetic manipulation.

There is a strong prima facie case for permitting, developing and promoting moral enhancements. Moral failings are, almost by definition, problematic and indisputably worth overcoming. At the individual level, we try and convince ourselves or others into being moral – keeping promises, respecting others' rights, acting kindly, and so on. And at a societal level, we endorse various policies aimed at promoting moral behavior (e.g., reducing crimes like theft or assault, preventing environmental damage, and subsidizing altruistic behavior through tax credits). Moral enhancement could be characterized as simply another intervention of this sort, one that could ultimately be more effective than previous efforts to induce morality and so is especially worth promoting. Nevertheless, unique difficulties exist for moral enhancement.

In what follows, I will investigate a particular set of difficulties that emerge from the existence as well as importance of moral disagreement, and how to overcome them. The

upshot will be that a) we cannot have confidence that at least certain forms of moral manipulations[1] are truly enhancements and b) large-scale programs of moral enhancement would be wrong-headed. However, I will argue that there is a path forward for moral enhancement. Moral enhancement can be best achieved by focusing on the reasoning processes that will reliably lead to moral improvement.

Before proceeding, it is important to specify exactly what moral enhancement amounts to. Tom Douglas defines moral enhancement as bringing about (sets or groupings of) morally better motives – motives being "the psychological — mental or neural — states or processes that will, given the absence of opposing motives, cause a person to act." (Douglas 2008, p. 229; see also Walker 2009, pp. 29-30, who defines moral enhancement in terms of better traits or stable dispositions) This definition, of course, leaves it open as to what makes a given motive more moral; it could be the mental content itself (for instance, the intention to help someone or desire that they be harmed), the goodness of actions that (tend to) result from the motive (for instance, the tendency to actually help or harm others), or some combination of the two.

It is important that this understanding of moral enhancement is substantively neutral. Persson and Savulescu (2008, 2010 and 2012) offer a decidedly non-neutral understanding, according to which "moral enhancement will consist in strengthening our altruism and making us just or fair, i.e. properly grateful, angry, forgiving, etc." (Persson and Savulescu 2008, p. 169); these traits are selected in part because they are useful in the avoidance of the harms caused by greater destructive power that results from cognitive developments. While this specification may be helpful in clarifying which interventions, in particular, will count as

[1] The term 'moral manipulation' refers to any alteration of someone's moral beliefs or alteration of motives or actions that have a clear moral dimension. Moral manipulations will then include manipulations that denigrate as well as improve people's moral beliefs, motives and actions. This is a term of art – it is meant to include interventions such as persuasion that are not typically thought of as manipulations, but nevertheless can change a person's moral thought and behavior.

moral enhancements, it does so at the cost of not being a suitably general definition; it will leave out features important to moral systems that identify the non-instrumental value of such traits, finding value in sources other than harm and benefit. Indeed, that very disagreement about such normative issues will motivate much of the discussion to come. We should employ, then, a substantively neutral definition like Douglas's.

Still, Douglas's definition is somewhat too narrow. There are at least two classes of interventions that, on at least some substantive views, might count as moral enhancement that do not fit well with a motivational conception. The first class involves mental states (especially moral beliefs) that, for whatever reason, do not necessarily cause a given behavior in the absence of countervailing motives. For instance, someone might be in the grips of a severe depression such that very few of her beliefs motivate her to action; the person would interact with others in essentially the same way whether or not, for instance, she held extremely racist beliefs. It is plausible, on at least some moral theories, that such beliefs are (in a certain sense) morally wrong, and it would be morally better for the depressive not to hold them. This points to a general issue: according to some views, some mental states (such as moral beliefs) may be morally relevant whether or not they lead to action. We should refine our understanding of moral enhancement to accommodate such views.

The second class of interventions are those that aim strictly at altering behavior, irrespective of whether they were caused by mental or neural states. This may seem like an irrelevant category; after all, pretty much all actions (excluding basic reflexes) are caused, in one way or another, by mental or neural states. That may be so – but it is important that the real objects of such interventions are actions, not mental states. This is an important distinction; it is quite plausible for a utilitarian to claim as a moral enhancement something that (say) makes a wealthy individual donate vast sums of their money to efficient welfare-promoting charity, irrespective of the mental or neural states involved. Technically speaking,

the intervention would be (most likely) be mediated by mental or neural states, but it would be at least somewhat misleading to emphasize those states, as Douglas's definition does. Such consequentialist accounts would be ultimately interested in moral enhancement via improving people's actions (and the effects of their actions), not their mental or neural states (indeed, the success of a given intervention would be measured primarily in terms of the good effects of the actions, not the states that caused them, just as the success of a cancer drug is measured in terms of its effect on a patient's health, not the pathway by which it achieved that effect).

There may be other morally relevant features besides these; it is important for the definition of moral enhancement not to presuppose any such substantive moral claims, and so be open to any potentially relevant features. To accommodate these various views, I propose defining moral enhancement as the moral improvement in someone's features. This is admittedly somewhat vague, so for the purpose of this thesis I will focus on what are the three sorts of features most plausibly relevant to moral improvement: mental states (especially beliefs[2]), motives and actions. This understanding of moral enhancement illustrates the plurality of ways that a given intervention might count as a moral enhancement, depending on one's background moral commitments. It is certainly broader than many of the previous understandings of moral enhancement given in the literature (as noted later, it will potentially encompass imprisonment), but it can still be used to adequately engage with others writing on moral enhancement insofar as narrower definitions still fall within its scope.

When it comes to determining what changes would count as moral enhancement, I will for the most part remain neutral between various substantive moral positions in normative ethics concerning what the good and the right consist in. Indeed, to do otherwise

[2] For the remainder of the paper, I will for the most part use the term 'moral beliefs' to refer to the relevant mental states. One might hold that some other sort of mental state (such as desires or judgments) is of central moral importance. I wish to remain neutral on this dispute, and use the term 'belief' primarily for simplicity. The discussion below should apply to whatever relevant mental state one puts in the place of 'belief'.

would be to beg the question against the arguments I will here develop. The key difficulty for moral enhancement is disagreement about morality; without such agreement, any answer for what makes a given property more moral may be cast into doubt. This disagreement is particularly problematic when we seek to morally enhance other people. We may have confidence in our own moral ideas, but as I will argue, we have reason to avoid pushing those ideas on others via direct biological moral enhancement.[3]

Two further distinctions are necessary. The first is between biological and non-biological forms of moral enhancement. Biological interventions are just those that essentially involve alteration of an individual's biology. This will include the ingestion or injection of certain chemicals as well as neurosurgery and genetic manipulation. Non-biological forms of enhancement involve actions like persuasion, inducement or even brainwashing using advanced (non-invasive) psychological techniques. Traditional attempts at moral enhancement have been of the non-biological variety. Biological moral enhancements, however, are relatively new and potentially more efficacious. The arguments in this dissertation will apply to both biological and traditional non-biological moral enhancements will be addressed in this dissertation, but some differences will remain. In particular, biological moral enhancements have much greater potential at effecting significant changes in people's moral behavior and attitudes – though whether such would be a moral improvement remains to be seen.

[3] Later on in Chapter 6, I will critique some forms of moral enhancement on the grounds that they threaten the value of disagreement. This value is arguably substantive, and so might appear to violate the neutrality laid out here. However, importantly, this neutrality is not completely thoroughgoing; it is, rather, limited to the definition of moral enhancement. Essentially, we do not want a definition of moral enhancement to presuppose a particular normative framework. However, it might still be possible to allow one's substantive values to influence a definition that still left the content of the good and the right open. This might seem to imply that an enhancement inculcating the value of disagreement in the population should be endorsed; however, that sort of enhancement would essentially be self-defeating due to its necessary conflict with the existence of dissent, which I will argue is something worth preserving.

The second distinction is between direct and indirect moral enhancements. A given intervention is a direct moral enhancement when it is designed to bring someone's beliefs, motives and/or actions in line with what the enhancer believes[4] are the correct moral beliefs, motives and/or actions. So, for instance, if an enhancer believes that it is wrong to kill an innocent, then he would be performing a direct moral enhancement by inculcating the belief that murder is wrong, or by inculcating the motive or inclination to avoid murdering. An indirect moral enhancement, on the other hand, is designed to making people more reliably produce the morally correct ideas, motives and/or actions without necessarily committing to the content of those ideas, motives and/or actions. And though indirect moral enhancements do not rely on particular substantive commitments, they will rely on the connections between certain processes and the correctness of moral beliefs, motives and actions.[5]

This dissertation will be organized as follows. This first part consists of an introduction and background to various forms of moral enhancement. Chapter 2 surveys the traditional, non-biological forms of moral enhancement and some of the issues that surround them. Chapter 3 will discuss some of the currently-available means of biological moral manipulation through biochemical interventions. This will not only demonstrate that biological moral manipulations are possible, but also indicate some of the difficulty with classifying any given manipulation as a moral enhancement.

---

[4] I will not assume that the enhancer's beliefs are correct, as this would be question-begging against the fallibility objection below. Indeed, some have argued that moral enhancement just consists in enhancing what one believes to be moral (see, e.g., Shook 2012). Still, even if one adopts an objectivist account according to which moral enhancement only occurs when the enhancer is correct, the further objections concerning reasoning and individuality will still apply.

[5] It might seem more intuitive to classify these two approaches as substantive vs. procedural. However, this distinction would be somewhat misleading. Indirect moral enhancement may involve some substantive commitments – not to the morality of particular ideas, motives or actions, but the connection between certain processes and moral outcomes. Moreover, indirect moral enhancement need not be purely procedural; it does not involve commitment to the idea that certain processes necessarily or inevitably lead to better moral ideas, motives or actions, nor that morality consists simply in following particular sorts of rational processes (c.f. Smith 1994). Instead, the position merely makes a probabilistic claim – that certain processes will tend towards good outcomes, though they may not give the best results in every case.

Part II will discuss several problems for direct moral enhancement that arise from moral disagreement. Chapter 4 outlines various levels of moral disagreement and the implications different theoretical accounts might have for a program of moral enhancement. These differing implications point to how unsettled a program of moral enhancement will be, and the difficulty of reaching a consensus on a wide range of topics. Chapter 5 discusses a deeper epistemic problem for moral enhancement, linking the present topic to general issues of disagreement among epistemic peers. I show that we should be skeptical of any claims of enhancement of people's moral beliefs and motives (though not necessarily actions) without becoming moral skeptics in general. Chapter 6 discusses a different problem for moral enhancement, in particular wide-scale efforts of enhancement, arising from the value of disagreement. Expanding on Mill's work in *On Liberty*, I argue that such a program would problematically lock us in to potentially-fallacious current moral beliefs, motives and behaviors, as well as suppress valuable features of society such as reasoned deliberation and individuality.

Part III shows how these difficulties can be overcome. Chapter 7 outlines a necessary shift from direct moral enhancement to indirect moral enhancement, which allows one to stay mostly neutral on substantive issues and avoid the issues arising from disagreement. It will further be argued that indirect moral enhancement focused on reason and rationality, rather than sentiment and emotion, is the most defensible. An important step in this is establishing that moral intuitions (broadly construed) are generally reliable, which I defend in Chapter 8. Chapter 9 argues for the rationalist approach to moral enhancement, relying on a broad notion of reasoning and defending the connection between reasoning and correct moral beliefs, motives and actions. Chapter 10 discusses the issue of akrasia – does removing weakness of will in moral cases constitute an indirect moral enhancement? I answer in the affirmative, but this case is much more contingent than the general reasoning ability discussed in Chapter 7.

The final Part IV discusses the upshots of the preceding arguments. Chapter 11 focuses on the implications of this view. Just as many have argued for liberal neutrality in public education, I argue for substantive neutrality in all forms of education (public and private) as well as any biological interventions aimed at moral enhancement. Chapter 12 will briefly summarize the main points of this dissertation and emphasize how this dissertation is more or less a moderate approach to moral enhancement, one that should appeal to proponents and opponents alike.

Indeed, my purpose here is not to decry any attempts at moral improvement. Instead, I want to point towards potential pitfalls with rushing headlong into moral enhancement program. In this way, my work bears some similarity to the work of John Harris (2011), who also critiques various substantive approaches to moral enhancement while expressing sympathy for a more indirect approach aimed at reasoning. However, as will become clear in Chapter 3, I disagree with the particular arguments deployed by Harris; moreover, my own account spends considerable time developing and defending an indirect approach to moral enhancement. Going forward, a relatively nuanced approach should be used when pursuing moral enhancement. This will involve sensitivity to the widespread substantive disagreements that pervade our society and suitable humility about our own deep-seated moral views. With such a sensitive approach, the difficulties engendered by moral disagreement may indeed be avoidable and allow for the responsible pursuit of moral enhancement.

# Chapter 2: Non-Biological Moral Enhancement

Until recently, moral enhancement at both the individual and societal levels has focused on non-biological means. Moral education is arguably the most prominent traditional form of moral enhancement, but other means include argumentation and various forms of inducement such as propaganda, social pressure, and motive manipulation. This section will survey these areas and note some of the difficulties that immediately arise. These difficulties foreshadow larger problems with moral enhancement as well as offer some suggestions as to how these problems could be overcome. Later chapters of this dissertation will endorse some of these solutions, though with markedly distinct arguments.

## Moral Education

The term 'moral education' is somewhat vague, and will overlap to some degree with the other traditional approaches discussed in this chapter. What distinguishes moral education, though, is the idea of teaching. This presupposes a teacher-student relationship, with one individual in a pedagogically superior relationship. On some accounts, this will mean the teacher is in some way more morally knowledgeable or expert. On other accounts, this will mean the teacher is better-equipped to guide the student towards the morally correct beliefs, motives and actions – even if the teacher him- or herself is not a moral expert.

### *The virtue approach*

The importance of moral education is nearly as ancient as philosophy itself. One of the most prominent in the early discussions of moral education is the work of Aristotle, who incorporated moral education into his overall theory of virtue ethics.[6]

On Aristotelian accounts, the exercise of virtue is the central good – not only morally, but also for people's own happiness or flourishing. This virtue does not consist (just) in doing the right thing, or being disposed to do the right thing; one must act through practical wisdom, which is the knowledge that allows someone to recognize and respond to the morally salient features of a situation and discern from there the proper course of action. This does not mean the virtuous person will always be contemplating every aspect of a situation, but he or she must be appropriately sensitized to those features, and indeed natural reactions have some rational underpinning. The truly virtuous individual will not, on reflection, be ignorant of what made some particular decision the correct one. Particular virtues such as courage, honesty or beneficence will just correspond to particular sorts of morally salient features (danger, truth, need, etc.); to be courageous is to (reliably) recognize and act on the appropriate degree of boldness in the face of danger.

Given the centrality of virtue not only to morality but flourishing in general, it was crucial for Aristotle to establish how to bring about virtue both in oneself and others. Aristotle argues for two connected means of inculcating virtue: habituation and role-modeling. Habituation involves acting virtuously even in the absence of the practical wisdom discussed above. The idea is that, just as one can gain practical knowledge of trades by practicing a trade, one can gain virtue by practicing virtuous activity. By way of example, Aristotle writes: "For abstaining from pleasures makes us become temperate, and once we have become temperate we are most capable of abstaining from pleasures." (*Nichomachean*

[6] Interestingly, Confucius – writing halfway around the world and two centuries before Aristotle – defends a similar approach to moral education as Aristotle. (Huang 2011)

*Ethics*, 1104a34-b2)[7] This does not mean habituation is sufficient, of course, but it is in the very least necessary. And such habituation could occur at the individual level, with parents compelling their children to act rightly, or at the social level, with the enactment of laws enforcing virtuous behavior in the hopes that, through enough practice, virtue will follow.

But how is the non-virtuous person supposed to know how to act virtuously[8]? Here, role-modeling is crucial. Role-modeling relies more explicitly on a teacher-student relationship than simple habituation, but is crucial for the latter to reliably induce virtue. The role model must be, first and foremost, a virtuous person. The student of virtue will then take cues from this virtuous exemplar – practice acting as they do, which is known to be virtuous. But more than simply having a role model is needed – there must be a connection between the role-model and the student of virtue. As Nancy Sherman observes, "We learn best from those with whom we can identify and from those whom we value positively." (Sherman 1999, p. 41) This connection both fosters a motivation to respect and emulate the role-model, as well as plays up the crucial emotional role in virtue – a positive emotional attitude towards virtue. Without the motivation to be virtuous, role-modelling and habituation would be pointless.

While this view is rather attractive, it faces at least three difficulties when used as a model of moral education. Firstly, virtuous action requires, crucially, acting not only rightly but for the right reason. But, as Robert George (1993) has pointed out, the Aristotelian account does relatively little to encourage this. Habituation on its own will just lead individuals to act out of habit or custom – not because something is virtuous. One might hope that the emotional connection noted above will supply the right sort of reason, but that

[7] Here and elsewhere in this work, I rely on Terence Irwin's (1999) translation.
[8] This discussion talks about 'acting virtuously' generally for the sake of simplicity. Of course, Aristotle's strategy is more particular – practicing a specific virtue inculcates that specific virtue, though there may well be cross-over between virtues. The arguments in this section, though, are roughly untouched whether one thinks about specific virtues or virtue in general.

will not work. The emotional connection would lead someone to act in a certain way because of that connection, not because of the virtuous nature of the act. This is once more the wrong sort of reason.

Secondly, virtue is not like technical skill, in part because there is no direct feedback. One could become a better driver, for instance, through practice because cases of failure are obvious, and the reasons for failure often relatively transparent. Virtuous action, however, does not have this quality. There is no direct 'feedback' for failing to act virtuously. One could look to imitation of role models, but that will be entirely external. One could only learn to act as a virtuous person acts via imitation. One cannot learn to think like a virtuous person simply through observation, because others' thoughts cannot be observed. This phenomenal gap poses great difficulties in general when trying to engage with others on moral matters, as will be discussed further in Chapter 5.

Thirdly, virtue education faces a familiar epistemic difficulty. Imitation of the virtuous is meant to guide habituation, but how does one know whether some role model is truly virtuous? The easiest strategy would be for the teacher to themselves be virtuous and, through adequate self-reflection, realize this fact. The teacher can then justify to him- or herself why others should be imitating them. However, that justification is not readily available to those who are not already virtuous – either the students themselves, or well-meaning but flawed third parties setting up a system of moral education for society. This problem becomes particularly acute when various people claim to be virtuous, but act in markedly different manners. Whom should those seeking virtue follow? The only clear path would be for the non-virtuous to gain the practical wisdom to recognize virtuous activity when they see it. However, to do so is to already have the tools at one's disposal to become virtuous. In that case, virtue education, habituation and role-modeling would have been

completely superfluous.[9] Virtue can be achieved without any of those strategies. Some small space might remain for the motive to act virtuously to be inculcated in someone who can already recognize virtuous activity. However, this would be to severely limit the scope of Aristotelian moral education.

This does not mean that a virtue ethics approach to moral education is hopeless. As will be discussed in Chapter 3 as well as Section III, turning one's attention to the cognitive aspects required for virtue – reasoning, reflection, wisdom, and so on – indicates a possible direction forward for inculcating the virtues.

*Substantive moral education*

Beyond the specific methodology of moral education, there has been vigorous debate over the standards employed. There are roughly two positions: 1) moral education should be substantive, relying on a clear understanding of what is right and wrong and inculcating beliefs, motives, and actions that are in line with that understanding and 2) moral education should be substantively neutral, presenting various moral positions but not presupposing one or another to be correct.

Substantive positions have enjoyed considerably more support, perhaps due to the natural connection to the possibility of moral knowledge. (See, e.g., Wall 1975, Lickona 1996, Strike 1999 and Gilead 2009) So long as one believes it is possible to know right from wrong, the standards of moral education would easily follow: impart that knowledge to

[9] Aristotle suggest that people have a natural ability to expand on a sketch of the good (1098a22-6), using that to identify the exemplars of the good. But even then, it is unclear why external models of superior moral calibre are needed - people should be able to discern and reflect on other people's actions, deduce they are moral or not, and act accordingly. I will return to this Aristotelian approach to education in Chapter 11, on implications.

others, and teach them to act in accordance with that knowledge. This is to invoke the notion of a moral authority and presupposes that the moral teachers have greater moral expertise than the students. It also may find appeal because it fits naturally with the most common form of moral education – parents telling their children what is right and wrong. This is purportedly justified because of the idea that parents are more knowledgeable than children about what is right and what is wrong.

Immediately, however, this approach faces a similar difficulty as the virtue ethics approach. Substantive moral education relies on the notion of an identifiable moral expert. Being able to single out parents over children is one thing. But what about disagreements between different parents about what to teach children, or what teachers in public schools should inculcate in their pupils? The fact that moral knowledge is possible leaves open the crucial issue of how to identify moral experts, if there are any. One would need to be a moral expert to identify a moral expert, once again raising a circularity of sorts that undermines a project of moral enhancement.

A promising solution to this issue would be to generalize moral expertise: all or almost all adults possess it concerning at least some issues. We can identify those issues by looking to overlapping consensus. Everyone (or almost) can agree that theft, assault, and murder are wrong, while kindness and fairness are worthy dispositions. Moral education would then coalesce around these issues, leaving topics about which there is at least some disagreement aside. Thus, when teachers tell students that murder is wrong, they are justified because of their moral expertise. However, teachers who tell students that (say) abortion is wrong are not justified – the level of disagreement undermines any claim to moral expertise, and so obviates the possibility of moral education in that area.

More on this overlapping consensus approach to moral education will be discussed in the latter portion of Chapter 5. Suffice it say, for now, that overlapping consensus faces a problem of justification. It is just not clear how general agreement could ground moral expertise of the kind needed on specific issues. If some form of moral objectivism is correct, whether there is widespread agreement that some moral position is irrelevant to the question of whether that position is correct. Similarly, the fact that everyone agrees on some issue cannot on its own make everyone a moral expert on that issue, which would be necessary to license the sort of moral education necessary for substantive approaches. General agreement may provide some evidence in favor of a proposition, but it is not clearly reliable enough to be a good standard basis and as a guiding principle for moral enhancement would be overly conservative.[10] An alternative grounding of moral expertise will be necessary – and it seems unlikely, whatever that ground ends up being, that it will coincide neatly with the present areas of moral agreement and disagreement.

*Neutral moral education*

The alternative neutral approach can avoid the issue of moral expertise quite handily. By not committing to a particular moral view, there is no need to justify moral expertise in any given area. This does not mean morality can never be 'taught' – moral issues and dilemmas can be discussed, but the teacher would not take the goal to be inculcation of particular moral beliefs, motives or actions. Instead, the goal might be to make students

[10] The problem of fallibility and moral progress will be discussed in Chapter 6. On the other hand, some reason to trust most people's moral intuitions will be given in Chapter 8. However, that argument notably does not necessarily apply to people's expressed opinions, which may be miscommunicated or misunderstood.

aware of various moral views, or encourage them to think critically and reflectively about moral questions.

There are three main reasons one might adopt such an approach. The first is due to either a substantive commitment to moral relativism. The substantive commitment would take the truth of moral claims to be essentially relative (either to agents or groups). Teachers would be mistaken in trying to get students to conform to their own moral beliefs because moral truth is essentially relative – the teacher's commitment to, for example, the immorality of the death penalty in capital cases could be perfectly consistent with the student's commitment to the rightness of the death penalty in capital cases. There cannot then be a privileged pedagogical position in regards to moral truth or rightness, though students could be assisted in becoming aware of the moral ideas that are out there and how they can be contemplated and discussed.

The disadvantage of this approach is that moral relativism is a very objectionable doctrine. It has very counter-intuitive implications, including the idea that moral disagreement is actually an illusion and much moral condemnation of others' practices is wrong-headed (we would have to agree that, when a cruel dictator says his killing of peaceful protestors was morally justified, he was speaking truthfully). Also, relativism arguably relies on a gross misinterpretation of moral language. Notions of right and wrong presuppose notions of fallibility, that someone can have a moral belief that is mistaken. Moral relativism denies this (again, either for individuals or for groups), and so problematically deviates from how moral language is typically deployed. There is a sense, then, that relativists have changed the subject when proposing an alternative, non-substantive program of moral education. They are no longer talking about morality as most people conceive it, but rather some alternative category of judgment. Relativistic education would then not actually qualify as moral education worthy of the name.

But even if these objections can be overcome, it is not as clear as might appear that relativists should reject substantive moral education. Firstly, a relativistic education could end up being overly-dogmatic, indicating to students that all moral ideas are equal. This dogmatism is inimical to the relativist's moral pluralism, and substantive education (to ensure open-mindedness) may be needed to avoid it. (Cohen 1983) Secondly, relativism presupposes at least one substantive moral truth – the truth of moral relativism. So, in the very least, it would license inculcating moral relativism in students. The teacher then needs to be a moral expert in at least one area, moral relativism, and the preceding issues are not avoided after all. And thirdly, relativists are not in a position to reject or object to a substantive program of moral education. Just as the question of whether some moral position is justified is relative to a person or group that holds the position, the question of whether some program of moral education is justified is relative to some person or group. A relativist might not herself create a program of substantive moral education, but she cannot consistently say that such a program is wrong or mistaken – that is all relative.

The second motivation for adopting neutral education comes from a commitment to moral non-cognitivism.[11] Non-cognitivism is roughly the position that moral statements are non-propositional and so cannot properly said to be true or false. Instead, non-cognitivists (such as A.J. Ayer, Simon Blackburn and Alan Gibbard) typically interpret moral statements as expressions of one's (non-belief-like) attitudes towards certain actions, ideas, states of affairs, etc. This interpretation naturally lends itself towards substantively neutral education. As no moral statements are true or false, it would be a pointless exercise to try and inculcate in students the correct moral beliefs, motives and behaviors. This is to be contrasted with areas such as science, math or history where teaching (and indeed expertise) is quite sensible due to the truth-aptness of scientific, mathematical and historic claims. Teachers could fairly

[11] This approach is admittedly hypothetical and may not have many serious supporters; the authors cited are content to have substantive first-order commitments on the quasi-realist grounds referenced below.

express their own moral views, but they would have no basis for thinking a student with whom the teacher disagrees is in the wrong and must be corrected. Moral education could then consist not in inculcating particular beliefs, motives or behaviors but instead instructing students about the various moral attitudes (and metaethical positions) that various people hold.

Still, even more than relativism, non-cognitivism only tenuously supports substantively neutral education. Quasi-realism, a variety of non-cognitivism espoused by Simon Blackburn, is arguably compatible with substantive moral education. According to quasi-realism, while moral statements are not truly truth-apt, we can nevertheless properly behave and speak *as if* they are truth-apt. This allows space, in particular, for the idea of moral mistakes and improperly justified moral statements. A quasi-realist, then, could defend a program of (quasi-) substantive moral education not on the grounds of that genuine moral improvement is possible but rather on the grounds that moral educators can legitimately make (at least some) moral claims that could form the basis of moral education. In order to justify the teachers' privileged pedagogical position, quasi-realists can help themselves to whatever justifications employed by cognitivists – only with the qualification that those justifications involve a sort of moral fiction surrounding the truth-aptness of moral claims.

The third avenue for defending substantively neutral moral education is distinctly political and draws heavily on Rawls. In *Political Liberalism*, Rawls argues that the sate (at least when it comes to constitutional matters and the basic political structure of society) must be neutral towards comprehensive doctrines (including moral and religious views). Though Rawls is not entirely consistent on whether this liberal neutrality applies to systems of education, Matt Waldren and Kyla Ebels-Duggan have recently argued that it does. Waldren notes that, under Rawlsian liberalism, "Neutrality applies to this [political] domain because it has three important features: membership in it is non-voluntary, the institutions that constitute

it are coercive, and these institutions profoundly affect the course of citizens' lives." (Waldren 2011, p. 4) Education in western societies is indeed generally non-voluntary and coercive (students are required to attend institutions either run by or meeting criteria set by the state), and doubtless education has a profound impact on citizens' lives. The basic idea is that students should not be compelled to enter into an institution that instructs based on comprehensive doctrines that cannot be accepted by all citizens (Ebels-Duggan 2013). Education, then, must be substantively neutral and avoid imposing particular moral ideas (which generally fall into the category of comprehensive doctrines) on students.

The political justification for substantive neutrality comes with some significant limitations. Most prominently, like the realist and non-cognitivist approaches, it relies on commitment to a specific wide-reaching and controversial theory. If one thinks that comprehensive doctrines have a place in determining basic political matters, the whole account will fall apart. And Waldren's particular interpretation relies crucially on treating educational institutions as part of the basic structure of society. It is not clear that this is the case, however. Note, for instance, that education curricula are not typically construed as constitutional matters – rather, they are typically determined by suitable experts in the respective fields. The selection of those experts (indeed, the criterion on which their expertise is to be judged) may fairly be subject to comprehensive neutrality (just like the selection of all public officials). But once selected, those experts have at least some free rein to determine how students should be taught. It may well be within the scope of those experts' mandate to include substantive moral education as part of the curriculum.

Ebels-Duggan, by contrast, takes liberal neutrality to apply to all policies, not just those constituting the basic structure of society. Education thus falls easily within the realm of policies that should be justifiable to all. But because of significant substantive disagreement over a wide array of policies, a strict application of this standard would make

the state incapable of acting in the face of controversial disagreement over a wide array of issues such as tax policy, reproductive rights, consumer protection and so forth. For this reason, Ebels-Duggan conceives of liberal neutrality as a defeasible value – substantive education may be justified because other values (e.g., social well-being) take precedence.[12]

Indeed, there may also be room for a liberal program of substantive moral education without adverting to non-liberal values. In particular, Rawls' liberalism is, at least in part, a conception of legitimate government. Legitimacy is a clearly substantive idea, and thus so is Rawls' liberalism. Like the relativist, the liberal neutralist will then be interested in inculcating at least one substantive idea – political liberalism. That is taken by Rawlsian liberals to be the correct political theory, and given its centrality in the Rawlsian state, perhaps must be taught to citizens and citizens-to-be.

Indeed, this idea of moral education through substantive political education is arguably the closest to how moral education is undertaken in western democracies. The main institutional form of moral education, at least in the U.S., comes from civics class. There, students are instructed not only in how their government functions but also why that government is just. Those justifications are not as sophisticated as a Rawlsian picture, to be sure, but they are unmistakably moral in character. For instance, students are typically taught that people have a right to self-determination, that democracy is needed to protect the interests of the people from the whims of rulers, and that rights like freedom of speech have either (or both) instrumental and intrinsic moral significance.

[12] Ebels-Duggan also astutely observes that, pragmatically, no education approach is completely neutral. Even presenting competing perspectives may lead one to conceive of an equivalence between those claims, which is a substantive position. This leads Ebels-Duggan to advocate greater parental latitude in educational approaches on the grounds that parents may reasonably object to a 'neutral' state education curriculum as insufficiently emphasizing their own world-view. However, a policy of parental latitude is itself non-neutral; conservative parents could reasonably object that progressive parents are allowed to unduly sway their children towards progressive views, which will over time lead to more progressive policies affecting the conservative parents and their children. Ultimately, any approach will practically involve some imposition of substantive doctrine; the liberal challenge is then to *minimize*, rather than *eliminate*, substantive influence in education.

All this goes to show that any defence of thoroughly neutral moral education faces severe difficulty. The main defences surveyed above rest on controversial theories and ultimately endorse only partially-neutral moral education. However, the prospects for more fully neutral moral education are not completely hopeless. Later on in this dissertation, I will defend substantively neutral moral education (and enhancement in general) through a different route – one that focuses on indirect rather than direct improvement, emphasizes the presence and importance of disagreement in our society and systematically avoids commitment to (most) controversial substantive ideas. The result is a program that will be more generally acceptable than the preceding views - and ultimately more practicable as well.

## Argumentation

A more philosophical approach to moral enhancement can be found in the process of moral discussion and argumentation. A moral argument, in this context, consists of at least two parties disputing some moral issue, theory, dilemma or other topic. Typically, at least party (and likely both parties) will be attempting to convince the other party of the truth (or falsehood) of some moral claim. This could be accomplished by laying down a strictly logical argument from premises to conclusions, but will more likely involve elucidation of a series of points and counterpoints that purportedly support or detract from various positions. Argumentation is plausibly the most common way people become involved in (attempts at) moral enhancement, but the question remains whether it is a truly tenable route to moral enhancement.

### *Upsides*

Argumentation, we might expect, will lead to moral enhancement in two ways. Firstly, one party may be convinced (or at least become more open to) the other party's point of view. This shift will presumably occur because of the strength of that other party's case, making the shift more likely to be in the proper direction. In turn, behavior based on those positions should shift as well. Secondly, even if neither party is convinced (the most likely outcome), the exercise will force both sides to justify their positions and avoid counterarguments. This will inevitably involve some tweaking, making moral positions more coherent and better-grounded.[13]

This method of moral enhancement is strongly philosophical. Indeed, moral philosophy essentially consists of arguments for and against various moral positions. True, most moral discourse bears little resemblance to the philosophical literature, but the key features are there: an ostensible pursuit of truth, attempt to convince others through reasoning, and tacit assumption that good arguments will reliably lead to the correct moral conclusions. Any philosopher who puts much stock into the utility and importance of moral philosophy, then, should be supportive of argumentation as a means of moral enhancement.

In fact, this approach goes back to the early Socratic roots of philosophy. The approach of Socrates, as recorded by Plato, was to tease out the positions of interlocutors then use a series of questions to explore the grounding and implications of those views – more often than not revealing contradictions or fallacies that demand revision of the original positions. Key, though, is that Socrates (especially the early dialogues) often avoids making positive claims and instead focuses on rationally challenging the claims of others. But even in the later dialogues, arguably more a format for the exposition of Plato's own views, the back-and-forth nature of discourse remains. Socrates is portrayed as being on equal footing

[13] Much of this turns on the relationship between reasoning and morality, to be discussed more deeply in Chapter 9, but for now this prima facie case for the enhancing nature of argumentation will have to suffice.

with his interlocutors, working with them in a mutually-enlightening search for truth. And though philosophical literature since Socrates has gotten away from this direct dialogic approach, the strengths of the Socratic Method can still be seen in the discourses that emerge seminar rooms, conferences and coffee breaks throughout the philosophical community.

Another notable feature of moral argumentation is its level playing field. Education enhancement is an inherently unequal enterprise. It presumes one party is on a moral high ground, so to speak, and able to (more or less) unidirectionally impart knowledge on someone else. As we have seen, this leads to difficulties in establishing who qualifies as being on higher moral ground – who are the experts. Argumentation, by contrast, treats all parties as more or less equal. Neither party is presumed to be a moral expert or have greater access to the truth, so there are no antecedent problems of selection or moral evaluation. Any moral 'superiority' is determined *in situ*, and purely on argumentative merits (again, which we can expect to track the truth).

And finally, argumentation is a commendably flexible form of moral enhancement. While other traditional forms of moral enhancement must presuppose some morally privileged position and engage in an enhancement program based on them, argumentation relies on no such presuppositions. This makes argumentation much more dynamic and flexible. Any given position will gain currency only because of the strength of the arguments surrounding it; once some party comes along and refutes those arguments, that position will inevitably lose currency. In this way, an enhancement program based around argumentation will be constantly evolving – hopefully, getting closer and closer to the truth over time. (This dynamicity also helps avoid the problem of moral stagnation discussed in Chapter 6)

*Downsides*

In theory, argumentation may be an ideal means by which to accomplish moral enhancement. However, it is not a very practicable solution. Primarily, a moral enhancement program based in argumentation faces the problem of futility –arguments alone will generally not effect true moral change.

It might be thought that there is an in principle problem with reason-based argumentation that renders much persuasion in itself ineffective. The problem arises because many moral disputes concern basic moral ideas that are not easily susceptible to argument – strong, primitive intuitions. Consider, for example, a debate over whether homosexuality should be criminalized. One interlocutor argues in favor of such criminality based on a strongly-held intuition that homosexuality is deeply immoral, and we should criminalize deeply immoral acts. The second interlocutor agrees with the second part of the argument, but does not share the intuition that homosexuality is immoral. The problem is, it is not clear how either interlocutor could persuade the other to change their intuitions about homosexuality. The intuitions were not generated through a process of reasoning in the first place; their primitive nature seems to make such attempts futile.

This initial form of the worry is somewhat deceptive, however. Each interlocutor does indeed have powerful tools by which to persuade the other to abandon their intuitions. In the above example, the pro-criminalization interlocutor may independently accept that acts outside the control of an agent cannot be immoral (no 'ought not' without a 'can not'). If the anti-criminalization interlocutor could convey (via empirical evidence) that homosexuality is generally not a choice, this would force the interlocutor to either give up the intuition about choice and morality, or the intuition about homosexuality (or a commitment to the consistency of intuitions). Though one cannot guarantee that the pro-criminalization

interlocutor will give up the anti-homosexuality intuition, it is at least possible. This points to at least two ways to accomplish persuasion: reveal an inconsistency and demonstrate an empirical misconception. There are doubtless other means of persuasion besides (see, e.g., Hunter 1974), but these will suffice for now.

Nevertheless, even if moral persuasion via argumentation is possible, we must still ask whether a program of moral enhancement based around argumentation is likely to succeed. Here, we face much more considerable difficulties. How often, really, have moral arguments themselves persuaded interlocutors? In philosophical discourse, there is a surprising amount of moral intransigence surrounding a large number of ideas. Very rarely will one interlocutor in a moral dispute concede that he or she was incorrect and abandon his or her original positions. Shifts do occur, of course, but they tend to be incremental and technical. And argumentation becomes all the more difficult as a means outside the realm of philosophical discourse. In the public sphere, it is not very common that one can identify the power of some moral argument winning over large segments of the population. More often than not, the dominant forces seem to be non-rational rhetoric, manipulation of emotions and appeals to special interests. In this environment, arguments on their own stand little chance of being effective means of moral improvement.

In addition, the construction of a true program of moral enhancement through argumentation would face structural problems. Here, the non-directional aspect of argumentation turns from an asset into a disadvantage. Recall, argumentation does not place one individual into a privileged dialectical position. This means that, whereas moral education fits neatly into specific educational institutions, institutions to foster adequate argumentation are more elusive. One could simply try to generally encourage discourse at the social level, but such would be vague and directionless (and arguably not too different from current social practice). Any more formal attempt to institutionalize discourse (say, by

setting up monitored debate arenas for people to come and debate) seem rather redundant with current informal institutions and environments (television, radio, blogs, internet chat, and of course personal interactions), and may actually end up stifling debate through over-regimentation.

Alternatively, one might abandon the multi-directional aspect of argumentation and focus on exposing as many people as possible to what one takes to be the most powerful and insightful (and correct) moral arguments. This could be accomplished, for example, in a television program or newspaper column where one espouses moral views with the aim to persuade the audience. While this more or less solves the institutional problem, it only does so by effectively becoming another form of moral education – one privileged interlocutor sharing his or her wisdom with the masses. So once again, we have the problems of moral education such as how to select such an individual, while the appealing egalitarianism of the argumentation approach has been abandoned.

We should not surmise from this that argumentation is to be abandoned; far from it. Arguably, reasoned argument has helped bring about to moral progress in a number of social areas – emancipation, enfranchisement, liberal democracy, and animal rights, just to name a few areas. Nevertheless, argumentation alone faces significant limitations as a practical means of moral enhancement. Perhaps argumentation can enhance from time to time, and even does so reliably, but the magnitude of the effect and difficulty putting in encouraging proper argumentation make it an unpromising avenue for further development.

**Psychological Influence**

The final category of traditional moral enhancement surveyed here is loosely classified as 'psychological influence'. This is a term of art used here to group together a number of disparate techniques that involve neither pedagogy nor argumentation. As such, they will be typically – though not entirely – non-rational influences. Non-rational influences are those techniques that seek to make people more moral using means that bypass or subvert rational processes – that is, they do not seek to achieve conformity through direct engagement with our reasoning or deliberative capacities. These techniques include propaganda, social pressure, punishment, motive manipulation and more indirect psychological interventions.

*Propaganda*

Rational processes like argumentation or directly engaging in people's thinking about morality through education can be cumbersome and unreliable. It is easy, then, to understand why many groups and societies would opt to influence people through means that do not directly engage with people's rational faculties. This saves the need to develop powerful arguments or advanced pedagogical techniques, and instead focuses on a few 'tricks' to affect the desired change. Like pedagogy, this is unidirectional – the influencers wish to use various techniques to make (generally large groups of) people conform to their ideas of moral uprightness. But unlike pedagogy, there is no formalized educational component nor the pretense of knowledge-impartation. This avoids the need to make strong claims that moral knowledge and its transmission are possible, allowing for much simpler public justification for the techniques.

A propaganda campaign is the clearest form of this sort of influence. Such campaigns are generally centrally organized with a goal of bringing around the public to a certain point

of view (typically political). They are especially prominent during wartime to drum up patriotic support and vehement opposition to a country's enemies, but are also be used in a variety of contexts to rally support for certain causes. Propaganda takes the same form as advertisements – television spots, billboards, pamphlets, etc. While a propaganda campaign aims at selling an idea rather than a product, the techniques are quite similar: employ pleasing imagery for what the influencer wants to support and displeasing imagery for what the influencer does not want to support, rely on and confirm the audience's stereotypes and impressions on the subject, evoke certain emotional responses to bolster support for a cause, and so on.

While these techniques may make propaganda effective, they also make it particularly disturbing. That is not because the techniques of modern advertising are intrinsically odious, but become so when applied to moral matters. It may not matter much why we like one particular product over another, but it is indeed of critical importance why we support one moral position over another. What suffices as adequate moral justification is a deep and complex issue[14], but some approaches can be ruled out. Propaganda, in particular, relies on techniques such as imagery and association that has little to do with the content of the subject matter. Any group, position or idea – whether good or bad – can be presented with ennobling or, just as easily, demonic imagery. This divorce between the content of propaganda and the techniques used to improve people's morality make it singularly unappealing.

Against this, some may say it does not matter why we come to have any particular moral ideas, motives or behaviors – only that we end up with the correct ones. Propaganda would just, then, be the most efficient way to bring about positive moral change. Two things can be said in response. In the first place, we should not be so quick to give up on the

[14] The extent to which emotion and sentiment should play a role in the formation of moral ideas, motives and behavior will be discussed in Chapter 7.

importance of how someone came to some ideas. Ensuring a tight connection between the means of influence and the content of morality will make the given technique more reliable at achieving improvement. That lack of connection in propaganda campaigns makes them overly susceptible to error and corruption on the part of the influencer.

Secondly, ignoring the means of influence exhibits a certain level of disrespect towards those that are influenced. Propaganda campaigns treat them as mere receptacles of a moral view, rather than as moral agents with ideas and insights of their own. Even education, which is similarly unidirectional, engages with students' reasoning capacities and treats them as reflective beings capable of being able to evaluate ideas on their own. Insofar as propaganda does not even attempt to engage with these faculties, it is even more strongly paternalistic than educational approaches.

*Social influence*

Another prominent means of non-rational influence is social pressure and conformity. This is much less systematic than propaganda campaigns, instead involving a disparate set of more-or-less uncoordinated influencers. Social pressure can be subtle or overt. It may involve simply having most of one's associates possess some particular moral view and, without much thought, falling in line. It could be moderate, expressing blame or admiration based on certain behaviors. Or, it could be more prominent, with the dominant social position being explicitly reinforced and non-conformers ostracized. Wanting to fit in, someone could adjust their behavior and even ideas. These forces are relatively powerful, and help explain how many individuals within groups end up with very similar views on moral subjects.

Social influence has some interesting advantages over other forms of moral enhancement. It is arguably more respectful than propaganda and pedagogy, insofar as it relies not on one morally privileged individual or group but rather a group of individuals, none of which have any clear privilege. This is a sort of de facto moral equality. There is, to be sure, another form of moral inequality – the individual is in a distinctly subordinate position to the dominant ideas and behaviors of the group. But this is less disturbing, as there is some legitimacy to the group (insofar as it is composed of a large number of individuals) taking precedence over the individual.

Also, a reasonable case can be made that ideas, motives and behaviors originating in social conformity will be more reliably moral than those coming from individuals on their own. Whereas other techniques of enhancement rely on a (relatively small) number of individuals coming up with the proper morality then imparting it on others, group conformity instead relies on entire societies to have something approaching moral consensus. This will help weed out deeply-flawed positions or mitigate the influence of a handful of psychopathic individuals wreaking moral havoc on society. There is, then, an indirect check on group moral consensuses – very unreasonable ideas will be less likely to spread.

However, this picture of group rationality is overly optimistic. Conformity might in principle allow for rational evaluation, but in practice it is a strongly non-rational process. Consider the famous Asch conformity experiments (1951). When a room full of confederates gave obviously-incorrect answers as to the relative length of a line, most experimental subjects ended up also giving the obviously-incorrect answer at least some of the time.[15] The main thrust is that the conformity impulse itself is not a very effective check against 'crazy' positions.

[15] A replication of the Asch experiments found the phenomenon was not limited to geometric or even empirical facts, but also applied to the formation of opinions on more subjective matters. (Crutchfield 1955)

Perhaps the subjects in the Asch experiments had some legitimate reason to think the confederates had figured out some trick and optical illusion. More generally, we might not be so disturbed by unthinking conformity to the group if the group's results are very reliable. Here, however, we find further trouble. While more eyes on a problem might seem to be a boon, it can also run people astray. The phenomenon of groupthink is a particularly stark example of this. In groupthink, the desire for conformity within a group leads people to minimize critical evaluation, suppress challenges to the conventional wisdom and avoid external influences that might offer fresh, unbiased perspectives on an issue. This phenomenon can lead to any number of political ills, with groups conforming to some odious ideology (like fascism) and ensuring everyone is brought into line – ignoring or suppressing any dissent against the dominant view.

Beyond these practical worries, there is also an issue of will-substitution. When simply conforming to what the group thinks, there is a sense in which the group's will (such as it is) substitutes the individual's will. The person's ideas are not properly their own any longer, but instead are simply a reflection of the conventional wisdom. In this way, conformity serves to dehumanize, insofar as it removes from a person control of their own ideas, motives and behaviors. Perhaps some individuals do not care so much for such control and self-direction, but it is generally crucial if we want to remain moral agents who can properly be said to be responsible for our thoughts and actions.[16]

*Punishment*

[16] The issue of individualism will be further explored in Chapter 6, which addresses the value of moral dissent.

Social pressure can sometimes involve coercive activity (conform or be ostracized), but by far the more robust coercive systems are legal sanctions (both civil and criminal) that seek to punish wrongdoers. It may not seem so at first, but legal sanctions can indeed serve as a form of moral enhancement. Insofar as any such sanctions aim to alter behavior through disincentivizing certain actions and restraining the activity of certain individuals, they effectively serve as moral enhancements.[17] So, if the law prevents people from murdering one another, it has effected a moral enhancement (there is less immoral behavior). This form of enhancement is somewhat narrower than the other forms discussed; it is only directly concerned with improving moral behavior, rather than moral motives or ideas.

However, it could be argued that being compelled to follow just laws also improves motives and ideas in a more indirect manner. Here we can return to Aristotle, who thought that legal sanctions that enforce moral conformity can be another form of habituation that helps train people to act in good ways. In this way, Aristotle writes, "Then perhaps also someone who wishes to make people better by his attention, many people or few, should try to acquire legislative science, if laws are a means to make us good." (*Nicomachean Ethics* 1180b24-6)

As noted earlier in this chapter, it is not at all certain that this habituation can really help bring about the internal moral changes Aristotle hoped for. Indeed, it could have just the opposite effect. Consider a law that would impose legal sanctions for failing to keep promises to friends. Putting aside concerns with privacy and enforcement, the law would have a troubling effect on reasons for action. Whereas presently a primary motivation to

[17] At the same time, purely retributivist punishment cannot be said to aim at moral enhancement. Certainly, a Kantian framework, which would exclude punishing as a means to obtain some further good, including moral improvement of the criminal or society, would therefore exclude using punishment as a means for moral enhancement. Similarly, more modern conceptions based around the intrinsic goodness of punishing wrongdoers (see, e.g., Moore 1987) do not fit into a framework of moral enhancement. This subsection, therefore, will only apply to justifications of punishment that allow for at least some consequentialist considerations.

keep such promises will be more or less moral – it's one's duty to live up to one's word. Under the law, however, these motives would change. The desire to avoid punishment could easily become the dominant reason one keeps one's promises, which is not the sort of motive that underpins truly virtuous activity. So, in general, any moral sanctions aimed to improve behavior run the risk of having the converse effect on ideas and motives, corrupting the reasons people act.[18]

Further doubts could be raised about the effectiveness of certain punishments at deterring crime. For example, there has been considerable debate over whether the death penalty in fact deters violent crime (see, e.g., Ehrlich 1975 & Passell 1975). Indeed, it is crucial for any justification of punishment as a way to reduce crime (and thus count as a moral enhancement) that the punishment actually reduce crime. However, it is outside the scope of this dissertation to evaluate those claims. Here, we must be contented to note that, while almost every society may have long ago accepted and instituted this particular means of moral enhancement, it is only a partial solution. Despite the punishments that are widely propagated, criminal and other sorts of immoral behavior remain (and in some areas, widely so). In the very least, we should be looking at other supplementary avenues for improvement.

*Indirect psychological influence*

The various forms of psychological influence discussed here thus far have, for the most part, constituted direct moral enhancement. As explained in Chapter 1, direct moral

[18] There is some evidence for the same sort of effect with rewards for good behavior. When people are offered monetary compensation for blood donation, donation rates fall. (see, e.g., Mellstrom & Johannesson 2008) This seems to be due to a reconceptualization of the effort; a previously altruistic framework (where donation is worthwhile for the sake of the greater good) is substituted for an economic one (donation is not worthwhile for the self-interest of the donor). This indicates a shift in the moral character of the donation – the incentives shift the reason for donation from altruistic to self-interest.

enhancement involves the enhancer attempting to bring the target of enhancement in line with some preconceived notion of the morally proper behaviors, motives and ideas. Indirect moral enhancement, by contrast, involves identifying some trait, capacity or other factor that is linked to moral behavior, motives and/or ideas and altering it in such a way that will reliably lead to moral improvement. One advantage of this approach is that difficulties in working out moral knowledge and expertise can largely be avoided.

Indirect psychological influences, then, involve specific psychological interventions that one expects would lead to moral improvement. There has not been a huge amount of interest in these sorts of means of moral enhancement within academia, and precious few attempts to apply them at the social level. Still, the progress that has been made is quite instructive.

One of the more successful attempts at such indirect psychological influence comes from impulsivity and reflectivity. A 1973 study, for instance, found a powerful connection between greater reflection (fewer impulsive responses) and moral maturity among children. (Schleifer and Douglas 1973) Moral maturity was measured in Piagetian terms, specifically through whether individuals take into account agents' intentions, or merely consequences (more on this rubric below). Why, exactly, the connection between impulsivity and maturity? The best explanation seems to be that certain features of moral maturity, such as the importance of intentions, requires some thought. Immediate reactions, at least among children, focus on the most salient conditions (observable behavior), but children who take more time can better appreciate the nuances involved in decision-making.

The 1973 study merely found a correlation, but it indicates a path forward for moral enhancement: make people less impulsive and more reflective. This hypothesis has, in fact, put to the test by Lopez & Lopez (1998). The researchers designed an intervention to

improve reflectivity in children by training them to give delayed responses, scan for details, plan out problems, and go through internal dialogue. They found that children who had gone through such a regimen (and had greater reflectivity as a result) had greater levels of moral development compared with children who had not (This time measured on a Kohlbergian scale of post-conventionality). Again, these results should not be too surprising, as taking more time to consider a problem can generally lead to more insight into the complexity of an issue. In addition, it points the way towards moral enhancement that does not involve commitment to specific normative ideas. And whereas most of the above interventions rely on theoretical or intuitive analyses, this is an example of scientifically-informed enhancement that could thereby be a much more reliable and effective means of moral improvement.

All that having been said, studies like the one just cited are not truly indirect. They each crucially involved theories of moral psychology (Piaget and Kohlberg) that actually contain fairly substantive moral claims concerning ideas, motives or behaviors. The concern over intentionality, for instance, presupposes a framework in which intentions are morally crucial – ruling out substantive moral theories such as utilitarianism that discount the importance of intentions. Similarly, an emphasis on distinguishing moral rules from conventional rules excludes relativistic moral frameworks and does not fit well with certain contractualist accounts (e.g., Scanlon 1998) that rely on mutual justifiability. That is to say, the studies presuppose a certain idea of what it is to be moral and attempt to use psychological interventions to make people fit that idea. Along the way, dubious appeals to authority (Piaget and Kohlberg) undermine the argument that such interventions can be effective means of moral enhancement. Also, the studies' focus on children's development belies the immensely greater controversy that would emerge in measuring comparative maturity in adults, who are ultimately the main target of regimes of moral enhancement.

Unfortunately, this shows how scientific robustness is in strict tension with indirectness. If a purported moral enhancement can be experimentally evaluated, that means there is some measurable moral endpoint. But any such definite moral endpoint implies the experimenter has already determined what are the morally proper behaviors, motives and/or ideas. Instead, a truly indirect moral enhancement will have to rely on more theoretical underpinnings that link the intervention with reliable moral improvement. Just such a theoretical argument is the aim of Part III of this dissertation (Chapters 7-11); once that theory has been worked out, we will be able to return to more practical, empirically-evaluable measures.

*Motive manipulation*

The preceding subsections have delineated various means of psychological influence by the external mechanisms involved. It can also help to delineate a further category, admittedly intersecting with many of the above interventions, in terms of its internal character. By bringing out this internal character, we will be able to better analyze a more subtle problem many of these approaches encounter. This internal character will be referred to as 'motive manipulation,' a term that comes from the illuminating article by Eric Cave ("What's Wrong with Motive Manipulation?", Cave 2007) that will be the focus of this subsection.

At the broadest level, all attempts at moral enhancements are manipulations (they seek to alter others' thoughts and actions). However, motive manipulation is a more narrow and problematic category. Cave defines motive manipulations as interventions that "…mobilize some non-concern motive of such an agent so as to induce her to behave or

move differently than she would otherwise have behaved or moved, given her circumstances and her initial ranking of concerns." (ibid, p. 132) Non-concern motives are neither pro- nor anti-attitudes, but can nevertheless incline a person to act. For example, an implicit (i.e., not consciously entertained but nevertheless effective) bias in favor of one's own race would be a non-concern motive. A propaganda piece that uses racial stereotypes to engender opposition to an international foe could then be an instance of motive manipulation; the implicit bias is deliberately triggered to alter the agent's attitude towards the foe.

What makes motive manipulation particularly problematic? Non-concern motives are, in a certain sense, outside a person's direct conscious purview and control. Insofar as those non-concern motives affect behavior, they do so by taking a person's thoughts and actions outside of their control. The greater the influence of non-concern motives, the less control a person has. So, when an external agent specifically seeks to take advantage of such a non-concern motive, they induce a person to lose control over their own lives. This is, as Cave argues, a problematic violation of our autonomy. Autonomy is quite literally self-government, which we generally find to be of great value. Motive manipulation takes the governing away from the individual, and puts it in the hands of these non-consciously-endorsed motives as well as an external agent. We should generally refrain from such manipulations (not necessarily as an absolute prohibition, but at least a strong prima facie consideration), out of respect for the value and importance of people's self-government.

These objections to motive manipulation may not be completely worked out. For instance, Cave's definition arguably casts a pall over too many actions. Consider, for example, the force of a forgotten promise. Strictly speaking, it is a non-concern motive (it is not presently a pro- or anti-attitude, but could under certain circumstances incline you to act). When a friend reminds you of a forgotten promise, a non-concern motive would then be mobilized to change your ranking of concerns (you will now keep the promise) – thus

counting as a motive manipulation. But it is hard to see anything particularly wrong with such a reminder, nor does it plausibly count as infringing on one's autonomy.

We may, then, have to narrow our understanding of motive manipulation, or at least what makes it problematic. One route is to exclude as problematic cases where the agent, while not consciously entertaining the non-concern motive, would hypothetically endorse the influence of such a motive should she become aware of it. Or, if one is wary of the normative force of hypothetical endorsement, one could rely on more generic attitudes (e.g., 'I want to keep my promises'). Any mobilization of non-concern motives that aims at respecting such a generic attitude (thus really aiming to bolster a person's self-government, not inhibit it) would not count as an autonomy violation.

Another issue is that the comparative degree of self-government in the presence or absence of motive manipulation is unclear. Arguably, motive manipulation involves just as much agential control as, for example, persuasion. In both cases, a person will be influenced – unintentionally or no – by innumerable unconscious biases, tacit considerations and suppressed premises. Motive manipulation simply replaces the influence of some of those non-concern motives with others. So, in the instance of the racially-tinged propaganda campaign, implicit racial bias may influence a person's behavior more because of the piece – but, say, subconscious reluctance to take political positions (another non-concern motive) may be suppressed. The agent would consciously endorse neither non-concern motive, so the solution from the preceding paragraph will not be of help here. In effect, we can reconceptualize motive manipulation as motive replacement. We're going to be subconsciously influenced beyond our control no matter what, and these interventions merely shift what those influences are.

Two things can be said in response to this. In the first place, it is not entirely plausible that motive manipulations will involve a one-to-one replacement of non-concern motives. Assuming that conscious attitudes have at least some significant influence on behavior,[19] there will be some baseline 'balance' between conscious and non-concern motive influence. By enhancing the influence of a certain non-concern motive, a manipulator would then typically be 'tipping the balance' in favor of non-concern motives. If the manipulation actually manages to change behavior, then it is reasonable to think that the shift in the balance was quite significant – enough to tip the scales. It is still theoretically possible that the entire shift in influence came through shifting around non-concern motives, but without further argument (and likely some fairly robust empirical evidence) we should be dubious of such an outcome.

Secondly, it is important that the motive manipulator is purposefully intervening to alter a person's behavior. While the influence of non-concern motives may on its own run afoul of self-government, it is still 'closer' to the person than the external agent. That is to say, implicit biases and other non-concern motives are still in a certain sense part of the agent. When another agent manipulates those biases, however, the primary impetus of psychological change is external to that agent. Even if the manipulator simply replaces the influence of one concern motive with another, the change is still a violation of self-government insofar as it further removes direct control over the agent's activity from the agent's own psychology. This intersects with the concerns discussed above over will-substitution (an agent moves from self-government to manipulator-government), and underpins a key problem with motive manipulation.

[19] One may wish to contest this, instead insisting that all conscious activity is simply a reflection of predetermined cognitive activity or unconscious mental processes and any ideas of personal control are an illusion. There is not space to argue against such a position here, and such arguments – which effectively deny the presence and thus moral relevance of autonomy – will have to be set aside.

Cave's argument that motive manipulation is morally problematic is, then, still relatively persuasive. Insofar as psychological influences will gravitate towards such manipulation, they will end up being morally problematic. Even if such were effective means of moral enhancement, the means itself could be deeply objectionable insofar as it ends up violating people's autonomy. We should, then, seek to focus on interventions that can effect moral enhancement while nevertheless respecting people's autonomy.

This chapter has sought to give an overview of the various means of traditional, non-biological moral enhancement. As we have seen, each poses problems of one sort or another. None of these problems are insurmountable, but they do point in the direction of what a defensible form of moral enhancement will amount to: one that avoids or overcomes issues of moral expertise and knowledge, is practically implementable through institutions, respects the autonomy of the enhanced, and has a solid theoretical grounding. Later sections in this dissertation will attempt to evaluate whether direct moral enhancement can, in general, overcome these issues and will ultimately argue that indirect approaches are the most promising way to achieve moral enhancement. First, however, let us examine some more modern and biological methods of moral enhancement and the (related) issues they raise.

# Chapter 3: Biological Moral Enhancement

We will now shift from the past of moral enhancement to its future. Indeed, much of the recent interest in moral enhancement has been in large part due to the prospect of biological interventions such as pharmaceuticals, genetics and neurosurgery that could make people more moral. These techniques may not be so different in kind from traditional forms of moral enhancement, especially those that focus on psychological influence, but they arguably have much more potential in terms of degree. While we may be able to marginally manipulate people's ideas behavior through propaganda, social pressure, punishments and motive manipulation, directly changing the way people's brains process ethical dilemmas and respond to particular situations through biological interventions could bring about much more dramatic change. This makes biological enhancements both more promising in terms of actually bringing about change, as well as more perilous because the effects of wrongheaded attempts will be much direr. But before turning to these potential pitfalls, we should look at whether manipulation of moral behavior, motives and ideas is indeed feasible, or consigned to the realm of science fiction.

## Empirical evidence for moral manipulation

One potential source of skepticism about the possibility of direct biological moral enhancement would be doubts surrounding the feasibility of manipulating the moral character of our mental states, motives and actions. John Harris (2011) has raised this concern vis-à-vis one prominent motive identified by Tom Douglas (2008) as uncontroversially immoral:

racism. Racism, Harris postulates, is too cognitive in nature to be directly manipulated using biological interventions; it is largely constituted by a set of false, prejudiced beliefs that are too cognitive to be affected through biological mechanisms, which will tend to operate at conative or emotional levels. The only feasible way to rid people of such immoral motives would be to either to provide indirect biological interventions, such as those that would improve rationality, or to employ old-fashioned educational or socialization techniques to correct people's false beliefs.

This objection could be dealt with by pointing out that that racism can involve emotional as well as cognitive attitudes. Even if directly manipulating cognitive attitudes is not feasible, an enhancer could instead focus on altering people's emotional attitudes – in particular, their basic emotional reactions to people of different races. (Douglas 2011) Perhaps this misses the force of Harris's objection, and the point is that the cognitive nature of racism is what really matters. However, it is dubious that the immorality of racism is so limited – presumably, it is important to not only correct racist beliefs but also change behavior. Moreover, it could very well be that changing people's emotional reactions to different races will also alter their cognitive attitudes about race. Perhaps people form negative beliefs about members of other races based in part on negative emotional reactions to members of other races. This is an empirical hypothesis and cannot be evaluated here, but it at least suggests that racism may be more prone to biological manipulation than Harris contends.

This debate between Harris and Douglas over mitigating racism points to a general question: to what extent is it actually feasible to manipulate the morality of people's ideas, motives and actions using biological interventions? If such manipulations are confined to the realm of science fiction, then the debate over biological moral enhancement would be of little practical importance. Indeed, objections to such enhancement would be idle – there is not

much reason to be concerned about interventions that are for all practical purposes impossible.

However, recent empirical findings clearly demonstrate that biological interventions can have an influence on the morality of people's apparent beliefs, motives and actions. I will now spend some time surveying some of these findings, especially concerning serotonin and oxytocin. These studies conclusively establish that moral manipulation is a real possibility. That is not to say these studies show that moral enhancement is clearly possible; in fact, the ambiguity of whether any given manipulation constitutes an improvement provides reason to doubt that we could be confident that any intervention counts as an enhancement at all. These issues will be discussed in much greater detail in Chapter 5.

*Serotonin*

Serotonin is a neurotransmitter that regulates a wide variety of cognitive and non-cognitive functions. It is perhaps best known for its effects on mood; the most widely-prescribed treatments for depression are medicines such as selective serotonin reuptake inhibitors (SSRIs) that increase the level of serotonin in the body. Such drugs have also had some success in diminishing the violent behavior of psychopaths.

These latter interventions provide some initial indication of how drugs such as SSRIs could be moral manipulators – specifically, by manipulating people's moral behavior. Much violent or aggressive behavior does indeed seem paradigmatically immoral, and aggression is another one of Douglas's paradigmatic cases of immoral motives. Research has shown that low levels of serotonin are correlated with increased aggressive behavior, while boosting

people's serotonin levels makes them more cooperative and less retaliatory. (Krakowski 2003)

Indeed, Serotonin's effect on retaliation goes beyond strictly aggressive behavior. The ultimatum game is one particularly useful way to test people's willingness to retaliate against others for treating them unfairly. In the game, one player is the proposer and the other the responder. The proposer is given a sum of money and must decide how to divide between the two players. The responder is then given two options: accept the offer, in which case the money is divided accordingly, or reject the offer, in which case both players receive nothing. The responder has nothing to gain monetarily from rejecting offers, yet players frequently reject offers in which they would receive less than about a 30% share of the money.

One might wonder whether affecting the rate at which people reject offers in the ultimatum game is truly a moral manipulation. Two things can be said here. One, there is an intuitively moral explanation to these rejections; players see low offers as deeply unfair, and wish to punish (or perhaps express disapproval to) the proposer. This is attractive not only because it explains the phenomenon, but it matches up well enough, based on introspection, to how we might feel in such scenarios. Moreover, two, those who reject low offers tend to explain their actions in morally-laden terminology – denigrating the character of the proposer and complaining about being treated unfairly. (Kravitz and Grunto 1992) But, it should be noted, claiming that there is a clear moral character to these rejections does not imply that the morality of these rejections is clear. That is to say, it is not clear whether responders are morally justified in rejecting the offers. This complication will arise in various guises for almost any biological interventions, and the implications of this moral ambiguity will be discussed in later chapters.

Returning to the topic at hand, recent studies have found that serotonin has an unambiguous effect on those playing ultimatum games. One study measured serotonin levels in the blood of players; responders who refused low offers had markedly lower levels of serotonin. (Emanuele et al 2008) Another set of studies found that experimenters can actually affect the likelihood that someone will reject a low offer. Inhibiting serotonin production by depleting people's tryptophan levels leads to lower acceptance rates, (Crockett et al 2008) while boosting serotonin levels by giving people SSRIs leads to higher acceptance rates. (Crockett et al 2010) Given the above analysis of ultimatum game rejections as distinctly moral actions, this implies that we can indeed manipulate the moral nature of people's actions.

But what about mental states like beliefs? As it turns out, serotonin does not seem to have an effect on people's explicit moral judgments concerning the fairness of offers; lowering someone's serotonin levels (and thus increasing the likelihood that he or she will reject a low offer) does not cause that person to judge low offers as more unfair. (Crockett et al 2008) However, serotonin has been demonstrated to have an effect on people's moral judgments concerning the permissibility of directly harming one person in order to save many others from harm. The classic 'trolley problem' illustrates this sort of dilemma: is it permissible to push someone from a bridge, killing that person, so that their body will stop a train that is about to run over five others trapped on the train tracks? This question is contrasted with one where the train is careening towards five people trapped on the tracks, but one could pull a switch that diverts the train towards a separate track, on which only one (different) person is trapped. The effects of the two options are the same in both cases; either you cause one person to die so that five may live, or you do nothing, allowing five to die. However, most people identify a moral difference between the two, claiming it is permissible to pull the switch but not push someone from the bridge.

Boosting people's levels of serotonin through the SSRI citalopram has been shown to make them more likely to disapprove of pushing someone from the bridge in the above trolley case, while it has no effect on approval of pulling the switch. Interestingly enough, this effect is only found in those who have high levels of empathy, as measured by personality tests. (Crockett et al 2010) Even so, it clearly demonstrates that moral beliefs can indeed be manipulated through biological interventions – though, once again, we should note that it is not clear whether it is in fact morally permissible to push someone from the bridge to save five others.

It is more difficult, if not impossible, to find empirical evidence that demonstrates that the moral character of people's motives can be manipulated, via serotonin or any of the other means discussed below. Such would require actually being able to observe or detect motivational states, perhaps using neuroimaging to detect the effect of a given intervention on people's brains. In addition, we would need a robust understanding of how those brain states cause or motivate someone to act as well as a robust account of how certain brain state/action hybrid has an explicitly moral character. Still, we may be able to reasonably infer from the fact that we can use serotonin to manipulate people's moral beliefs as well as the actions that we could (perhaps more speculatively) manipulate the moral character of their motives, which are something like the union of mental states and actions.

*Oxytocin*

Sometimes referred to as the ‘love hormone’, oxytocin is a neuromodulator that, like serotonin, primarily affects brain activity. It has been associated with generally pro-social and affective behavior, strengthening interpersonal bonds and perhaps underpinning behavior such as kin and group loyalty. (MacDonald and MacDonald 2010) These effects themselves have a vaguely moral character – arguably, much of morality is based on concern for other members of society – but it is possible to tease out more particular ways oxytocin can manipulate morality.

Trust is a crucial aspect of many value-laden human activities, including friendship, loyalty, promise-making and cooperation. While one can be both too trusting and not trusting enough, it seems clear that the degree to which one trusts other people will have a significant effect on the moral character of one’s interpersonal relationships. This effect may be somewhat less specific than some of the others discussed here, but at the same time it is conceivably much more far-reaching.

One way to measure trust is through an investment game: one player, the investor, is given a certain sum of money that he or she can either keep, or give to the second player, the trustee, as an investment. If the investor gives the money to the trustee, the experimenter triples the amount received; so, if the investor gives $1, the trustee receives $2 extra from the experimenter. The trustee then has the opportunity to split the proceeds with the investor, give an unequal portion back, or keep everything for himself or herself. An alternative lottery game (used as a control, as it does not measure trust but just risk-aversion) has a similar setup, only the trustee is not a person; whether the investor receives his or her money back is explicitly determined at random.

Giving investors a boost in their oxytocin levels has been shown to increase their likelihood of providing trustees with investments in the trust game; it has no such effect on

the likelihood of investing in the lottery game, indicating that oxytocin increases the extent to which people trust other people. (Kosfeld et al 2005; Baumgartner et al 2008) One plausible explanation for this effect is that oxytocin reduces the extent to which investors fear that the trustees will betray them. But whatever the explanation, the implications could be far-reaching; oxytocin could make marriages, friendships and other relationships more trusting, inhibiting suspicions that are both well- and ill-founded. Boosting oxytocin could be a moral manipulation insofar as differing levels of trust can have a profound impact on the moral nature and value of relationships.

Oxytocin has also been found to have an effect on generosity. In ultimatum games, proposers who were given oxytocin ended up offering responders significantly more money compared with those who were given a placebo. (Zak et al 2007) Interestingly, the same study did not find the any effect of oxytocin on offers in dictator games (where responders do not have the option to reject offers), leading the study authors to conclude that oxytocin did not boost altruism directly, but rather made people more generous in situations where they had to actively consider other people's perspectives. At the same time, a separate study found that people who have genes associated with greater production of oxytocin tend to give fairer offers in dictator games. (Israel et al 2009)

This suggests at least two separate means by which one can manipulate people's moral behavior – in particular, the extent to which they are willing to sacrifice their own interests for the greater good – through oxytocin. One could provide them with oxytocin, which seems to affect the extent to which people react empathetically when they consider others' perspectives. Further study on this subject would be needed, but this could easily end up affecting the extent to which people respond to the moral claims of others or, more directly, to others' perceived distress. The other, more speculative means would be through gene selection and manipulation (perhaps at the embryonic stage of development); we could

select for genes that promote oxytocin, which have been shown to promote people's altruistic behavior.

If the preceding makes oxytocin seem like an all-around great hormone, it is important to note its darker side. In a series of experiments, Carsten De Dreu and colleagues (2011) found that Dutch subjects who had been given oxytocin were more ethnocentric than those given a placebo. One of the three ethnocentrism tests carried out by de Dreu and colleagues is particularly worth mentioning: subjects were asked about hypothetical dilemmas like the trolley case, in which responders were asked if they would sacrifice one in order to save five. In some cases, the person to be sacrificed had in-group (Dutch) names, while in others the person to be sacrificed had out-group (Arab or German) names. Those on placebo showed no discrimination; they were equally willing to save people with in-group names as those with out-group names. However, those who had been given oxytocin showed a significant preference towards those with in-group names. Strikingly, oxytocin administration generated discriminatory judgment (which is rather morally problematic) where none had existed before.

In addition to complicating potential enthusiasm for oxytocin, this may seem to vindicate Tom Douglas's claim (2008 and 2011) against John Harris, mentioned above, that racist tendencies can indeed be manipulated directly through biological means – indeed, even through manipulating cognitive states themselves, as in the case of judgments about moral dilemmas. We should be cautious with such claims; De Dreu and colleagues (2011) found that the ethnocentric effect was primarily caused by in-group favoritism, not out-group derogation (the latter of which seems more characteristic of racism). Nevertheless, critics like Frances Chen and colleagues (2011) are wrong to suggest[20] that such 'positive'

[20] "Goodwill is not a fixed pie, and increased goodwill to in-group members does not necessarily imply any change in goodwill to out-group members." Chen et al 2011, p. e45

ethnocentrism is unproblematic. We live in a world of finite resources; the actual effect of ethnocentrism, whether motivated by in-group favoritism or out-group derogation, is to discriminate against out-groups – whether in interpersonal relationships, hiring, policy or any other arena.

This effect of oxytocin on in-group bias should not be too surprising. Such biases seem motivated by strong bonds to one's own people, and oxytocin has been shown to strengthen those very sorts of bonds. These may have apparently good effects, such as by increasing altruism, as well as quite bad effects, such as by boosting ethnocentrism. Still, it should be clear that the effects are explicitly moral; oxytocin directly affects the moral character of people's mental states and behavior – and probably, as with serotonin, motives as well.

*Other interventions*

Serotonin and Oxytocin are perhaps the best-studied biological means of moral manipulation, but that does not mean they are they only feasible ones. Giving male proposers testosterone in ultimatum games, for instance, has been shown to reduce the generosity of their offers. (Zak et al 2009) Nor should we expect biological interventions to be limited to complex pharmacological treatments. One rather shocking study found that the timing of food breaks had a profound effect on Israeli judges' leniency; immediately after a food break, parole judges found about 65% of cases in the parolee's favor, but that percentage drops steadily as the day goes on, reaching 0% immediately prior to the judges' next meal break. (Danziger 2011) Providing food to judges in order to maintain the fairness (or at least consistency) of their decisions might stretch the meaning of 'biological' intervention, but it

demonstrates the potentially very diverse range of means by which people's moral beliefs, motives and actions could be manipulated.

It should be evident, however, that all manner of diverse interventions will likely run into similar complications as those found with oxytocin and serotonin. These problems are of two sorts: the value or desirability of many intervention's influence on the moral character of people's mental states, motives and actions will be rather ambiguous and the subject of intense moral disagreement (as with trolley cases and ultimatum games); and, the effect of any intervention will likely not be domain-specific but will rather affect the moral character of a wide variety of mental states, motives and actions, to such an extent that there will be ambiguity and intense disagreement over the proper balance of these different moral features (as with oxytocin's effect on generosity and ethnocentrism). Chapter 4 will offer further arguments that these difficulties are a general problem for most biological interventions that directly manipulate morality, casting doubt (for primarily epistemic reasons) on the possibility that we can use such interventions to morally enhance people.

Before turning to those concerns, however, let us examine briefly the current state of the debate over moral enhancements.

**Debates surrounding biological moral enhancement**

Prior to recent debate concerning moral enhancement, there have been long and fruitful discussions of potential problems with biological interventions aimed at enhancing cognition, health, beauty and other areas of living. From various corners, objections have been leveled against the prospect of widespread human enhancement. Some of these worries (such as concerns about the unnaturalness of the interventions or the failure to accept our

given nature) are, I believe, rather confused and unconvincing, while others (such as concerns that enhancement technologies, if not properly regulated/distributed, will exacerbate already-endemic injustices and inequalities) raise serious problems that proponents of enhancement should attend to. Nevertheless, I will for the present discussion bracket these background concerns over enhancement. My purpose here is to discuss unique problems raised by the prospect of moral enhancement not characteristic of biological enhancement in general. I will, then, assume for the sake of argument that biological enhancements are not in general impermissible and in fact are often worth promoting (insofar as things like intelligence and health are beneficial to individuals and society; the uncontroversial nature of such claims of benefit contrasts with controversies surrounding morality).

Recently, several unique arguments both for and against moral enhancement have been raised. On the 'pro' side, Persson and Savulescu (2008) have argued that moral enhancement is critically needed to avoid catastrophic loss of life that is becoming ever more likely as our technological capabilities develop. As knowledge is spread (and people become smarter, perhaps via cognitive enhancement), the risk grows that some rogue terrorist group or doomsday cult could develop and deploy a nuclear or biological weapon of mass destruction. (ibid, p. 166) At the same time, reformed moral dispositions could lead people to act to significantly mitigate environmental damage as well as alleviate global poverty. (Persson and Savulescu 2010, pp. 663-5) On these consequentialist grounds, they propose a program of widespread moral enhancement, in particular improving people's sense of altruism and justice.

While Persson and Savulescu's arguments are compelling and have broad appeal (any plausible moral theory would laud the goal of preventing terrorist attacks or reducing global poverty), it is not clear how their goal could be achieved without compulsory moral enhancement – that is, without governments compelling citizens (without their consent) to

undergo certain biological treatments that may not even, strictly speaking, be in any given individual's interests. The alternative of compelling enhancement of neonates (who cannot sensibly be coerced) would similarly violate the reproductive autonomy of parents who disagreed with the enhancement program. Passive programs such as putting enhancers in the water supply would similarly run into the problem where people's bodies are being significantly manipulated without their consent. Perhaps some individuals in a state of akrasia – realizing, say, they should give much more money to charity but failing to do so – would be motivated to take interventions that improved their good behavior, but far too often such bad behavior is rationalized; certainly, a violent terrorist would not typically think they are doing anything wrong and would refuse any intervention that inhibited them from achieving their goal.

One could argue that the stakes are just too high and such compulsion is a necessary evil. However, the immorality of widespread compulsory biological moral enhancement is indeed incredibly weighty; it involves the gross violation of personal autonomy, both in terms of bodily autonomy as well as (arguably more prominent) freedom of thought, and calls to mind dystopic states such as those found in the writings of Huxley or Orwell. And the benefits of such widespread moral enhancement are not entirely clear – we may well get the proper moral balance wrong. To avoid these problems, widespread enhancement could take softer, less coercive forms. Further problems with such softer programs will be discussed in Chapter 5.

Conversely, Fabrice Jotterand (2011) has argued that at least one sort of improvement – moral enhancement via neural manipulation – is not actually possible because we can only become better, more moral people through careful, reflective exercise our moral agency, something not characteristic neural manipulation. Jotterand explicitly adopts a virtue ethics framework (contrasted with Persson and Savulescu's consequentialism) and she holds that

true virtue requires a certain sort of interaction between emotion and cognition; such cognition involves the right sort of reasoning and deliberation.

Two responses to Jotterand are warranted. One, a proponent of enhancement could simply admit that one cannot enhance the virtues, but there are other forms of moral enhancement (such as improving people's moral beliefs or actions) that are nevertheless possible. Indeed, Persson and Savulescu's aim of preventing suffering and death seems unaffected by Jotterand's worries about virtue. Two, even within a virtue ethics framework, moral enhancement may be possible. As Barbro Elisabeth Esmeralda Fröding (2010) has argued, some forms of manipulation (such as cognitive enhancement) can actually enhance virtue by promoting the sort of rational deliberation that is so central to a virtuous life. This could involve strengthening inferential abilities, removing cognitive biases, and improving one's ability to attend to and comprehend all the salient facts. Indeed, on certain Aristotelian accounts, there is not a clear distinction between virtue and rationality. Insofar as the rationality could be enhanced, so could virtue.

Importantly, the sort of indirect moral enhancement suggested by Fröding (as well as Douglas and Harris), improving rationality or reasoning, rather than affecting the morality of mental states, motives or actions themselves, could potentially answer some of the concerns explicated below concerning disagreement. That is because such rationality enhancement would be value-neutral,[21] sidestepping problems arising from our deep moral disagreements. The prospects of such indirect moral enhancement, however, are outside the scope of this paper, and must be left for discussion at another time. The important point here is that Jotterand's objections to moral enhancement are unconvincing, and even if Persson and

[21] That is to say, one can commit to a particular view of rationality, and even the idea that more rationality will lead to more accurate moral views, without being committed to any particular substantive normative claims.

Savulescu's argument fails, the strong prima facie case for moral enhancement outlined in the first chapter remains.

One might worry instead that, even if there is a strong prima facie case for moral enhancement, it comes at too high a cost – it limits our freedom to be or act rightly or wrongly. This freedom to err critique comes from Harris (2011), who thinks such freedom is so crucial that, even in the face of the existential risks suggested by Persson and Savulescu, it must not be sacrificed. There would, on this view, be a strong right to do wrong – not only in the sense that people should be permitted to do certain wrong acts, but also that they have a moral right to act wrongly that would be infringed by enhancements that led people to be more moral. This right is exemplified through free speech laws (people are allowed to say deeply immoral things as long as their speech does not incite others to do harm) and permissions to engage in vices such as gambling and smoking (though these may be more plausibly understood as blameworthy rather than immoral).

Somewhat surprisingly, Douglas (2011) replies by conceding that direct moral enhancements may indeed violate people's freedom to be immoral, but a) that violation can be outweighed by the good brought about by moral enhancements and b) the objection leaves room for indirect moral enhancements that improve people's reasoning abilities (Harris argues for the freedom to be immoral, not the freedom to be irrational). But if the ability to act wrongly is valuable, shouldn't the freedom to act irrationally be valuable as well, implying the objection also has force against cognitive enhancement? After all, Harris's concern seems to be rooted in concern that moral enhancements will violate people's autonomy, which covers a much broader range of thought and action than morality alone. Conceding Harris's point, then, may prove too much, implying broad limitations on what sorts of enhancements are permissible.

However, we should be skeptical of the claim that moral enhancement runs afoul of the freedom to be immoral. In the first place, (non-coercive) direct moral manipulation does not obviously limit someone's freedom. Intuitions often arise or are (at least partially) caused by people's genes or their environments. In the face of such causation, we have two choices. One, we could be compatibilists and say that freedom is consistent with such external influences, in which case having another person (as opposed to one's genes or environment) significantly influence one's moral character should not make a difference to one's freedom. [22] Or, two, we could be incompatibilists and say that any external influences inhibit freedom; but then, the moral enhancer would only be swapping one external influence (themselves) for another (genes or environment). On either view, the freedom of the person (including the freedom to be immoral) being enhanced would be unaffected.

Moreover, Harris's conception of the right to do wrong is rather incoherent. On his view, it is somehow permissible or valuable to do that which is impermissible or of disvalue. The more plausible interpretation of the right to do wrong is that, in certain cases (such as non-threatening speech), people should be allowed to be immoral or act in a blameworthy fashion. This understanding of the right to do wrong rules out the permissibility of widespread coercive moral enhancement (which is quite reasonable) but leaves plenty of room for the permissibility and even governmental promotion of moral enhancement. The latter understanding of the right to do wrong is not only coherent but also quite compelling, and it should be favored over Harris's definition.

Another of Harris's arguments, though, is deeper and more troubling for proponents of moral enhancement. Harris writes, "[T]he sorts of traits or dispositions that seem to lead

[22] The compatibilist could insist that there is a morally salient difference between external influences and gene manipulation, such that the latter is autonomy-infringing while the former is not. In this case, moral enhancement could be said to be an improper violation of autonomy – but this point seems to go beyond freedom to be immoral, and into general autonomy concerns. There will be further discussion of issues related to autonomy (in particular, individuality) in Chapter 6.

to wickedness or immorality are also the very same ones required not only for virtue but for any sort of moral life at all." (Harris 2011, p. 104) This worry is all the more pressing in light of the empirical work reviewed above. Interventions tend to have ambiguous and controversial effects on the morality of people's mental states, motives and behavior. While Douglas (2011) interprets this as a problem of risk – any given moral manipulation might actually end up making people morally worse, not better – it is actually a much broader epistemic problem arising from deep disagreements about morality. In the next chapter, I will spend some time discussing how broad and relevant this disagreement can be; Chapters 5 and 6 will then explain why this disagreement poses a serious, perhaps insurmountable challenge to certain forms of moral enhancement.

# Part II: The Problems of Direct Moral Enhancement

With a clearer understanding of the nature and scope of moral enhancement, we can now turn to the negative part of this work: my arguments against direct moral enhancement. These arguments use the issue of moral disagreement to explicate the flaws of direct moral enhancement. Moral disagreement is pervasive and has a number of implications outlined in Chapter 4. Depending on one's metaethical, normative and practical views, one would recommend significantly different approaches to moral enhancement. Practically speaking, this means that (in a liberal democracy) it will be very difficult to get people on board with a particular, specified project of moral enhancement. Chapter 5 explains this worry, and also addresses the obvious response that we could simply seek out overlapping consensus on moral issues. Chapter 6 presents a more normatively-laden argument that we have reason to value disagreement on Millian grounds, and that this value is threatened by widespread direct moral enhancement. Such a project essentially suppresses dissent, making moral progress more difficult, subverting valuable reasoning processes and threatening people's individuality.

# Chapter 4: Moral Disagreement

Intense disagreement pervades most areas of moral discourse, including moral philosophy. This section will provide a broad survey of this disagreement. It almost goes without saying that there is indeed wide and deep moral disagreement. Still, it is important to see how this disagreement is not merely idle or theoretical; as we will see, these disputes have direct implications for how a project of direct moral enhancement might be carried out. Even if the arguments in Chapters 5 and 6 are found unconvincing, this section will provide a loose understanding for how various divergent normative commitments will lend support towards different modes of direct moral enhancement. These disagreements will provide a basic practical problem for direct moral enhancement: how to offer a generally-acceptable program moral improvement when divergent views at every level of morality lead to differing implications for moral enhancement. Disputes in metaethics, normative ethics and applied ethics will each be discussed, in turn.

**Metaethics**

There are a variety of disputes about the nature of morality and moral concepts, some of which will be relevant to moral enhancement. I will separately address the relevance of substantively neutral and non-neutral metaethical theories. Substantively neutral theories are those that do not entail any particular normative conclusions (or entail the rejection of such conclusions), while non-neutral theories do have at least some such implications.

Along the way, I will also discuss various forms of moral realism and anti-realism. Moral realism is the position that moral claims can be true and are at least sometimes true, independent of our opinions about morality. Moral anti-realism is the negation of that claim. Moral anti-realism encompasses a range of popular metaethical theories, including: expressivism, according to which moral claims express attitudes, not propositions (see, e.g., Ayer 1936 and Gibbard 1990); subjectivism, according to which the truth of moral claims depends on whether one's attitudes towards them (see, e.g., Hume, *The Skeptic*, esp. I.XVIII.8); and error theory, according to which moral claims, though truth-apt, systematically fail to correspond to true propositions (see, e.g., Mackie 1977). There is also a further view, called quasi-realism, put forward by Simon Blackburn (1984, 1993). Quasi-realism is expressivist, but seeks to accommodate the legitimacy or appropriateness of moral claims.

We must also distinguish between first-order moral claims and second-order moral claims. First-order moral claims are the province of normative and applied ethics; they are claims about what is good and bad, right and wrong. These will also sometimes be referred to as substantive claims. Second-order moral claims are the province of metaethics; they are claims about first-order moral claims – their nature and content, in particular. This distinction will be useful, but as we will see, some philosophers doubt that the line between the two is very thick.

*Non-neutral theories*

The mechanisms by which many non-neutral metaethical theories can have repercussions for moral enhancement are fairly straightforward. Insofar as non-neutral

theories have normative implications, those theories should imply something about what individual beliefs, motives or behaviors are good, and any disagreements over that theory would be of relevance to whether and how moral enhancements should be pursued. This can occur when a metaethical theory involves commitment to a particular normative theory. Perhaps the most prominent instance of such a theory would be G. E. Moore's Benthamite utilitarianism: the meaning of the term 'right' is to maximize happiness. (Moore 1903) This metaethical claim straightforwardly implies a form of utilitarianism, which if correct would have direct repercussions on how moral enhancement should be carried out. Yet, if one doubted Moore's understanding of rightness (as many do), these further claims would also be put into doubt – making a Moorean defense of utilitarian-based moral enhancement contingent on his controversial metaethics. Indeed, normative entailments from the meaning of moral terms are relatively unpopular among contemporary philosophers.

More strikingly, certain metaethical theories could entail first-order normative skepticism. First-order skepticism about morality, for present purposes, just refers to the idea that all first-order moral claims[23] are false. If such skepticism is correct, then the entire enterprise of moral enhancement is fundamentally mistaken. We would be aiming at something – moral improvement – that is ultimately impossible, like trying to find Shangri-La or travel faster than the speed of light. Any claims made about purportedly morally enhanced people would be mistaken, and there would be no (moral) point pursuing such enhancements at all. True, there would be nothing wrong with moral enhancement if skepticism were true – but there would also be no normative reason to pursue it. The rejection of moral skepticism, then, can be seen as a necessary condition for the justification of moral enhancement.

[23] Second-order skepticism is also a possible view (indeed, it is an implication of most anti-realist views), but the relevance of first-order skepticism is the subject of this section. The question of whether skepticism about second-order moral claims entails skepticism about first-order moral claims will be addressed later in this section.

Ronald Dworkin (1996) has argued that all metaethical claims are non-neutral, and furthermore that all anti-realist claims imply moral skepticism. If Dworkin is correct, then the whole range of anti-realist metaethicists are committed to a view that would ultimately undermine the prospects for moral enhancement.

Dworkin's strategy is basically to deny that there are any truly 'Archimedean' metaethical claims – claims about morality that are not themselves normative claims. This means that there are no truly 'second-order' moral claims; any allegedly second-order claims can be shown to simply reduce to first-order moral claims. He does this by arguing that a) any allegedly second-order metaethical claim can be reduced to a first-order moral claim or set of claims, and b) alternative interpretations of these claims are not plausible. (Dworkin 1996, pp. 97-112) The implication of Dworkin's argument is the idea that anti-realists, in particular, must deny a wide range of moral claims. Take the common example, 'abortion is wrong'. When expressivists say that claim is neither true nor false, but merely the expression of some attitude, they seem to be denying the truth proposition that there is something morally objectionable about abortion – and that denial (allegedly) amounts to a substantive claim. Expressivists would deny that this claim is truly substantive, since they deny that moral claims are truth-apt across the board; but Dworkin insists that, even if they are denying other things as well, the denial of the truth of impermissibility amounts to an endorsement of moral permissibility.[24] Similarly, subjectivists would be committed to the claim that abortion would not be wrong if no one thought it wrong (also a substantive claim), and error theorists would characterize the claim as, like all moral claims, simply wrong (again, substantive). Looked at another way, anti-realism necessitates the denial of the truth of certain moral statements, which in turn makes it impermissible (on pain of inconsistency) to hold certain

[24] Blackburn's quasi-realism can potentially avoid this worry. By design, quasi-realism accepts that claims about permissibility – though not strictly speaking true or false – are not mistakes. We could comfortably make them and discuss them with others. Quasi-realism may in this way be able to maintain substantive neutrality, and so would not have particular implications for moral enhancement.

substantive moral beliefs, which is itself a substantive implication of anti-realist views. (Fantl 2006)

Importantly, anti-realists tend to reject Dworkin's claim that their theories imply first-order skepticism; instead, they prefer to see their theories as substantively neutral, including concerning whether moral skepticism is correct.[25] On behalf of such neutrality, Paul Bloomfield has suggested Dworkin overlooked a way of interpreting certain metaethical claims, such as those of the form, 'there is a right answer as to whether x is (morally) bad'. One could reasonably interpret such statements as only asserting that there is some fact of the matter concerning x's goodness or badness, without committing oneself to whether x is good or bad. (Bloomfield 2009, p. 292) This appears to be neutral as to the substantive question, whether x is bad, but makes a distinct metaethical claim about moral realism. Appearances can be deceiving, however. Consider the denial of that claim: 'there is no right answer as to whether x is bad'. That denial does have substantive implications of the sort mentioned above: the claims that x is bad and x is good are both false. Such leads to the sort of permissiveness characteristic of moral skepticism – there would be nothing wrong with doing x.[26]

[25] Mackie is arguably an exception. Though he insisted on a distinction between first- and second-order moral claims, he believed that "ordinary moral judgments include a claim to objectivity, an assumption that there are objective values in just the sense in which I am concerned to deny this." (Mackie 1977, p. 35) Everyday first-order moral claims, then, also involve or assume false second-order moral claims, casting any given moral claim into doubt. This led Mackie not to full-throated nihilism, but to a sort of constructivism: "Morality is not to be discovered but to be made." (ibid, p. 106) This construction is bounded, in particular by constraints that a given moral principle be universalizeable. There is not space here to tease out the implications of Mackie's attempt to reconstruct morality, but such a reconstruction would doubtless have consequences for how a project of moral enhancement – perhaps the purest way we could 'make' morality – would be carried out.

[26] This dispute might turn on how one understands substantive permissibility. A purely negative understanding that amounts to the denial of the proposition that something is wrong or impermissible would bolster Dworkin's case. Denial of truth-aptness involves denial of truth, so the expressivist does seem committed to the fact that abortion is permissible in the purely negative sense. However, we could understand permissibility more positively – with 'permissible' being a sort of moral property on a par with other substantive terms like 'good or 'bad'. The expressivist would deny there are any true instances of such positive permissibility (such claims are not truth-apt), and so without contradiction deny both that abortion is permissible and impermissible (both permissibility and impermissibility imply non-existent moral properties).

Bloomfield alternately suggests that the claim, 'there is a right answer as to whether x is bad' could be interpreted more broadly as a tautology: either x is bad, or it is not, or there is no right answer. This would make the negation trivially false. (ibid, pp. 292-3) Yet, this interpretation faces two problems. Firstly, it has the air of paradox – it implies 'there is a right answer as to whether x is bad' is consistent with 'it is false that there is a right answer as to whether x is bad'. Secondly, if the claim really is a tautology, it seems unlikely to be an accurate characterization of what anti-realists are claiming. Tautologies are vacuously true; presumably, metaethicists in general do not take their theories to be vacuous but rather reveal something interesting about the nature of morality. So, even if the air of paradox could be dispelled, it is unlikely that anti-realists could rely on this sort of interpretation to defend the neutrality of their theories.

Jamie Drier has suggested an alternative theory that fares better against Dworkin. He proposes a version of subjectivism according to which moral facts are determined by the dispositions we have here and now. (Dreier 2002) Recall that subjectivism typically appears to imply the substantive conclusion that, if people did not have a certain negative attitude towards some moral proposition, it would not be wrong. Dreier elegantly avoids such substantive implications because his suggested theory is immune to counterfactuals: the fact that people in a different time or universe might have different moral dispositions has no impact on the truth of moral claims, as moral claims depend only on our current moral dispositions.

Presentist subjectivism succeeds in refuting Dworkin's sweeping claim that there are no distinctly metaethical theories.[27] However, does such a presentist subjectivism have any normative implications? True, it makes no claims about what dispositions people actually

[27] Though it does not address a softer claim, that most metaethical claims are reducible to substantive claims. Indeed, Dreier does not explicitly defend the soundness of presentist subjectivism, so Dworkin could still conclude that there are no plausible Archimedean theories – any plausible theory could be reduced to first-order moral claims.

hold and is in itself compatible with any normative claim someone might make. However, if we establish certain empirical facts about current moral views, substantive implications will become clear – the moral facts are just those that people currently hold. Thus, presentist subjectivism is not neutral when it comes to practical matters. In the case of moral enhancement, it would seem to recommend in favor of enshrining and reinforcing current moral views in people's psyche.[28]

From the preceding, we can see how disputes over non-neutral moral theories can influence the debate over moral enhancement. Non-neutral metaethical theories like Moore's might directly imply certain normative commitments that will determine what enhancements count as moral. Additionally, if anti-realist theories are non-neutral, as Dworkin maintains, then those who accept such theories are committed to a moral skepticism that will severely limit the prospects of moral enhancement – indeed, moral enhancement would be impossible.

*Neutral theories*

Suppose that Dworkin is wrong and there are indeed neutral metaethical theories – theories that do not imply moral skepticism. We can still draw several interesting lessons from disputes over these neutral theories which can shape moral enhancement – if not over what it means to be moral, then over the appropriate means to determine whether some intervention counts as a moral improvement.

One lesson is that differing theories will offer differing standards of normative justification. Even if competing metaethical theories do not have differing implications about

[28] This sort of enhancement program would, in particular, run afoul of the problem of moral stagnation discussed in Chapter 6.

particular normative conclusions, they may have differing implications as to how normative claims are to be justified. (Ehrenberg 2008) While realists will appeal to the moral reasons that there are, expressivists will appeal to their attitudes, subjectivists to personal dispositions and error theorists might deny that proper justification is even possible.[29] The position one takes on metaethics, then, will distinctly inform the proper way to justify one course of moral enhancement over another – whether one should appeal to objective facts, subjective opinions, or something else.

Aside from justification, different metaethical theories make somewhat different descriptive claims about our moral psychology. Expressivists, in particular, are committed to a non-cognitivist (that is to say, non-propositional) account of moral judgments. The truth or falsity of non-cognitivism will be crucial in figuring out how to target certain moral enhancements. When trying to improve people's moral ideas, do we try and alter neural processes associated with belief-states, or more affective neural centers? The non-cognitivist would likely favor the latter, while the cognitivist would choose the former (at least as an end-state; they could agree that affects will affect judgments).

Also, realists can more easily be fallibilists about people's moral beliefs, motives and actions than expressivists or subjectivists (though not necessarily error theorists, who also embrace fallibilism), making claims that certain people's moral beliefs or propositions are mistaken. Expressivists deny the truth-aptness of moral claims, making the idea of 'moral mistakes' mysterious, while subjectivists have defined morality in such a way that one can never be wrong, from one's own point of view (except, perhaps, in cases of self-deception).

[29] The metaethicist might attempt to be neutral about justification as well, claiming instead to be embarking on a purely descriptive project that is compatible with any forms of justification. However, metaethicists tend to offer a more revisionary approach, in which a given metaethical view will influence the appropriate way to ground moral claims. Some justificatory stories – those that rely on purportedly objective, mind-independent facts – will simply not be tenable due to their realist metaethical undertones. Blackburn's quasi-realism is meant to avoid this sort of revision and allow that moral claims are not mistakes, but even under his account we need to be careful about how such claims are structured and grounded.

Fallibilism is also one plausible justification for moral enhancement: we make certain systematic moral mistakes, and enhancement can help overcome those mistakes. But if expressivists or subjectivists are right, such fallibilism will be difficult to defend, and one motivation for moral enhancement undercut. Of course, expressivists and subjectivists can condemn certain moral dispositions as wrong-headed, and perhaps motivate enhancement on those grounds, but making general moral claims about such dispositions will come into conflict with their antirealism. (see Sturgeon 1986, p. 140,n. 43)[30]

Accepting or rejecting subjectivism, in particular, would have an enormous impact on how moral enhancement would be carried out. While subjectivism may be compatible with almost any particular claims, it takes the truth of those claims to be contingent upon people's actual dispositions. This will either lead to moral enhancement of beliefs or motives being impossible. If moral truth is contingent upon with whatever moral dispositions people have, then almost any revisions to the moral character of beliefs or motives will count as good, at least from the enhanced person's point of view. If someone at time t1 thinks that murder is permissible, then at time t2 she is altered to think that murder is wrong, she would in a sense be correct at both instances (again, from her point of view). But then, moral 'enhancement' is impossible – any alteration will simply change a person's moral attitudes, not their actual moral status. Dreier's (2002) presentist subjectivism even makes such moral alterations impermissible; people's present moral dispositions and motives are necessarily correct, and any changes would be a step in the wrong direction, towards immorality. As mentioned above, such a theory actually might recommend preventing future dissension from present

[30] Blackburn (1993) has tried to account for fallibilism by interpreting judgments of fault as attitudes about what moral sensibilities it might be better or worse (in terms of whether it would elicit more or less approval) to have; though as Sturgeon (1986, pp. 130-1) has pointed out, such fallibilism ends up being difficult to reconcile with consequentialism, where mistakes are grounded in good or bad consequences rather than (only) people's attitudes. This is not so much an objection to Blackburn's view as an observation that his view makes consequentialism more difficult to defend – which would in turn speak against an enhancement program based around consequentialist principles.

moral attitudes and 'locking in' society to our current moral beliefs. The problems associated with such interventions will be discussed further in Chapter 6.

Arguably, moral enhancement of actions would still be possible for subjectivists if moral beliefs stayed constant from time t1 to t2 but the agent acted more in line with her moral beliefs at time t2 (essentially reducing akrasia). Still, this would limit moral enhancement to a very small set of psychological dispositions. If subjectivism is correct, moral enhancement would in effect be a rather limited, conservative enterprise, not actually changing the way people think about morality but just making their beliefs and actions more consistent.

Moral disagreement itself will also be informed by one's metaethics. Much more will be said about the implications of such disagreement later, but for now it is enough to note that differing theories will present differing accounts both of whether and how moral disagreement is possible. Realists have a straightforward account of such disagreement, according to which people are just disputing the facts (one side asserts a truth-apt proposition, the other denies it); error theorists similarly will think people are disputing facts, but (almost always) both sides are wrong. Subjectivists and expressivists, however, must either deny disagreement exists ('I think x is wrong' does not refute someone else saying 'I think x is right') or give a nuanced, nonstandard account of disagreement. Which account of disagreement one adopts will help determine how disputes over moral enhancement are to be characterized and, ultimately, adjudicated; subjectivists and expressivists may be more inclined to take an 'agree to disagree' line and therefore let people enhance as they please, while realists might gravitate more towards the need for general agreement and consensus before proceeding.

## Normative ethics

Normative theories are systematic attempts to characterize the grounds for and conditions of moral character and behavior. It should be unsurprising that disputes over normative theories would have significant implications for the proper way to perform moral enhancement; while such theories might not always have clear entailments when it comes to particular moral dilemmas, they at least give general characterizations of morality that could significantly inform the direction of moral enhancement. Three normative theories will be considered here: consequentialism, Kantian deontology, and virtue ethics. These are not of course exhaustive of normative thought, but they are representative of the central disputes in the literature.

### *Consequentialism*

Consequentialism has been succinctly defined by Samuel Scheffler as "a moral doctrine which says that the right act in any given situation is the one that will produce the best overall outcome, as judged from an impersonal standpoint which gives equal weight to the interests of everyone." (1988, p.1) This highlights two features of consequentialism relevant for moral enhancement: teleology and agent-neutrality. I will discuss the implications of each feature in turn, then note some common grounds for disputing consequentialist theories.

Teleology defines right actions in terms of their capacity to bring about good outcomes or states of affairs; actions are right just insofar as they produce the best outcomes.

This is, according to teleology, the only way in which actions can be right – it entails a denial of other normative theories of rightness. Teleology is itself neutral about what the good is; more particular consequentialist theories like utilitarianism fill that in. Thus, while consequentialists tend to concentrate on external consequences for actions, it is actually consistent with there being (dis)value in actions themselves, or the mental states behind them. Teleologists will take goodness to at least be rankable: some states of affairs are better than others.

The fact that consequentialist teleology is itself neutral as to the content of the good makes its intrinsic implications for moral enhancement somewhat limited but nevertheless discernable. As an exclusive normative claim, it implies a rejection of certain other theories discussed below, which is significant in itself. In addition, the necessity of ranking states of affairs (in terms of whatever the good is) indicates that improving people's 'maximization' tendencies – inclinations to sum up the morally relevant effects of various options and weigh the results against one another – would count as an enhancement, if a blunt and imperfect one. However, combining any given particular theory of the good with consequentialism will have more obvious implications: moral enhancement would involve making people more likely to take actions to maximize that good. If certain moral beliefs and motives facilitate such maximization, or one's theory of the good includes the intrinsic value of certain beliefs or motives, then instilling or strengthening them would also be a robust form of moral enhancement.[31]

Utilitarianism (which holds that utility is the good) is the most famous version of consequentialism, and it would have particular implications for moral enhancement. Persson

[31] Consequentialists could resist the importance of such maximization tendencies; perhaps the best way to maximize the good is to impose non-maximizing dispositions. Such views are sometimes deployed by consequentialists to avoid embarrassing conclusions about, e.g., sacrificing the one to save the many. However, the fact that consequentialists (at least the ones we know about!) are honest about their position and argue that others should adopt their position indicates that consequentialists tend not to put much stock in the benefits of promoting non-consequentialist worldviews.

and Savulescu's (2008) proposal gives a rough approximation of what utilitarian moral enhancement would look like: changing people's dispositions such that they produce the most utility for the population at large. This will likely involve inducing people to support involve greater public cooperation, more resources geared towards improving welfare, and perhaps even making people more likely to feel happy with the same amount of resources (engineering what Nozick (1974) calls utility monsters). As with other forms of consequentialism, it need not involve inducing people to actually be utilitarians; if utility can be best brought about if no one believes in utilitarianism, then inducing people to reject utilitarianism might in fact be a utilitarian moral enhancement.

Agent-neutrality's implications are more direct than those of teleology. To say that morality is agent-neutral, as consequentialists do, is to say that the rightness of actions does not depend on one's viewpoint; agent-neutrality is to be contrasted with agent-relativity, where the rightness of actions is crucially dependent upon the agent's standpoint. It might be that while, from an agent-neutral standpoint, it is better that one person is murdered than two, it might be wrong from an agent-relative standpoint for a given agent to commit one murder in order to prevent two more – even if preventing the two deaths is clearly a better state of affairs. Consequentialism, as an agent-neutral theory, more easily justifies 'sacrificing' the one to save the many. Agent-neutrality also implies that, in general, one cannot morally justify privileging one's own interests over the interests of others. So, for instance, when choosing whether to improve one's own lot or the lot of two others by an equal amount, consequentialists would generally require someone to improve the lot of the two others; the fact that one's own interests are at stake (as opposed to someone else's) is not in itself morally relevant.

We can discern two possible implications of consequentialism's agent-neutrality for moral enhancement. One, agents should perhaps be made more willing to sacrifice the few in

order to save the many. Qualms about pushing the person off the bridge in the trolley cases discussed in the previous chapter may have to be suppressed, at least when the public good is at stake. Two, agents should be made to weigh their own relative interests about the same as the interests of others. This directly implies agents should be more altruistic, but may also imply that we should also put less emphasis on special relationships (friends, family, country, etc.) that lead to an agent-relative emphasis on the interests of the few as opposed to the many. Admittedly, not all consequentialist theories will have these implications – in particular, they can be resisted by those that claim the preservation of one's own interests or the interests of those close to you as an intrinsic good. Still, that resistance will only go so far – such values will often be in tension with the general requirement that the goodness of states of affairs be maximized for all agents, not just oneself or close relations.

As with the other theories, it is not possible to discuss all the objections to consequentialism here. Instead, I will just mention two of the more prominent objections, to give a flavor for the disagreement over consequentialism's soundness. One critique focuses on consequentialism's aggregationist tendencies. The value of one life can be stacked up and weighed against another, much like piles of gold. This is said to not respect the separateness of persons – it treats producing the best state of affairs within one person in the same way it treats producing the best state of affairs between two people. This leads to cases of impermissible sacrifices and disrespectful usage of people. (see, e.g., Rawls 1988, Nozick 1974, Scanlon 1998) Interestingly, this objection suggests a converse sort of moral enhancement – inhibiting people's maximizing tendencies when faced with moral quandaries, somewhat in line with the sort of Kantian deontology discussed in the next section.

The second critique is that consequentialist theories – utilitarianism in particular – demands too much. The lack of agent-relative permissions within consequentialism has somewhat extreme consequences: one must spend one's resources not for oneself, but in a

way that is best for the whole world. Some consequentialists such as Peter Singer (1972) have embraced these implications, including the fact that the well-off are required to give away most of their resources to help those worst-off, until the point that their giving resources away makes them worse off than those they are helping. Such theories suggest an extreme form of moral enhancement, a radical altruism where people's inclinations to hold onto their own money is drastically suppressed. Others such as Samuel Scheffler (1994) find these implications disturbing and posit, on grounds of personal integrity, that people can sometimes permissibly fail to live up to the demands of consequentialism. Similar objections could be raised against an extreme altruism enhancement – that it led to excessive self-abnegation and the loss of personal integrity.

*Kantian deontology*

Kantian moral philosophy is deep and complex, and only an approximate sketch can be offered here. Kant's account of morality can be called deontological, insofar as it does not derive rightness from what brings about the best state of affairs but instead from absolute rules – rules that follow from the formulations of what Kant called the categorical imperative. This section will focus on Kant's first two formulations from the *Groundwork of the Metaphysics of Morals*, the formulation of universal law and the formula of humanity, and draw out some of their implications for moral enhancement.

The formula of universal law is as follows: "Act only according to that maxim through which you can at the same time will that it become a universal law." (IV:421) Though Kant ultimately justified this based on a certain understanding of people as rational, free agents capable of contemplation and assessment of theirs and others' autonomy, it also

has intuitive appeal by being similar to the golden rule – do unto others as you would have them do unto you (though Kant rejected this analogy). The formula entails a certain method for determining whether some proposed plan of action is morally acceptable: consider whether, if everyone to systematically acted according to the same ruleset that permits your plan of action, you could rationally will such an action. If such universal maxims are indeed willable, then the action is permissible; if not, it is forbidden. This will, for instance, lead to the requirement – and note, it is an absolute requirement – to pay back loans because a world in which one was permitted to not pay back loans is one where loans would not be issued in the first place, making it impossible for you to coherently will non-loan-repayment to be a universal maxim. Other prohibitions, such as against lying or suicide, are meant to follow.

One possible route for Kantian moral enhancement, then, could be to instill this particular decision procedure – following the formula of universal law – into people's moral behaviors. While it is not clear Kant's metaphysics would have admitted this is even possible (Kant conceived of our wills as uncaused causers), more modern Kantians (e.g., Hill 1989) would be more amenable to such amendments. This may seem to over-intellectualize moral decision-making, but Kantians would have strong reasons to promote such enhancement – Kantian deontology not only requires people to follow the rules, but follow them for the right reasons, to wit, an understanding of the universal law. By instilling this reasoning procedure, people would not only get the right result, but have the right motives for action.

Two other sorts of enhancements could follow from an endorsement of the formula of the universal law. One, people should be made to be more inclined towards moral absolutism. Kantian moral laws are absolute and exceptionless; there is never any circumstance in which it permissible to lie, cheat, etc. Common-sense morality is more pliable than this; we often admit of exceptions and extenuating circumstances, and perhaps these are signs of a certain moral weakness. By making people more amenable to absolutism,

they should be quicker to accept Kantian formulations and therefore act correctly. And two, people should be made more honest. This is a particular normative implication of the formula of universal law – perhaps the clearest – and it would seem everyone would be better off if they simply ceased to lie, at all times. True, this sort of direct alteration of inclinations to lie might not get people to tell the truth for the right reasons (due to consideration of the formula of universal law), but acting in the correct way should be a strong second-best.

Let us now turn to the other formulation to be considered here. The formula of humanity takes the following form: "Act that you use humanity, in your own person as well as in the person of any other, always at the same time as an end, never merely as a means." (IV:429) This can be taken as an expression of respect for persons and their intrinsic, final value; a similar sentiment underlies some of the anti-aggregationist arguments mentioned in the previous subsection. This formula is absolute like the formula of universal law, but it has somewhat different direct implications. It prohibits acts such as theft and murder on the grounds that those acts do not involve treating people as ends, that is, as the autonomous beings that they are; instead, thieves and murderers use others purely as means to some personal end. This does not prohibit all interactions done for personal profit, of course; one can engage in economic exchanges for self-interested reasons, so long as, in doing so, one evinces respect for others' humanity (by, for instance, not cheating them).

Again, Kantians have reason to support moral enhancement programs that would instill this formula in people. Targeting 'respect for others' may be difficult – the notion of respect is difficult to grasp even conceptually, let alone through an enhancement intervention – but making people value other people more might have this effect. Easier might be enhancements that reflect the various prohibitions the formula of humanity entails. Various dispositions to steal, cause violence, or ignore other people's interests could be suppressed. In trolley cases, this would perhaps mean inhibiting support for pushing the person off the

bridge to save the many – in essence, making people more averse to harm others. Of course, such harm aversion would have to be tempered, as Kantians (like others) would want people to come to the defense of others, when necessary. Exactly how much to temper such inclinations is a difficult puzzle, one which relates to the central difficulties facing moral enhancement, to be discussed below.

The implications of Kant's formulas have been deemed unacceptable to some. Famously, Kant admitted to one of his contemporaries that his theory implied you should not deceive a vicious murderer at your door about the whereabouts of his innocent target, even if that is the only way to save the innocent's life. This will strike many as implausible; it shows the weakness of absolutist principles, in that there will inevitably be unpleasant extremes. Some Kantians, like utilitarians, are willing to accept their theory's dubious implications, but to others the implications cast serious doubt on the wisdom of such absoluteness. This serves to underline one of the central disputes between deontologists and consequentialists: are values to be respected, or promoted? This debate cannot be adjudicated here, but it goes deep and, even after hundreds of years of dispute, does not appear to be near resolution.

*Virtue ethics*

Virtue ethics, introduced by Aristotle in the *Nicomachean Ethics* and revived in the 20th century, can serve as a counterweight to the extremes of certain forms of Kantian deontology and consequentialism (though it is arguably more in line with non-Kantian forms of deontology than consequentialism). It is also superficially more amenable to enhancement, as virtue ethics shifts the focus from general states of affairs and actions to character and traits, the very targets of many potential enhancements. Virtues are not just

good traits, however, somewhat complicating efforts for moral enhancement within virtue ethics. Chapter 2 discussed the typical role-model approach to moral education endorsed by virtue ethicists.

Chapter 2 briefly examined some of the problems for virtue ethics and moral enhancement, in particular the fact that part of being virtuous is coming to virtuous action via one's own practical wisdom, not just some brute (perhaps inculcated) emotions. One way around this is to enhance practical wisdom by improving reasoning ability. This would be a general effect, one that would not zero in on one particular virtue but should (in theory) make someone more able to contemplate and analyze morally relevant facts as well as draw the proper inferences from there to action. Such enhancements would not be sufficient for practical wisdom, of course; people have to be committed, to a certain extent, to employing their reasoning facilities for virtuous purposes. But enhancing people's rationality could at least help those so committed. This approach would be in line with the reasoning approach defended in Chapter 9.

The regulation of instincts and emotions could possibly be another way to help those committed to virtue, but who might need some assistance. The clearest way would be to resolve pathologies – irrational fears, compulsive transgressions, antisocial impulses, and so on. Those traits are not sensitive to the relevant facts and will most likely inhibit any attempts to become reliable and virtuous in one's behavior. Similar results might come from the regulation of non-pathological tendencies in those who nevertheless exhibit vicious behavior, but this will be more controversial. It is just that sort of emotional regulation that Jotterand (2011) is concerned with; indeed, such does run the risk of substituting external influences for personal wisdom, thus inhibiting the emergence of true virtue.

One difficulty with virtue ethics is that the theory itself can give relatively little guidance as to what the appropriate mean for a given virtue is. This is to be contrasted with particular deontological or consequentialist theories that at least attempt specific prescriptions. The theory does suggest how one might become virtuous, but the precise balance between competing poles of a virtue remains somewhat obscure. The theory can sometimes seem to simply recommend, 'you'll know it when you see it' – but only if you are virtuous. This is perhaps not too surprising – virtue ethics is meant to be a guide to life, not a guide to action. But that makes discerning how to morally enhance virtue rather difficult.

A similar problem can be said to beset moral enhancement in general. What is the precise, morally ideal level of any given hormone or neurotransmitter? How much altruism is too much altruism? What traits are good, and what are bad? Some might purport to have an answer to these sorts of questions, but even within normative theory the answers are very elusive. The difficulty of these problems will spell trouble for moral enhancement.

**Applied ethics**

The preceding positions, perhaps due to their theoretical nature, have primarily been disputed among philosophers. However, most everyday moral disputes concern more practical matters. These disputes can, at times, seem intractable, and that is not due to a lack of careful consideration; philosophers have weighed in on such issues and often seem no closer to resolution than the general population. I will consider only a small set of these disputes, in bioethics and politics, but the same general considerations should apply to most moral issues that are the subject of public debate.

*Bioethics*

First and foremost of bioethical disputes is the abortion debate. On the pro-choice side, there are two primary forms of argument: one, fetuses do not have a right to life because such a right requires certain (psychological) capacities they lack (in other words, fetuses are not people; see, e.g., McMahan 2002); two, women have a right to the use of their body that trumps the interests of fetuses. (Thomson 1971) The opposing pro-life side tends to emphasize the strength of fetuses' right to life, sometimes by appeal to the basic wrongness of killing living humans (Lee and George 2005), and other times by appeal to the fetuses' potential to grow into people with the relevant psychological capacities. (Marquis 1989) While the debate has been going on for decades, it shows no signs of slowing down or becoming resolved.

Another dispute arises over the end of life – whether euthanasia of varying sorts should be permitted. This dispute can sometimes be focus on a distinction between killing and letting die. Given many seem to accept that the terminally ill can permissibly decline life-saving treatment, does consistency demand they also could permissibly request and be given life-ending interventions? (Rachels 1975, Sullivan 1977) Others are more concerned to debate whether having doctors perform such procedures violates their professional role as healers. (Emanuel 1997) Disputants must be sensitive to the distinction between voluntary and involuntary euthanasia, and active and passive euthanasia, though there is not space here to get into the intricacies of these differences.

With these disputes in mind, those on various sides might promote moral enhancement in the name of their cause. A pro-life activist might hold that inducing people to oppose abortion counts as an enhancement (which a pro-choice activists would vigorously

dispute), just as a proponent of euthanasia might want to induce people to support it (again, much to the chagrin of their opponents). Still, the idea of moral enhancement working by simply instilling beliefs is somewhat far-fetched. In the previous chapter, current research indicates current methods of modulating moral beliefs, motives and actions affect broad swaths of the human psyche. Perhaps future research will narrow those domains down, but it seems unlikely they will be so refined as to affect such particular propositions.

Instead, disputants might seek to identify core values that support each of their positions, which would be more amenable to enhancement. Interestingly, both the euthanasia and abortion debates share a common (if vague) set of values that appear to be in dispute. Pro-life and anti-euthanasia positions could be seen as promoting the intrinsic, absolute value of human life; this suggests both an inclusive stance as to what humans have the right to life, and a deontological prohibition against killing. They might then consider enhancing strict aversion to killing other humans, no matter what stage in life and what they request. Conversely, supporters of abortion and euthanasia might want to enhance people's commitment to freedom, personal choice, and the value of autonomy.

The upshot is that opposing sides on hot-button issues like abortion and euthanasia will not only disagree about the issue at hand, but in light of that disagreement, may dispute what manipulations are moral improvements (insofar as they make people support their causes) and what are morally odious (insofar as they make people oppose their causes).

*Politics*

Political party and ideology infects a large amount of modern public discourse, including the above bioethical debates; few policy issues avoid inevitable partisanship, and the demand for political unity may well see as its next frontier a program of moral enhancement.

The liberal-conservative dichotomy is the most prominent and recognizable feature of political discourse. Roughly, we can understand liberalism as support for greater social freedoms and tolerance, coupled with economic programs that aim to assist the worse off in society, promote equality (particularly through redistribution) and impose economic regulations to prevent the harms of the free market. Conservatism, by contrast, involves support greater social unity and adherence (particularly to tradition) as well as confidence in the morality and efficiency of the free market. A wide range of policy disputes follow from these basic characterizations, but for now let us consider them at this broad level. Each side would generally count making someone more in line with their own ideology as a moral enhancement.

At first glance, there is a certain tension within both liberalism and conservatism that threatens the coherence of inducing liberalism or conservatism. Liberal parties tend to believe in free social policies and controlled economies, while conservative parties tend to believe in the opposite. Neither consistently endorse the impulse to let people be free or to control them, and any broad-brush manipulation would have unclear results on the ideological spectrum. By making people more supportive of control, they would become more liberal on economic issues but more conservative on social ones, and vice-versa for freedom. Now, libertarians, who support both economic and social freedoms, can take advantage of the single-mindedness of their ideology (support for individual freedom) and

promote enhancements of that sort, likely assured it will lead more to support their cause. But more mainstream liberals and conservatives must take a more nuanced approach.

Still, such a nuanced approach should be feasible. Economic issues have a distinct enough character from social issues that it may be possible to isolate the different sorts of manipulable dispositions. Inclinations towards tolerance or a reduction of in-group bias, for instance, is mostly only relevant for social concerns, and could be part of a liberal program of enhancement. Perhaps one would have to go issue-by-issue to discern all of the politically relevant dispositions. This may take time, but should not be out of the reach of researchers, perhaps even in the near future.

**Section summary**

This section has explained in some detail the sorts of moral disagreements that pervade moral discourse. This disagreement, as has been seen, is both deep and broad, covering issues both practical and theoretical. Moreover, these moral disagreements entail differences in the sorts of enhancement programs that would count as moral. Kantians and consequentialists, realists and anti-realists, liberals and conservatives – their moral disputes inevitably lead to disputes about the proper way to go about moral enhancement. The next two chapters will discuss what this pervasive disagreement means for moral enhancement. However, even if those chapters are ultimately unconvincing, I hope the preceding has at least been an illuminating discussion of how those committed to various views might be inclined to engage in moral enhancement. This could provide a road map to the promise and perils of future interventions as research in moral enhancement progresses.

# Chapter 5: Disagreement and the Feasibility of Moral Enhancement

## Verifiability

It is no coincidence that there is disagreement over a diverse range of topics in morality; there would seem to be something about moral thought that makes such agreement difficult. But what might be the reason for such disagreement?

One explanation might be the way people are committed to particular values. These values form a core part of people's lives; group membership and identity are often built around a shared set of moral sensibilities. People are unwilling to give up moral beliefs because doing so might alienate them from their community or their sense of self. (see, e.g., Simpson 2012) Empirical beliefs tend not to be so central to people's lives – though, notably, sometimes they are. Historical Catholic opposition to heliocentrism might be an example of similar empirical intransigence, as perhaps is modern-day doubts about global warming. If the analogy holds, then proponents of moral enhancement have little to worry about; just as intransigent opposition to heliocentrism and global warming models are were not reasons to give up on astronomy or climate science, moral enhancement can persevere in the face of such (arguably irrational) opposition.

However, the analogy is not especially strong. While people do indeed have many ex ante commitments to various values, moral disagreement goes beyond such loyalty.

Disagreements persist in areas far outside what might be called people's moral identity, in areas of metaethics or normative theory that are rarely considered by non-philosophers. As we have seen, some areas of ethics are politicized, but even without such partisanship about hot-button issues it is doubtless that widespread moral disagreement would persist.

Instead, the relative disagreement over moral matters appears to be due to the non-empirical nature of normative claims. Empirical facts (at least most everyday facts) are subject to a sort of widely-accepted system of independent verification not available to moral claims. This is roughly a general trust in sensory observations combined with inductive reasoning; it is exemplified by the scientific method and is the grounds by which we can be confident in the opinions of experts. (McGrath 2008) There may be some skeptics of such empirical reasoning, but they are few and far between (and, if consistent, would probably not last long due to a tendency to fall into clearly-marked pits).

Moral claims, by contrast, lack just such a widely-accepted method of verification. Some such methods have been proposed, either as metaethical or normative theories, but as we have seen there is little agreement over them. There are, to be sure, some widely-accepted constraints on moral claims, primarily revolving around logical consistency and coherence. Most everyone would agree it a mistake to believe that some particular action is both immoral and (in the same sense) not immoral. Similarly, someone would be making a mistake to both hold that all murder is wrong as well as claim that a specific instance of murder was not wrong. So, there is at least some independent check on (a set of) moral claims.

However, one cannot justify moral claims based purely on such uncontroversial requirements of consistency. Some further, inevitably controversial premises will be needed. Empirical investigations have the benefit of sensory observations and inductive inference.

However, such inputs are not sufficient for most sorts of moral claims, particularly not the moral claims relevant for moral enhancement. One must also supply further reasons to accept a given moral claim, and these further reasons typically take the form of moral intuitions (roughly defined as basic, non-inferred, self-evident moral beliefs).

These non-verifiable moral intuitions appear to be often at the crux of moral disputes. Debates over metaethics often come down to what seems to various interlocutors to be the best understanding of morality or moral terms, normative theories turn on differing basic ideas about what the good consists in, and applied ethics can frequently reduce to basic gut reactions concerning the morality of particular cases. Focusing on intuitions can sometimes be frustrating, as they are not easily amenable to argumentation or alteration, but many disputes do ultimately seem to just boil down to competing intuitions, and there is no commonly-accepted standard by which to evaluate those intuitions.[32] This leads to the sort of intractable debates characteristic of the previous section.

**What to Do When People Disagree**

From an objective perspective, that sort of intractable disagreement is more or less irrelevant. It may be that people disagree vehemently about morality, but (if we are realists or quasi-realists), there will be some correct view. Moral enhancement should just be in line

[32] Sidgwick (1907) famously proposed several criteria by which to evaluate one's own intuitions. However, they do not adequately explain how to adjudicate between different interlocutors' opinions. The third criterion (internal consistency) is merely logical and won't help explain which (among competing inconsistent intrusions) to give up. The first two refer to clarity and reflection, which are important but will be of little use in engagement with others whose minds one cannot read. And the fourth, downgrading certainty in the face of disagreement, is both controversial and unhelpful – how much should one downgrade? Are there circumstances where no downgrade is needed because someone is crazy? Does one have to downgrade if one's interlocutor is contradicting themselves? Ultimately, all of Sidgwick's criteria serve as grounds for tinkering with one's confidence in intuitions, but not a full standard by which all intuitions can ultimately be measured against one another to determine which are correct.

with whatever that correct view is. However, this provides little insight into how to establish and evaluate a proposed program of moral enhancement – just as the directive 'do the right thing' is sensible yet practically useless. The more common and action-guiding course in the face of disagreement is to simply stick to one's guns and act on one's considered views of what is right. Moral enhancement, on this line of thinking, would simply follow whatever the considered moral views of the enhancer are, in ways suggested in the previous chapter.

There are some theoretical issues with just sticking to one's guns. Most obviously, disagreement should prompt people to reflect on their own views and consider the potential flaws that others might point out. Beyond this, some have suggested that, in the face of disagreement with your epistemic peers (people just as reliable as you are at getting the right answer), you should give equal weight to your and your interlocutor's views. After all, if you really are peers, you should each be just as likely to get the right answer, and the reasonable thing to do is split the difference between the two positions. (Elga 2007, Christensen 2007) But there are ways around this problem. One could insist that it is perfectly acceptable to use disagreement as grounds for discounting someone else's reliability, allowing one's own considered judgments to stand even in the face of disagreement. (Enoch 2010) Or, it could be emphasized that epistemic peerhood is in general quite rare, making the issue unlikely to significantly undermine the prospects for moral enhancement. (King 2012) [33]

A broader problem has already been suggested in Chapters 2 and 3: many direct moral enhancement programs will operate in circumstances where the enhancer and the enhancee disagree about the moral matter at hand. This is especially predominant in moral education and propaganda, but also could easily emerge in the newer forms of biological enhancement. Indeed, it is not too difficult to imagine the same proponents of traditional

[33] Later on in Chapter 8, I will endorse a more limited implication of epistemic peerhood – accepting others' basic moral intuitions as equally likely to be correct as one's own. This does not imply, however, that one needs to substantially revise one's moral views in the face of peer disagreement – one may have reason to doubt others' reports of their internal intuitions, or various factors influencing non-basic intuitions.

enhancement programs being attracted to employing pharmaceutical aids to make their task easier. As noted in earlier chapters, this sort of enhancement treats the enhancer as in a morally privileged situation. But why should we take the enhancer to be in this privileged position? In principle, there is nothing distinguishing the enhancer and the enhancee other than the power of one to alter the ideas, motives or behavior of the other. Yet that power is in itself morally irrelevant. Without some further argument, we would have no reason to think one party or the other is more likely to get things right. The risk, then, is that enhancers will go in the wrong direction.

It may appear that some of the same issues arise when one is just trying to determine what are the correct moral ideas, motives or actions. Many moral ideas, motives and actions are generated in the face of disagreement; to do otherwise would invite the charge of skepticism, which I hope to avoid here. Doing so privileges oneself in a certain way; one is putting one's own ideas ahead of those of others. This is even the case when one adopts the sort of equal weight view discussed above; even if one becomes ambivalent (adopting, say, a 50% credence) on an issue when faced with peer disagreement, there is a disagreement between oneself and one's interlocutor (assuming he or she does not also endorse the equal weight view). Conceding entirely to one's interlocutor may not even avoid the issue, insofar as many (if not most) moral disputes there will be multiple adherents on either side of an issue. No matter what one does, some moral position will have to be put ahead of another.

Still, moral enhancement is different from those ordinary cases of disagreement. With most moral issues, the issue is simply how to think, be or act. With moral enhancement, the question becomes how to make others think, be or act. In the latter sort of cases, privileging the enhancer over the enhancee becomes much more problematic because the enhancer proposes not only that he or she is right, but that the other party be made to conform to the view of the enhancer. In some forms of moral enhancement such as punishment, where the

primary mechanism is entirely external and has substantial independent support (e.g., retribution and containment of dangerous individuals), this is not so objectionable. But many other forms, such as propaganda and pharmaceutical interventions, the mechanism is primarily mental, where the enhancer attempts to alter how a person's mind operates. By seeking to bring someone's mind into line with the enhancer's (or how the enhancer thinks the mind should operate), the enhancer in a certain sense dominates that person. The position of privilege is not merely "Think what you want, but I'm right." Rather, the enhancer is effectively saying, "I'm right and you're going to think I'm right one way or another." This is to treat the enhancee disrespectfully, and can in extreme forms deny people the freedom of thought.[34] The enhancer might try to justify his or her action on the grounds that the enhancee's views are themselves disrespectful (in virtue of being morally incorrect). However, disrespectful people still deserve to be treated with respect – they don't lose their moral status just because they disagree with the enhancer.[35]

A further issue to consider is that disagreement poses some practical problems for any program of moral enhancement. Even if the enhancers are convinced that they are correct, as long as there is a basic reluctance among those they wish to enhance their efforts will be difficult. For traditional forms of moral enhancement like education and propaganda, this means that the targets of the enhancement will be relatively resistant to the various entreaties that are offered. Novel biomedical interventions might circumvent such resistance, but only at the greater cost of suspicion at the degree of internal invasiveness that would be required with such techniques. People may be able to get on board with invasive interventions when the outcome is indisputably a benefit, such as the disease prevention afforded by vaccines.

[34] A related point concerning the value of individuality will be discussed in the next chapter. There is a difference, however; here, the worry is about the enhancer's respect towards the enhancee and the problematic relationship between the two. The next chapter will deal more with the importance of preserving disagreement itself.

[35] There may be cases where otherwise-disrespectful treatment is justified, as in the case of punishing criminals. The issue of punishment will be addressed again in later chapters. But it would be problematic to extend those justifications to the population at large – it would be to treat everyone the enhancer disagrees with as a criminal.

But in an area like morality, where we can expect little agreement over the wholesomeness of the outcome, general backing of a particular intervention will be incredibly difficult.

**The Overlapping Consensus Response**

The above argument against the feasibility of enhancement of moral beliefs or motives relies on the fact that there is pervasive moral disagreement. I will now take some time to consider perhaps the most obvious reply: while there are indeed many areas of disagreement, there are also areas of agreement. If moral enhancement focused on those areas of agreement, one could argue, the problem outlined above can be avoided. An even more modest approach would be to limit enhancements to only cases where the enhancer and the enhancee agree that some change would be a moral improvement.[36] While this response suggests how the problem of disagreement could be overcome in theory, overlapping moral consensus, I shall suggest, will ultimately offer little guidance for prospective moral enhancements.

*Forms of overlapping consensus*

[36] This form of enhancement comes quite close to simple reduction of akrasia (acting against or failing to act on one's higher-order or considered judgments of what one should do). Akrasia reduction will be discussed in detail in Chapter 10 as a form of indirect enhancement, but removing akrasia is subtly different from simply enhancing what someone thinks should be enhanced. With the latter, the enhancer and enhancee identify a specific belief, trait, disposition, etc. and attempt to improve it. This makes it a direct form of enhancement and susceptible to a number of the problems of overlapping consensus models of direct moral enhancement enumerated below. Removing akrasia, by contrast, is indirect insofar as it does not identify specific beliefs, traits, dispositions, etc., but rather focuses on helping a person act on their considered or higher-order judgments, whatever they may be.

Common moral ground might be sought through appeal to common sense morality – general moral truths, such as the badness of suffering or the importance of promise-keeping, to which everyone, or most everyone, agrees. Moral enhancers could focus on these areas of agreement, and leave the problem cases aside. This approach, however, runs the risk of conventionalism – simply counting some manipulation as a moral enhancement because most people believe that it is.[37] While some might simply accept adopting a conventional moral standard, proponents of moral enhancement will need to find other grounds for what counts as a moral improvement. Such conventionalism is in itself arguably implausible, and moreover it seems unlikely that everyone could agree to a conventionalist standard of morality.

To avoid this difficulty, we might seek common moral ground through more systematic moral theory. There have indeed been a number of attempts to find common moral ground in the philosophical literature. One of the most prominent comes from Tom Beauchamp and James Childress's Principles of Biomedical Ethics. (2009) Beauchamp and Childress argue that, while various normative theories such as Kantian deontology, consequentialism and virtue ethics differ in their content, they all support the same rough mid-level principles: autonomy, justice, beneficence and non-maleficence. All mainstream normative theories are in agreement that those four principles are important and valuable; it is (at least prima facie) good to respect autonomy, be just, act beneficently and not harm others. While these mid-level principles are somewhat general, they nevertheless have specific content and can be a strong guide for action. Beauchamp and Childress's four principles have become a very popular tool among bioethicists to reach consensus and avoid sticky

[37] One could try to ground convention in the population's general deliberation over an issue, perhaps giving it some reliability. But then, it is arguably the deliberation – rather than society itself – that is providing the basis for the standard. Moreover, in practice, conventionalism is unlikely to be the output of a reasoned, open and fair debate – too many biases and irrationalities pervade social interaction to have such an optimistic expectation.

normative disputes. Similarly, they could be used by proponents of moral enhancement – we could enhance people's respect for autonomy, sense of justice, inclinations to help and aversions to harm.

Derek Parfit (2011) argues for an alternative form of reconciliation. While his goal is not to generate an overlapping consensus, an upshot of his work is that a number of major normative theories are actually in agreement with each other. According to Parfit, three major normative theories – consequentialism[38], Kantian deontology and Scanlonian contractualism (an action is wrong if " its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced general agreement." (Scanlon 1998, p. 183) – are actually extensionally equivalent. Kantian deontology is aligned with contractualism because universal principles are just those that no one could reasonably reject. And, more controversially, Parfit believes that no one could reasonably reject consequentialist principles (act according to maxims that, when followed, would make for the best state of affairs). While this leaves out at least some prominent normative theories such as non-Kantian deontology and virtue ethics, it would avoid a good portion of normative disagreement. Moral enhancement, then, would be undertaken under a consequentialist framework.

Another strategy, typically adopted by defenders of moral enhancement, is to identify more specific traits that would generally be accepted as morally good or bad – somewhat in line with common sense morality. If such agreement could be found over individual traits, then enhancing such traits should be relatively uncontroversial and avoid the problem of

[38] Parfit believes his account only applies to rule-consequentialism (always act in accordance with principles that, if followed, would generally make things go best) on the basis that act-consequentialism (always do what would lead to the best state of affairs) would undermine the value or certain intentions and motives not directly aimed at bringing about the best consequences. (Parfit 2011, p. 406) This is unconvincing, as bringing about certain motives or intentions (if that would make things go best) is reasonably within the scope of act-consequentialism. Parfit's theory, then, applies to both rule- and act-consequentialism.

disagreement. This strategy also has the advantage of specificity; instead of identifying generic principles, it identifies specific domains that may be more amenable to direct enhancement. So what are some of these generally agreed-upon traits? Walker (2009) believes there is an overlapping consensus that truthfulness, caring and justice (though what justice means is unspecified) are morally good; Persson and Savulescu (2008), while not explicitly arguing for an overlapping consensus, single out altruism and justice (which they take to be, at its core, originating in tit-for-tat reciprocity and equality) as uncontroversial traits; and Douglas (2008), flipping the equation, claims aversion to other racial groups and the impulse towards violent aggression are generally agreed to be morally bad motives.

*Problem 1: Disagreement over agreement*

The most obvious problem with using the sort of overlapping consensuses developed above is that there will be disagreement over either the truth of accounts themselves or whether they are truly uncontroversial or generally accepted. The fact that there are so many different accounts of overlapping consensus implies an initial difficulty here; while the varying views are not strictly speaking incompatible, they suggest some level of disagreement over what the precise nature of a purported overlapping consensus is.

Beyond the diversity of overlapping consensus accounts, serious theoretical objections could be made to the accounts above. While Beauchamp and Childress (2009) remain popular, there are a number of detractors. The mid-level principles account rests on the theory of a common morality (there are at least some norms that all 'morally serious' people share). Some worry that this will make the account overly conventional or

descriptive, which is problematic for reasons discussed above – conventionalism is a controversial moral approach and would make it difficult to challenge the conventional wisdom, as sometimes seems morally required. (DeGrazia 2003) Others worry that the common morality focuses too much on rule-following and does not adequately capture the complexities involved in leading a moral life. (Karlsen and Solbakk 2011) And still others find that, in practice (especially in Western societies), following *The Principles of Biomedical Ethics* leads to an undue emphasis on autonomy over other values. (Holm 1995; Tsai 1999)[39]

Similarly, Parfit's attempted unification is open to a number of criticisms. There is, of course, inherent difficulty in Parfit's task of showing how three seemingly-distinct and mutually exclusive mainstream normative theories not only overlap but are, for all practical purposes, equivalent. Scanlon in particular rejects Parfit's claim that consequentialist principles could not be reasonably rejected, partly based on a reluctance to allow for aggregation that would lead to the few to be sacrificed for the sake of the many. (Scanlon 2011; see also Otsuka 2009) And, it is arguably even more difficult to reconcile Kantian deontology and consequentialism, which have traditionally been seen as incompatible rivals. This reconciliation has been criticized on the grounds that it distorts Kantian deontology to the point that it is no longer reconcilable (Morgan 2010) as well as falsely claims that consequentialist principles are the only ones whose universal acceptance could be rationally willed. (Ross 2010)

Picking out particular traits, as defenders of moral enhancement tend to do, can avoid such theoretical disputes because they do not specify a particular grounding for what makes any given trait good or bad. However, this apparent advantage also comes with serious costs

[39] This is arguably not a failure of the theory but in its application – if people properly attended to the work, autonomy would not be privileged over justice, beneficence or non-maleficence. Still, we are concerned here with a various theories' application for moral enhancement; it would be important, and problematic, if in practice following Beauchamp and Childress would lead to over-emphasis on enhancing something like respect for autonomy.

– the justification for why any given trait is good or bad becomes unclear. This mirrors a critique of earlier editions of *The Principles of Biomedical Ethics*, which omitted the common morality as a unified justification for the four principles and made the four principles seem somewhat arbitrary. (Clouser and Gert 1990) There are then two horns of a dilemma for defenders of overlapping consensus: justify the overlap, risking again the problem of disagreement over that justification, or do not justify it, which implies there is no particular reason to accept one set of traits as good over another.[40]

While it may be possible to overcome these objections, they suffice to show that significant moral disagreement remains concerning the soundness of various accounts of overlapping moral consensus. Thus, theoretical attempts to find overlapping consensus cannot be used to avoid the problem of disagreement.

*Problem 2: Disagreement over content*

Another difficulty emerges when we examine more closely the supposed agreement over the goodness of various traits. Justice, in particular, is subject to intense disagreement over its content; the apparent conception employed by Persson and Savulescu (2008), approximating reciprocity, is somewhat counterintuitive. The literature on justice is indeed quite diverse, and a number of varying accounts can be given. Even if we were to all agree with Beauchamp and Childress (2009) that justice is generally accepted as good, deep disagreements will emerge when we try and specify what, precisely, justice consist in. Recall, from Chapter 3, serotonin's tendency to reduce rejection of seemingly-unfair offers in

[40] The proponent of moral enhancement might claim that justification does not matter for practical purposes. This might allow them to pursue moral enhancements without much public outcry, but at a significant cost – the enhancer's actions would be morally unjustified, undermining any claims about the supposed moral nature of any given enhancement.

ultimatum games. (Emanuele et al 2008, Crockett et al 2010) Does this mean serotonin makes someone more unjust? Such would rely both on an account of justice that involves valuing equality (contra Persson and Savulescu) as well as punishment; both claims would be subject to considerable disagreement.[41]

Beneficence might seem to be an easier case; people can generally agree that beneficence consists in promoting others' interests, even at considerable personal cost. Nevertheless, disagreements lurk below the surface. In particular, there will be disagreement over what counts as promoting someone's interests – is it beneficent, for instance, to paternalistically impose a live-saving treatment on someone who refuses medical care? Also, it there will be disagreement over the scope of beneficence. For example, is friendship an instance of beneficence? Consider when someone prioritizes the interest of one's friends over one's own interests as well as the interests of non-friends. Are they acting beneficently?

These questions of content will emerge for most relevant moral traits. The difficulty of such questions should not be underestimated, and may provide problems for any attempts to carry out moral enhancement under the purview of apparent agreement.

*Problem 3: Disagreement over balance and mean*

Defenders of moral enhancement tend to think of morally good traits or dispositions in isolation; tendencies towards altruism or justice are considered as independent domains that could be enhanced without affecting other morally relevant traits. However, we can see from the empirical evidence presented in Chapter 3 that, at least at present, biological

[41] This problem persists even if the enhancer focuses on prima facie reasons for action. While that may leave scope for disagreement over the weight of contrary reasons, we still face the question of the content of such reasons as well as their relative weight – particularly releveant in the next subsection on balance and mean.

interventions that moral manipulations are not domain-specific. Oxytocin, for instance, seems to bolster cooperation (Zak et al 2007, Israel et al 2009), but also increases in-group bias. (De Dreu 2011) Even if we were to agree that cooperation is good, while in-group bias is bad, the appropriate balance between them would be unclear. Indeed, the main currently-studied biomedical means of moral manipulation all involve tradeoffs of this kind, and will doubtless engender disagreement over the proper balance of traits.

One might counter that those moral tradeoffs just a practical limitation of currently available interventions. With enough research, perhaps domain-specific interventions can be developed, such that no such tradeoffs would be necessary. However, there is some reason to doubt the likelihood that domain-specific interventions can be developed. There is, instead, an arguably theoretical link between various competing moral values. In-group bias, for example, is arguably just a form of cooperation – specifically, the inclination to cooperate more with the group of which one is a member. It then seems likely that to excise in-group bias would also involve excising particular inclinations to cooperate with specific groups. Other conflicts between values are well-known, depending on how the concepts are spelled out; beneficence can conflict with non-maleficence (e.g., whether it is permissible to sacrifice the few for the sake of the many), autonomy with justice (e.g., what degree of autonomy-violating punishment is justified), and so on. Without prior agreement over the appropriate weighting of these values, determining the proper balance will inevitably lead to intense, and potentially irreconcilable differences over how to properly carry out moral enhancement.

And just as there will be disagreement in determining the proper balance between various theories, there will also be disagreement over the appropriate level or mean of a particular trait on its own. Justice or lack of violent aggression might seem like a good thing, but must be taken in moderation. Too much concern over justice might lead to someone being overly-punitive and draconian, while too little aggression could lead to failing to

prevent grievous harms to others. Determining the appropriate level might not be strictly necessary for everyday action, but such would be crucial for an intervention aimed at altering people's moral beliefs, motives or actions – we would need to know whether any given individual has too much or too little of a given trait. Yet the sort of disagreement that would emerge over that level will inhibit the possibility of moral enhancement.[42]

The issue of morally appropriate balance and mean indicates a problem for moral enhancement beyond the argument concerning access to others' mental states, one that generalizes to the moral nature of actions as well as beliefs and motives. Normative theories are not, in general, well-equipped to properly assess the appropriate balance or mean of various moral traits. Utilitarians may have a clear standard of goodness, maximizing utility, but how can one clearly show that possession of one trait over another will have the best consequences for overall utility? At least with evaluation of individual actions, the consequences can be somewhat contained to generate reasonable expectations. However, traits affect so many different domains of action and will have such diverse effects that evaluation will become epistemically quite challenging, perhaps out of reach. Virtue ethics could somewhat avoid this by noting that part of practical wisdom consists in being able to identify the proper balance or mean of various virtues. However, the question would then remain – what would the practically wise individual recommend as the appropriate balance and mean of various traits? Perhaps a defender of moral enhancement could claim to have such virtues, or appeal to alleged experts, but it would be extremely difficult to validate such claims. In any event, defenders of moral enhancement have not, to date, relied on appeals to moral experts to validate their claims.

[42] Determining the appropriate level may not be a problem for certain traits, like racism, where there is clear agreement that the complete absence of the trait is ideal. Still, in such cases the issue of the appropriate balance between various traits (as discussed in the previous chapter, making people less racist might also make them less cooperative with their group) would remain.

The upshot is that disagreement reveals a deeper degree of uncertainty concerning how, specifically, to carry out moral enhancement. Without a systematic, principled way to adjudicate the proper balance and mean of various traits, it is not clear how we could confidently claim any given intervention as moral enhancement. We could try and rely on our intuitions, but those are likely to be too vague and unreliable to ground most forms of moral enhancement. While determining the moral nature of particular actions may be possible, the additional complexities characteristic of our moral traits will make any confidence an enhancement misplaced.

# Chapter 6: The Value Problem

**Overview**

The purpose of the previous chapter was to argue that the existence of moral disagreement poses difficulties for moral enhancement. But even if those worries were overcome, disagreement remains a problem for direct moral enhancement in a different way. Specifically, we have some reason to be wary of moral enhancement because of the value of moral disagreement. Moral disagreement – while potentially inhibiting consensus-building – is actually an important feature of society, one which would be under threat by some programs of moral enhancement.

Unlike the previous chapter, the arguments below are primarily designed as an objection to widespread programs of moral enhancement – programs aimed at morally improving a large portion of a society (be it local, national or global). The weak claim that individual, isolated instances of moral enhancement are permissible or justified (e.g., Douglas 2008 & 2011) will not be directly addressed (though the subsection on individuality might be applicable to individual instances of moral enhancement). Instead, the primary target will be the stronger claim that we should embark on large-scale efforts to morally enhance entire groups of people (Persson and Savulescu 2008, 2010 and 2012; Walker 2009). There is, to be sure, a somewhat compelling prima facie case for that stronger claim. Social policies generally have as their ends the betterment of the group; moral enhancement of the group should lead to a more moral society, which is surely desirable. But as we will see, that betterment may come at a serious, indeed prohibitive cost.

There are a variety of ways large-scale moral enhancement might be brought about. Most obvious would be federal coercion, with the state mandating that its citizens (or future,

unborn citizens) undergo interventions such as moral education or pharmaceutical injections aimed at moral enhancement. As discussed previously, directly coercive policies are rather problematic for reasons other than disagreement – it is generally considered illegitimate for the state to interfere with people's bodies (though notably not in all quarters; see, e.g., Fabre 2006). There are, however, other 'softer' forms of large-scale enhancement that might not seem so objectionable. Incorporating moral instruction into public school curricula would not be substantially more coercive than current educational policies. The state might also selectively ban certain interventions it considers detrimental to moral belief, motive or action, while allowing those that it judges to be moral improvements. Alternatively, the state could subsidize certain enhancements or put forward propaganda campaigns aimed at promoting widespread adoption of interventions aimed at moral improvement. The arguments in this chapter will apply to all such methods of widespread moral enhancement, as they all employ a similar, objectionable approach to achieve moral improvement – bringing as large a portion of the population as possible in line with what the promoter of moral enhancement considers to be moral.

The emphasis on widespread moral enhancement suggests that this chapter will involve primarily political considerations. Widespread enhancement would most plausibly be the project of a governmental agency, one that can coordinate a single enhancement program over a large population. Still, the arguments below are meant to be more general. States are not the only entities that could embark on a program of widespread moral enhancement. A private charity, for example, could put up funds to subsidize certain forms of moral enhancement – essentially paying people to alter themselves to become more moral, in the eyes of the charity. A less well-organized but potentially more effective campaign might involve fostering an environment where one is pressured by one's peers to undergo a widely-supported procedure meant to enhance morality. The standards would, in such cases,

be set and (non-coercively) enforced by a non-state entity, and would still be vulnerable to the objections outlined below.

The following will, to a large extent, mirror arguments put forward by John Stuart Mill in *On Liberty*, especially chapters 2 and 3. The main aim of *On Liberty* was to defend the harm principle (a necessary, but not sufficient, condition on the acceptability of interference with an individual's liberty is that the intervention prevent harm to other individuals). In order to argue for that principle, Mill offers strident defenses of freedom of opinion as well as the value of individuality. This means not just the freedom to hold a dissenting opinion but also the freedom to criticize others and have the idea debated in public without censorship – thus applying not only to attempts to morally enhance beliefs and motives but also actions. While of course Mill was not thinking of the sort of biological interventions available today, his arguments are very much applicable to these relatively recent developments and speak powerfully against widespread moral enhancement.[43] Certain aspects of Mill (e.g., his utilitarianism) are not meant to be endorsed here, but I take his arguments to be generally sound and convincing.

**An initial worry: self-fulfilling enhancement**

Before delving into Millian concerns, let us consider an initial problem for a certain form of moral enhancement. This example is not meant to be devastating for the defensibility of moral enhancement, but instead illustrate some of what makes such interventions

[43] While this discussion will focus on moral enhancement, Mill's arguments actually speak against any intervention (biological or otherwise) seeking to alter people's beliefs or motives through non-persuasive means. Attempting to bring about cognitive enhancement by simply instilling the correct empirical beliefs would also, then, be objectionable for the same reasons detailed below.

disagreeable and point towards a systematic analysis of why we might object to widespread moral enhancement.

The interventions discussed in Chapters 2 and 3 affected a diverse range of moral domains. However, imagine we were able to develop an intervention that (perhaps among other things) is able to inculcate the belief in a particular proposition – that moral enhancement is worth promoting. It may be too fanciful to imagine this proposition can be directly or specifically implanted in a person's mind; however, it is not very far-fetched that certain interventions might make someone significantly more disposed to accept that proposition than they otherwise would be. We will, for present purposes, put aside what other effects on a person's psyche this intervention might have, and focus on the inducement of this particular belief.

A proponent of moral enhancement will typically take the proposition that moral enhancement is permissible and should be promoted to be true. This is, moreover, a moral belief. Working for the moment under the assumption that moral enhancement consists in inculcating correct moral beliefs, the proponent of moral enhancement then would, by his or her own lights, have strong reason to include inducement of this belief as part of a widespread program of moral enhancement. The result would be that the program of moral enhancement would end up with a sort of self-fulfilling or self-promoting character. Widespread moral enhancement would lead to widespread support for moral enhancement, even of a sort that was initially popularly opposed or only barely supported.[44]

A similar effect would occur if moral enhancement consists in improving motives or behavior. Since moral enhancement is, according to supporters of enhancement, worth promoting, a large-scale program of moral enhancement would include the inculcation of the

[44] A recent survey in the US suggests that, currently, there would not be much popular support for widespread enhancement programs not aimed at improving health. (Hays, Miller and Cobb 2011)

motive or disposition in others to actively promote moral enhancement. Once again, the result is large-scale unification on the issue. A population previously lukewarm towards moral enhancement (perhaps willing to permit it but not especially inclined to promote it) would, after wide-scale enhancement, become enthusiastically and uniformly motivated to support and perform further moral enhancements.

So what is wrong with this unification scenario? From the perspective of the proponent of moral enhancement, it may seem quite desirable – everyone has come to have the proper belief, motive or behavior in regards to moral enhancement. One problem might be the over-promotion of moral enhancement. The selection for promotion of moral enhancement might lead to excessive resources being devoted to moral enhancement, resources that should actually be spent elsewhere (perhaps because more moral enhancement has diminishing moral returns, so to speak). This problem could be overcome by more carefully fine-tuning the enhancement, such that people support enhancement to a moderate degree, but not so much that they devote too many resources to it.

There is, however, a deeper problem with this scenario. It is intuitively disconcerting. The sort of unification imagined, and the means by which it is brought about, has a dystopic aura about it – something that seems straight out of *A Brave New World*. Even though we have limited the enhancement to one specific domain, the resulting society has been fundamentally altered. This alteration is rather unique to moral enhancement; similar universal enhancement of people's health or intelligence, for instance, does not seem to have the same effect. These vague concerns do not constitute a real objection, but they do suggest further investigation; we should not rush headlong into a large-scale moral enhancement program without carefully examining the rational grounds for these concerns.

Indeed, these worries are not limited to the present example. The same result – a societal shift towards unification around a particular set of moral ideas, motives and/or behaviors - is part and parcel with the aims of any moral enhancement program. There may be a number of reasons for discomfort with this result, but I believe the central problem with unification is the effective suppression of moral dissent. As the next section will lay out in detail, moral unification through moral enhancement is objectionable because it is too great an infringement of the value of moral disagreement.

**The sources of the value of moral disagreement**

The value of moral disagreement can be derived from three somewhat interrelated sources: moral fallibility, reasoning, and individuality. Moral fallibility will entail a strong instrumental reason to preserve moral disagreement in a society, while moral reasoning and individuality are values threatened by the absence of moral disagreement. I will discuss each in turn; together, they constitute a cost to moral enhancement that will not be outweighed by the alleged benefits.

*Moral fallibility*

We are, without a doubt, fallible creatures. This is especially true when it comes to morality; the level of disagreement over moral issues should make us reluctant to claim true certainty about many moral claims. Indeed, proponents of moral enhancement will admit as

much about the general population – if people were not morally fallible, there would be no need for moral enhancement. This, however, holds true for the moral claims made by proponents of moral enhancement themselves. Any given program of moral enhancement, then, will run the serious risk of being wrong-headed.

Why should this be a particular problem? After all, every government action – indeed, every action – is subject to moral fallibility. Such does not seem to be a general reason against action, so perhaps it is not a problem for a program of widespread moral enhancement. But Mill points out that certain sorts of actions are problematic in the face of fallibility – specifically, actions intended to stamp out dissent:

> There is the greatest difference between presuming an opinion to be true, because, with every opportunity for contesting it, it has not been refuted, and assuming its truth for the purpose of not permitting its refutation. Complete liberty of contradicting and disproving our opinion, is the very condition which justifies us in assuming its truth for purposes of action; and on no other terms can a being with human faculties have any rational assurance of being right. (Mill 1999, p. 62)

The idea is that, by suppressing dissent, we cut off a crucial avenue of coming to adopt the correct moral beliefs and policies.[45] Without dissent, conventional wisdom will go unchallenged and moral progress becomes essentially impossible. This might not be a problem if we were infallible (i.e., already knew all the relevant moral truths), but because we are not, such actions will prevent the revision of morally odious policies that, at the time of suppression, seemed perfectly sound. Dissent is instrumentally valuable, then, as a constant

[45] Mill may have gone too far in claiming that allowing disagreement is necessary to justify any moral proposition. If one believes there are self-evident moral truths, those truths could be rational to hold in the absence of any critiques. Still, this is compatible with the notion that vigorous disagreement, which forces people to offer arguments and justifications for their positions, is necessary for the justifiability of a wide range of non-self-evident truths, and that it can indeed significantly bolster the justifiability of even self-evident truths.

check on the validity of the conventional moral wisdom of our time. Morality, in other words, should be allowed to evolve.

This position is particularly compelling if one is generally optimistic about moral progress. If we expect moral ideas held by the public to, by and large, become more and more in line with the truth, then there is very strong reason to want to preserve the ability of moral ideas to evolve. And there is some reason to think that human history has generally trended towards more morally upright positions. The gradual trend of societies towards toleration and civic inclusion of marginalized groups (e.g., women, minorities, foreigners, and more recently animals) as well as the spread of democratic political values are examples of this trend. But even if one thinks, on the contrary, that public morality has either not progressed, or indeed has regressed, one should still support preserving the ability of society to evolve. Such evolution does indeed run the risk of further regress. However, preventing moral evolution means shutting off any opportunity to improve or to return to the (allegedly) morally superior days of yore. In other words, it means preventing one form of moral enhancement, the sort that occurs over generations and can lead to massive social improvements. This should be a serious concern for anyone who thinks that moral enhancement is a valuable enterprise.

Mill's intention was, admittedly to argue against traditional forms of suppression of ideas (such as censorship and religious persecution). Still, the point is equally valid against programs of large-scale moral enhancement, where the ultimate effect is to eliminate or vastly reduce dissent. This is an inevitable result of a large-scale program of moral enhancement; just as health enhancement stamps out disease and cognitive enhancement stamps out intellectual deficiency, moral enhancement stamps out moral beliefs, motives and actions that differ from the enhancers'. Yet, unlike disease and intellectual deficiency, those supposedly deficient moral states serve a valuable purpose, to challenge the conventional

wisdom and ensure that "the means of setting it right are kept constantly at hand." (ibid, p. 63) This even applies to enhancements aimed purely at improving behavior (e.g., Persson and Savulescu 2008); while such do not strictly require making people assent to what the enhancers take to be right, it is difficult to imagine ensuring widespread compliance with a particular moral standard without ensuring widespread agreement with that standard.

Put another way, widespread moral enhancement could lead to moral stagnation. Mill uses an unfortunately ill-informed analysis of Chinese culture to illustrate this risk (Mill 1999, pp. 118-9), but the point still stands – true moral uniformity would mean we become 'locked in' to whatever moral propositions we happen to believe now. Indeed, the risk of this stagnation is arguably much greater in the face of biological enhancements than the sort of censorship and suppression Mill was opposing. Perhaps for the first time, we have the technological capability to not only restrict the spread of ideas and behaviors, but the very thinking of those ideas in the first place. In a generation or two, reasoned discussion and debate of previously-controversial moral issues (of the sort that might illuminate fallacies and lead to change) could come to be seen as pointless. This sort of stagnation would cut off the possibility that a society could come to revise its previous moral ideas.

Even if the full stagnation of moral ideas does not materialize (admittedly, such would require immense, likely coercive federal policies), direct moral enhancement is problematic to the extent that it impedes moral progress. A private organization promoting direct moral enhancement to bring a few minds to its way of thinking may seem innocuous. But like any social problem, this becomes a serious issue if such 'conversions' are widespread. The large-scale shift towards the organization's way of thinking via direct interventions reduces the diversity of thought in society in a way that does little to preserve true moral progress. An alternative scenario where a number of different organizations promote different moral enhancement programs is still problematic, to the extent that divergence from moral positions

endorsed by such organizations is suppressed. Furthermore, we should be suspicious of the moral reliability of the motives of well-funded organizations (as well as governments), who have strong incentive to push people not towards what is actually more moral, but towards what better serves those organizations' interests.

Against this, a proponent of moral enhancement might simply insist the instrumental cost of such stagnation is outweighed by the instrumental benefit of moral enhancement. Avoiding catastrophes like nuclear war or environmental devastation through moral enhancement (Persson and Savulescu 2008 & 2010) is too important to let quibbles about cessation of moral progress get in the way; better to have a morally stagnant society than no society at all. Interestingly, Mill anticipated this objection, and has a useful reply. Such claims about the importance of uniformity have been made countless times in the past to justify programs of suppression; for instance, even the wise philosopher-emperor Marcus Aurelius held that doctrinal unity was so supremely important to social cohesiveness that Christianity must be brutally suppressed. But we, like Marcus Aurelius, are fallible; while we can tolerate such fallible action insofar as errors can be corrected over time, the unique aspect of suppression is that it prevents such correction. One might think that moral enhancement of a certain sort is necessary to avoid one form of catastrophe; but we cut off the possibility of performing moral corrections that might avoid further catastrophes down the line. Such corrections are indeed necessary to ensure not only our well-being, but perhaps even our very survival.

*Reasoning*

Dissent is important for ensuring moral progress, but it also has value independent of such down-the-line effects. We may want a society that not only has the correct sort of moral beliefs, but has them for the right sort of reasons. This value is somewhat hard to pin down, but Mill argues it is important for people to come to their ideas (especially moral ideas) through reasoning rather than external authority. (Mill 1999, pp. 80-1) This is directly an argument against social conformity and in favor of rational deliberation, but it also indirectly suggests that moral enhancement problematically cuts off our reasoning processes. Instead of coming to believe or act on a given moral proposition because it is the most reasonable, we would come to believe or act on it because a particular external agent (the enhancer) said it is best.[46]

It is quite plausible to think that there is value in the process itself of deliberating over a moral proposition, both within one's own mind and in discussion with others.[47] Part of this value might be instrumental (promoting better, more accurate ideas, as per the preceding subsection), but there is also a compelling sense in which reasoning is valuable in itself for the reasoner. This idea goes back at least to the *Apology*, where Socrates famously argues that "the greatest good of a man is daily to converse about virtue…and the life which is unexamined is not worth living." (38a) We need not adopt this apparently extreme a view of the value of moral discourse and contemplation to agree with its core insight – there is

[46] At times, Mill seems to say that, beyond the importance of reasoning, it is important for opinions to have a foil in order that ideas would have a "clearer perception and livelier impression of the truth." (Mill 1999, pp. 59-60) The value of such 'liveliness' is significantly less convincing (and less clear) than the value of coming to a view through reasoning and deliberation. And, in any event, it may just be that such liveliness is not meant to be an internal characteristic, but rather just one way of explicating the value of (lively) deliberation over ideas. In any event, I will set aside the 'liveliness' conception of the value of disagreement for the present discussion.

[47] This is not to say that the notion or value of reason is exhausted by such a process. Rational intuitionism, for instance, takes basic, intuitive, non-procedural grasp of certain concepts to be a part of reason, and arguably there is value in itself to properly grasping such basic concepts. (Audi 2004) The present argument is consistent with such a view, so long as one leaves room for the existence and value a procedural component to reason.

something intrinsically good about such reasoning, something worth promoting and protecting.[48]

Yet, widespread direct moral enhancement would in all likelihood reduce that discourse. People tend not to debate or reflect much on issues about which there is no doubt or dissent. Discourse and contemplation are motivated in part by the existence of disagreement. This makes sense, after all; the purpose of such discussion tends to be the discovery of truth through reason. Disagreement prompts one to either doubt whether one's own ideas are actually true, or (perhaps more commonly) attempt to correct the opinions of others through discourse. But even the latter motivation can lead to real revision – in the process of defending a sincerely held position, one might come to notice its flaws and correct accordingly. As we have seen, widespread moral enhancement would vastly reduce that disagreement; it would therefore also lead to significantly less moral discourse (at least over issues pertinent to what is being enhanced) that seems so valuable.[49]

Perhaps one could try and preserve that value in a morally unified society by promoting debates where one side pretends to hold views that everyone takes to be false, and the other holds views that all assent to. But that misses out on the apparent value of the discourse (indeed, it comes close to the sort of sophistry decried by Socrates). Reasoning consists not just speaking certain words, but actually and sincerely entertaining the possibility that one is wrong – and being open to revision of one's beliefs or actions in the face of error. Structured discussion without true disagreement would only be a vague facsimile of the lively

[48] One recent suggestion comes from Alison Hills (2009): there is value in moral understanding, which is usually not acquired via the moral testimony of others but through one's own reflection and deliberation. Direct moral enhancement would be problematic for the same reason as accepting moral testimony – it subverts or works around those personal deliberative processes. Reasoning, on this account, is not merely instrumentally valuable in bringing about moral understanding; it partially constitutes moral understanding.

[49] It is of course possible to have too much discourse, in the sense that one spends too much time/resources debating when a decision should be made. However, the proper way to avoid inappropriately prolonged debate is not to directly induce agreement, but rather improve people's ability to discern when a debate has run its course. This skill is a component of sound reasoning (as I will discuss later on in Chapter 9), and improving it will be a form of indirect rather than direct moral enhancement.

moral debates that pervade modern society – more akin to the 'Two Minutes of Hate' from *Nineteen Eighty-Four* than true deliberation.

Even without a unified society, moral enhancement will be problematic to the extent that one's ideas can no longer be said to be the product of one's own reasoning. Direct moral enhancement makes one's own reasoning process more or less obsolete; even if one goes through the motions of reasoning, the enhancers have 'rigged the game', so to speak, to ensure the desired outcome. This will be the case even with some non-coercive direct moral enhancement. To the extent that people voluntarily give up on reasoning, they will be abandoning something that is of immense personal value.

One might try to sidestep these issues by having direct moral enhancement target basic intuitions that operate prior to any reasoning processes. The value of reasoning could be preserved because the intervention does not affect reasoning processes at all – only the intuitive inputs into that process. This approach does indeed avoid the present argument concerning the value of reasoning, but is liable to another serious objection. The difference between the enhancer and the enhancee, at this stage, would be over what are the proper basic moral intuitions to hold. Why should we believe that the enhancer's basic intuitions are any more reliable than the enhancee's? Faults of reasoning cannot be appealed to, as these basic intuitions operate independently. This issue will return again for more extended discussion in Chapter 9, but for now I will just note that there does not seem to be a principled reason to privilege the enhancer over the enhancee's intuitions, and so there is not much reason to think that any alteration of those intuitions in favor of the enhancer's would lead to moral improvement.

Alternatively, one might promote a form of direct moral enhancement that operates on reasoning processes. Suppose, for instance, that an enhancer believes altruism is good, and

wants to make people more altruistic. But this enhancer also believes that the best way to bring about more altruism is to improve on certain reasoning processes (in line with the strategy discussed below). Such a strategy, on its face, could respect the importance of people's ability to reason and think for themselves. However, this strategy still falls into the pitfalls listed above. The forms of reasoning inculcated will be narrowly construed as those leading to the enhancer's way of thinking. This runs the risk of the enhancer making everyone think the same, in order that everyone come to be in agreement with the enhancer's way of thinking. What's more, the narrow conception of reasoning promoted by the enhancer would be justified not based on the merits of the reasoning process itself, but the conclusions that process draws. This puts the cart before the horse in ensuring that, ultimately, we do not come to believe and act in certain ways because those beliefs or actions are supported by the best reasoning processes, but instead because they are supported by whatever reasoning processes the enhancer thought efficient at bringing about certain effects.[50]

*Individuality*

A related source of the value of disagreement is the importance of individuality. It is not only important that one's beliefs, motives and actions are the result of reasoning, or lead to good outcomes; it is important that those beliefs, motives and actions are one's own. Mill takes this individuality to be a singular human value:

[50] The primary problem here is that reasoning processes are selected not because they are in themselves the best or most reliable procedures, but because they produce certain results. Below, I will suggest that improving reasoning can be an acceptable form of indirect moral enhancements, insofar as the opposite is true: reasoning processes are selected not because of the particular results they generate but because they are morally reliable in virtue of the processes themselves. The key difference is that the indirect moral enhancer will make no reference to specific, desired outcomes of a given reasoning process, letting the agent's reasoning process itself – rather than the enhancer –be the real determinant of the agent's beliefs, motives and actions.

> It is not by wearing down into uniformity all that is individual in themselves, but by cultivating it and calling it forth, within the limits imposed by the rights and interests of others, that human beings become a noble and beautiful object of contemplation. (Mill 1999, p. 109)

Uniformity is a threat to such cultivation insofar as it makes people more like steam-engines – passive, mechanistic beings under the instrumental control of another – than men and women of true character. (ibid, p. 107) Though Mill did not use the term, it could be said that uniformity is a threat to autonomy, insofar as it would make acts and thoughts less autonomous and more driven by the enhancer's motives. Diversity of opinion – especially moral opinion – would importantly preserve that autonomous character, insuring that people retain their individual nature that we value so much.

It might be thought that Mill's concern for individuality cannot motivate an objection to certain forms of large-scale moral enhancement. Mill, at times, emphasizes the dangers of conscious conformity – to be like a steam-engine in to be the sort of person that allows others' opinions to be a substitute for one's own. Perhaps it is just that sort of individuality – not ceding to the opinions of authority figures – that is valuable. If that is the only issue, moral enhancement can avoid running afoul of individuality. Specific moral beliefs, motives and behaviors can be inculcated, after all, without inculcating the tendency to not think for oneself. A program of moral enhancement does not require that people assent to the enhancers' opinions because of their moral authority; it could alter people's moral intuitions directly, such that people never even consider an authority figure when undertaking moral deliberations.

Still, the essence of Mill's critique points to another way in which individuality is valuable. Individuality is not just about the structure of conscious moral thought; it is also

about the origin of those thoughts. It is important that we ourselves, and not others, are the originators of those thoughts. The steam-engine analogy is disturbing not because steam-engines conform, but because they have no independent will of their own. A society that imposes its will on the actions and thoughts of its members thus robs those people of their individuality; the group's will (or the will of the group's thought-leaders) is substituted for the individual's will. Similarly, moral enhancement replaces the individual's will with the will of the enhancer.[51]

This danger is apparent in traditional forms of moral enhancement, but the nature of biological enhancements makes the will-substitution especially problematic. Society has traditionally affected thoughts only indirectly and externally – putting pressure on those who dissent (either with coercion or softer methods), while rewarding those who assent. People will tend to adjust their ways of thinking in response to those social forces, but at least that adjustment is within people's control. Biological moral enhancement admits of no such individual control. In these cases, we are as close to thought-control as we have ever been in human history.[52] This is not the sort of imminent, constantly-active mind control imagined in science fiction. However, it is close, insofar as certain alterations to people's moral ideas (especially when inculcated prenatally, through genetic manipulation) can be traced almost entirely to an external enhancer. There is a way to retain individuality in the face of external influence – through one's behavior in the face of moral disagreement. But without that disagreement, there is little hope of people retaining their moral individuality in an enhanced

[51] Douglas (2014), though favourable towards moral self-enhancement, has recently offered some support for a similar objection to third-party moral enhancement from a Kantian perspective: "Where A imposes a brute conformity enhancement on B, B's subsequent conduct might be thought to originate not in B's deliberation, but in A's, and this might be thought to detract from its moral worth." (Douglas 2014, p. 9)

[52] Some analogy may be drawn to intense psychological brainwashing. Yet there is a sense in which direct biological manipulation is even more inhibiting of individuality than brainwashing. At least with externally-stimulated psychological techniques there is the opportunity for the individual to deploy their reasoning capacity and reject the influence. Biological manipulations, on the other hand, are much more direct in how they affect one's mind; one's brain-chemistry is itself altered, and the intervention itself consists in manipulation of one's very personhood, rather than that alteration being a contingent effect.

society. Maintaining disagreement in a society is then important because it is a way in which we can maintain our individuality.

Against this, it might be pointed out that already people's moral beliefs, motives and actions are entirely caused by external forces. No one is an uncaused causer. Individuality, then, is an illusion with no real value; there is nothing bad about coming to have ideas or inclinations due to moral enhancement because the alternative is to have those ideas or inclinations due to the vagaries of genetics, environment and society. We all are essentially like steam engines already (albeit with consciousness) – because there is no way to be otherwise, it is pointless to oppose any action on the grounds that it might make us lose control.

This objection, of course, paints a rather bleak portrait of human nature. In addition to making us not much more than self-aware steam engines, it implies the notion of moral responsibility is mistaken. In order to ascribe moral responsibility to an agent, one must be able to distinguish between cases where an agent is and is not responsible for his or her actions. We can contrast the following cases: (1) Alex punches Ben because she intensely dislikes Ben and (2) Alex punches Ben because Carlos has utilized a mind-control device that compels Alex to do so. There is a clear intuitive difference in Alex's moral responsibility for hitting Ben in the two cases, but the view that individuality is an illusion would deny this. The fact that Carlos has substituted his will for Alex's in (2) makes no material difference, as there were some analogous external influences (genetics, upbringing, etc.) to determine Alex's actions in (1) as well. But this denial of a differences a deeply implausible implication, one that suggests we should reject the individuality-as-illusion account. Instead, should prefer an account where the origin of some action (whether internal or external) is be morally relevant.

Compatibilism is just such an account. There is not space to fully explicate or defend this view here, but it is quite appealing to think that notions like responsibility can be preserved in the face of all these external determinants of action. There is something valuable and morally transformative about volition and personal choice, such that we can indeed properly attribute responsibility to people whose actions result from such personal choice and see responsibility as reduced when personal choice is reduced. Yet distinguishing (1) from (2) involves something more – not just valuing personal choice, but privileging certain origins of choice over others. Cases like (2) are particularly problematic because another agent is involved – one agent's will is substituted for another's. This particularly denigrates the extent to which the agent can identify with her own actions, affecting not only moral responsibility but the broader moral connection between who she is and what she does.

If that is right, then it is no stretch to attribute value to individuality and object to direct moral enhancement on the grounds that it involves will-substitution. Having a certain moral nature due to another's will diminishes this value of individuality, in a way that having that moral nature due to natural chance or reasoned deliberation does not. (Indeed, we might form a hierarchy of value: belief or action due to reason is better than it being due to nature; but both are to be preferred to being due to another's will) Certain direct interventions will be more deleterious to individuality than others, but they all rely on a privileging of the enhancer's moral view over that of the enhancee. In this way, direct moral enhancement involves substituting the enhancer's will (qua deciding what is good or desirable) for that of the enhancee.

*Summation*

Taken together, our fallibility and the value of reasoning and individuality indicate serious instrumental and non-instrumental costs engendered by a program of widespread moral enhancement.[53] Such would morally stagnate society and leave its members without the moral deliberation and individuality that we rightfully find of value. How weighty these considerations are is a matter of debate. For my part, I believe the above arguments show that the value of moral disagreement is sufficient to make widespread promotion of moral enhancement an untenable and wrongheaded policy. But perhaps, despite what has been said, one still thinks that the catastrophic consequences of not engaging in moral enhancement are too great to avoid embarking on widespread moral enhancement. To such people, I would nevertheless urge that they take these costs seriously and do all they can to mitigate them, to whatever extent possible.

**Limitations**

The above concerns are meant to apply to a wide range of traditional and novel techniques of moral enhancement – including moral education, propaganda, and pharmaceutical interventions. However, it must be conceded that not all forms of direct moral enhancement discussed earlier will be susceptible to these objections. In particular, argumentation and punishment – under the right circumstances – need not problematically suppress dissent.

[53] As suggested previously, similar arguments may speak against non-biological forms of widespread moral enhancement such as propaganda campaigns (insofar as they attempt to manipulate rather than persuade). Still, biological enhancement is especially problematic. It works by altering the structure of people's thoughts directly, much closer to brainwashing than psychological manipulation. At least with propaganda or social pressure, one can always deploy one's reasoning capacities to reject the influence. With direct biological enhancement, there is a sense in which we would be powerless to reject it (especially if our moral character was altered prenatally). Moral progress, discourse and individuality would then be under particular threat from a widespread program of biological moral enhancement.

*Argumentation*

Let us consider argumentation first. A primary motivation for many making arguments is to persuade others of one's point of view. In the case of moral arguments, the goal is to directly bring others in line with one's moral opinions.[54] Like other forms of moral enhancement, successful argumentation would then have the effect of suppressing dissent. It may then appear that the above considerations would militate against making moral arguments to others. This would be a mark against the argument, insofar as it is widely accepted that persuasion through non-manipulative arguments are well within the realm of acceptable discourse. In fact, it would imply there is something objectionable about this very work – insofar as I am making an argument aimed to persuade its reader, I would be objectionably attempting to stamp out dissent.

Fortunately for the morality of this work, argumentation (under the right circumstances) need not run afoul of the above objections. Argumentation is actually in a unique position to overcome three problems of most forms of moral enhancement – fallibility, reasoning and individuality. Most generally, argumentation might aim to reduce dissent, but the effect of widespread encouragement of argumentation is quite the opposite. Allowing and encouraging people to debate all manner of moral topics will lead to more vigorous disagreement and dissent. Argumentation could have deleterious effects if it was one-sided. A state might, for instance, only allow arguments in favor of certain preferred moral positions. It is uncontroversial to suggest that such a move would be morally wrong, not only for the Millian reasons laid out above but also due to the coercive intrusiveness on people's everyday life. But as long as argumentation is undertaken in an environment that allows and indeed fosters disagreement, there is little risk to valuable dissent.

[54] A more high-minded motivation might instead be to simply arrive at the correct ideas. In such cases, Millian worries would be weaker though still present. In any event, the considerations in defence of argumentation aimed at persuasion should equally apply to argumentation aimed at truth.

We can see this point more clearly by attending to each facet of the Millian argument in turn. Fallibility is worrisome because a program of moral enhancement may lock us into mistaken moral views. But as noted in Chapter 2, one of the apparent downsides of argumentation as a form of moral enhancement is it would be practically difficult to use it (without employing manipulations) to bring about widespread convergence on a large range of issues. And even if such convergence did occur, argumentation is perhaps the ideal means. Argument and discourse can effectively identify the flaws with certain ideas and bring out the strengths in others; insofar as it engages with reasoning, it will more reliably ensure we end up with the proper moral ideas.[55] Furthermore, being locked into a particular way of thinking is of little risk; so long as the basis of societal convergence is argumentation, that society will never truly be 'locked in' to a particular moral view – as different, more powerful arguments can always (and will likely, given people's often-contrarian natures) be brought to bear.

It almost goes without saying how argumentation does not run afoul of the value of reasoning. Providing arguments to others is perhaps the best way to foster the reasoning process. Coming to change one's views because of a particularly persuasive argument is perhaps the paradigmatic 'right sort of reason' for having a view. And engaging with others in an attempt to bring about such a change fully respects and honors the importance of people using their reasoning facilities to come to a given opinion. If the Socratic imperative to have a considered life is taken seriously, there will be little reason to object to vigorous argument.

Individuality is somewhat trickier, as changing one's views in the face of an argument arguably makes one's interlocutor the originator of one's views. Yet, the causal sense in which an interlocutor originates her opponent's view does not run afoul of the central concern, that one person's will not substitute another's in coming to have a certain view. The

[55] Chapters 7 and especially 9 will engage much more deeply with this connection between reasoning and morality.

central question is the extent to which one's own agency mediates the shift. In the case of manipulations like propaganda, part of the problem is that people's agency is being subverted by appealing to subconscious influences and non-rational capacities. But argumentation respects agency, as non-manipulative argumentation can only be accepted if one exercises one's agency in evaluating the relevant argument and deciding, for oneself, which side is ultimately the correct one. This involves no will-substitution, and so argumentation is relatively immune to the Millian objections that have been leveled against moral enhancement thus far.

*Punishment*

Recall that punishment can be properly characterized as a moral enhancement when it aims to reduce the proliferation of immoral activity, as well as indirectly use legal structures to shape people's moral thinking. The latter goal more properly falls under the category of indirect moral enhancement, which will be addressed in more detail in later chapters. But reducing crime by disincentivizing it can be fairly categorized as a direct moral enhancement, aimed at bringing society into conformity with a certain set of ideas about proper moral behavior. It stamps out dissent (qua immoral activity), and so it may appear that the above arguments imply punishment is morally objectionable.

Punishment is distinct from most of the other forms of moral enhancement discussed because of its very external character: it does not (essentially) aim to change behavior by making people think certain behavior is wrong. Rather, it effects moral enhancement by making acting rightly more in line with self-interest. This is perfectly compatible with continuing to hold one's own views about the morality of particular activities, and so dissent is still possible.

Again, we can look at each of the three sources of objections to moral enhancement. The external character of punishment means it will not significantly contribute to becoming locked into a certain point of view. Laws have been routinely changed in the past, as societies shift in their views about a wide range of issues. There may, to be sure, be a certain amount of conservatism – people clinging to the way things are simply because that's the way things are – which makes the existence of punishment inhibit progress. Yet these forces have only a limited effect. Public defenses of existing forms of punishment almost always appeal to reasons other than the fact that punishment exists, as people generally realize that such defenses are inadequate. Resistance to change is, ultimately, just another cognitive bias that should be avoided, and it would be a bridge too far to argue against punishment simply because it can foster that bias.

Reasoning and individuality are more problematic for punishment. If one acts morally due to fear of being punished rather than because the action is moral, then one is clearly not acting for the right sort of reason. Those reasons should flow from the morality of the action, not (merely) personal interest. And coercive intrusion is a powerful way to subsume the will of another under that of an authority figure. There is no getting around the prima facie wrongness of this kind intervention. But then, that should be unsurprising – punishments (such as imprisonment and fines) are acts that in themselves would be prima facie wrong, whether done by the state or an individual. Further arguments are generally put forward to justify punishments, typically either falling into consequentialist considerations of prevention and deterrence of greater wrongs or more deontic claims about the guilty deserving a certain amount of punishment.

The consequentialist grounds for punishment look strikingly similar to the arguments that Persson and Savulescu (2008) put forward for widespread moral enhancement. People need to be interfered with to prevent greater wrongs from being perpetrated on society. If we

accept the consequentialist grounds for punishment, then, this would imply that the Millian problems with moral enhancement could simply be overcome by greater societal benefits from enhancement. This would outweigh the force of the above objections. However, this assumes that consequentialist forms of punishments are justified. On the contrary, it is far from clear that current punitive laws are appropriate if social benefit was the sole goal of imposing punishment. In particular, it matters little on the consequentialist picture whether a given punished individual is actually guilty; what is most important is that the individual is perceived to be guilty. This would justify a legal system that routinely punishes innocents, so long as that has enough effect on the deterrence of wrongful activity.

To combat this, we could insert consequentialist considerations against punishing the innocent into the calculus. But if this modified consequentialist picture is to be taken seriously, it would be most sensible to severely limit the amount of punishment that is meted out. Likely, only restraint of very dangerous individuals (on the model of incarceration of criminally insane individuals who, though not responsible for their actions, nevertheless need to be kept away from society to prevent great harm) would be justified. Thus, on this model, punishments qua moral enhancement would indeed have to be severely curtailed, just as the above arguments stress great limitation on general moral enhancements.

Alternatively, one might think punishments are justified because wrongdoers deserve it. There is not space here to critically engage with the force of these arguments; suffice it to say that, if such desert-based arguments are accepted, then punishments can be justified even though most forms of direct moral enhancement are not. This is the case because retributive punishment is not actually an example of moral enhancement. The aim of a punishment is no longer to improve moral activity; rather, it is to ensure that wrongdoers get their just deserts. Conformity is not the goal of the retributivist; proliferation of active dissent is perfectly compatible with desert-based punishment.

Punishment, then, either should be curtailed significantly like other forms of moral enhancement (on the consequentialist model), or permitted and recognized as not at all a form of moral enhancement (on the retributivist model). The former option might seem radical, but I think it is not so radical as allowing the widespread punishment of innocents. The above arguments against moral enhancement, then, are not under threat of overly absurd implications in either the realm of argumentation or punishment.

**Preserving dissent**

A proponent of moral enhancement might indeed take that advice and suggest a program of moral enhancement that preserves moral dissent. This could be accomplished in a number of ways. One would be to only enhance a subset of society, leaving the rest to dissent as necessary. Another would be to limit the domains of enhancement, leaving disagreement over a large set of issues untouched. A third would be to have widespread moral enhancement be accompanied by inhibition to conformity and promotion of reasoning ability and deliberative tendencies. Do these compromises alleviate the concerns raised above? I will consider each in turn.

*Enhancing a subset of society*

The above arguments assumed that moral enhancement was targeted at all members of a society. However, that need not be the case. If only, say, half or three-quarters of

society were morally enhanced, the remaining dissenters might be sufficient to ensure progress. This could help avoid the risk of moral stagnation and allow for sincere deliberation and discussion between the non-enhanced and the enhanced. The threat to individuality (among the subset of the morally enhanced) would be unaffected, but perhaps that on its own would not be sufficient to outweigh the benefits of moral enhancement.

This solution does indeed partially mitigate the harms of moral enhancement – but only partially. Insofar as a large amount (but not all, or even almost all) of disagreement is stamped out, moral progress becomes that much more difficult. Indeed, if a majority becomes morally enhanced in a democracy, they could (and probably would, given their moral unity) easily form a political party and take control of their country in perpetuity. The non-enhanced would debate them, but it is unlikely they would be swayed – they would be too unified a group (reinforcing each other's beliefs) and too confident in their moral intuitions. This capture also has disturbing implications for democracy – the non-enhanced would be essentially disenfranchised, unable to have their opinions represented because they would be constantly out-voted by the unified, morally enhanced majority. This result could even occur when a minority were enhanced; political parties often form around core, minority 'bases', and the more unified that base is, the more influence they would have over that party (they would be better able to form coalitions and agree to throw their support behind individual platforms and/or candidates).

Additionally, having a subset of the population become morally enhanced would essentially be recreating the problem of the absence of moral disagreement on a smaller scale. At least in modern societies, people tend to form associations (political or otherwise) with others with whom they agree. Widespread moral enhancement would lead to the creation of these subcultures, which would problematically engage in a form of groupthink. Their agreement would lead to a lack of critical reflection on the issues, such that even if the rest of

society manages to morally progress, they could not. And in addition, they could not have truly sincere moral deliberations amongst themselves – likely to be their primary interlocutors – favoring instead agreement and self-reinforcement, as unified groups tend to do.

A targeted subset for moral enhancement could be more tenable. We can imagine a policy whereby convicted criminals could, as a condition of their parole, undergo moral enhancement intended to reduce recidivism. The limited size makes suppression of disagreement among them not so dangerous to society; the choice to enhance is at least somewhat voluntary (they can decline parole); and it is not clear how much deliberation goes on among criminals in the first place. We should nevertheless be very careful with such solutions; insofar as they do inhibit a criminal's individuality and freedom of thought, it should properly be seen as a form of punishment, and must thus meet strict standards of proportionality. It may in fact rise to the level of cruel and unusual punishment to suppress dissent among convicted criminals in this way, though that will of course depend on one's understanding of the standards of punishment. But in any event, this solution would be a far cry from truly widespread moral enhancement that some propose.

*Enhancing a limited number of domains*

Instead of targeting a subset of society, one might target a subset of moral domains. This is already implied by proponents of moral enhancement who advocate enhancing particular traits that people generally agree are valuable. Disagreement and deliberation over a wide swath of issues would remain and moral progress could be made (moral philosophers

would not be put out of work); it would only be on that subset of traits that unity would emerge.

Such a response, however, misses the force of the objections raised above. Recall that Mill was not objecting to social policies that affected every domain of thought, but indeed specific aspects – religious precepts in particular. Suppressing dissent in one domain is not permissible simply because there are a variety of other domains that remain unsuppressed. Trying to suppress dissent in one area is in itself objectionable; true, it is not as objectionable as enhancing a wide swath of moral beliefs, motives and behaviors, but objectionable enough to warrant rejection of such suppression.

In addition, it is not entirely clear one can so easily isolate moral domains and ensure suppression of dissent in one area will not lead to suppression of dissent in other areas. To illustrate, suppose we focus on morally enhancing people's altruistic tendencies. This may have the direct effect of making people more likely to give to charity or make self-sacrifices for the greater good – but it will also color other moral judgments. For instance, it could easily infect people's ideas about how just it is to impose costs on others for the greater good. 'I'm willing to make that sacrifice,' someone might say, 'why shouldn't I demand others make similar sacrifices as well?' The unity of these down-the-line implications of enhanced altruism might not be as strong as the unity surrounding altruism – but they would likely lead to greater conformity, in the direction of lessened dissent over the sacrifice of others for greater social benefit. Indeed, the enhanced traits (since there is general unity around them but not others) might form a new set of 'core principles' from which people attempt to derive most of their other moral ideas. Yet that growing consensus comes at the cost of suppressing

the sort of dissent that would lead people to challenge and critique our ideas surrounding the nature and value of altruism.[56]

*Enhancing deliberation and reasoning simultaneously*

I have argued that moral enhancement problematically inhibits moral disagreement insofar as it will prevent necessary challenges to conventional wisdom, reduce people's tendency to reason and deliberate, and inhibit our individuality. The proponent of moral enhancement could take these problems on board, and seek to mitigate them directly. While embarking on a program of moral enhancement, they would simultaneously reduce people's tendency to conform and improve their critical thinking abilities. Such shifts should help prevent moral stagnation (people would be less willing to conform and more willing to critique even their enhanced ideas), inhibition of deliberation (people would be more willing and able to engage in sincere deliberation) and inhibition of individuality (with less conformity and more reasoning, people would be in a better place to think and act based on their own will and not others'). According to this proposal, moral disagreement would be preserved even though we embarked on moral enhancement.

This is an attractive response, but it is ultimately either ineffective or self-defeating as a defense of direct moral enhancement. The relevant question is, what would the effect be on moral disagreement? If these ancillary enhancements do indeed lead to a reduction of moral disagreement, then the basic problems remain – moral progress would be more difficult,

[56] In Chapter 5, we saw that there is a difficulty in determining the appropriate mean and balance of different traits. This section illustrates that a similar problem emerges for the value of disagreement – it is important to preserve reasoned disagreement and deliberation over how traits should be valued and balanced against one another. Widespread moral enhancement runs the risk of suppressing such disagreement, to the detriment of progress, reasoning and individuality.

deliberation more pointless and individuality inhibited. They would be lessened somewhat, but still present. And if the ancillary enhancements do not lead to a reduction in moral disagreements, we have to ask – what was the point in promoting the moral enhancements in the first place? By the enhancers' lights, without lessened disagreement, it would seem that no moral improvement had occurred.

The proponent of moral enhancement could reply that, even if levels of moral disagreement remained unchanged, there could nevertheless be moral improvement. This could occur when serious debates (over altruism, justice, etc.) remain, but our premises or assumptions have shifted. We would still vigorously disagree over ethics and policy, but would have been nudged in one direction or another. Yet this response only masks the true cost of the shift – that people no longer take as seriously whatever inclinations have been abandoned. Moreover, it is essentially forcing social progress in one direction (the direction in which society was nudged) at the expense of preventing social progress in another (the direction away from which society was nudged). Society is admittedly still able to right the ship, but given the massive shift in people's assumptions, such becomes difficult; and if enhancers were wrong and the shift was in the wrong moral direction, social progress towards the proper moral mindset becomes more feasible.

The attractive aspect of this response, though, is that it suggests an alternative route to moral enhancement: indirect enhancement. Indirect moral enhancement will be further developed in the next chapter, and it serves to indicate one final reason against widespread moral enhancement. Not only is direct moral enhancement problematic in its suppression of disagreement, but there is another, more indirect intervention available (and perhaps even more within the grasp of present biotechnology) that could have a similar effect of moral enhancements while avoiding those objections. Proponents of widespread moral enhancement, then, should turn their attention away from the sorts of direct interventions

discussed in Chapters 2 and 3 and focus on bringing about a more moral society through other, more palatable means.

# Part III: The Prospect of Indirect Moral Enhancement

In the preceding chapters, we saw how the existence and value of moral disagreement could pose problems for a program of direct moral enhancement. The plurality of views

concerning moral matters makes securing general agreement on the form moral enhancement will take difficult, and in any case problematically presupposes a position of privilege or moral superiority of the enhancer over the enhancee. These problems persist for the most part even if one focuses on areas of supposedly broad moral agreement. Moreover, we have positive reason to preserve dissent against a widespread program of moral enhancement (which would as a matter of course aim for moral conformity) for a number of reasons: it ensures moral progress, it preserves reasoning and deliberation, and it maintains individuality within society.

Even if one thinks these concerns do not speak decisively against direct moral enhancement, they should at least raise concerns and be mitigated to whatever extent possible. The following chapters will explore an alternative strategy of moral enhancement – indirect moral enhancement – that can largely avoid these objections while also promising significant moral improvement among the enhancees. Chapter 7 explains how we could avoid the problems of disagreement by identifying particular processes that reliably lead to better ideas, motives or behaviors – not in virtue of bringing about particular outcomes, but because of those processes' intrinsic features. If we can accept that our basic intuitions are generally reliable (Chapter 8), then it is possible to develop a program of indirect moral enhancement without commitment to views about the morality of particular ideas, motives or behaviors. Chapter 9 outlines how improving people's reasoning is just such an indirect, normatively-neutral approach; if the basic inputs into one's deliberation are reliable, then helping people be more coherent, knowledgeable, understanding, critical and unbiased should generally lead to more moral outcomes. And Chapter 10 extends this to behavior, arguing that reducing akrasia (acting against one's all-things-considered judgments) will also reliably lead to moral improvements.

# Chapter 7: Overview of Indirect Moral Enhancement

## Preliminaries

Recall the structure of indirect moral enhancements provided in the introduction: An indirect moral enhancement is designed to make people more reliably produce the morally correct ideas, motives and/or actions without necessarily committing to the content of those ideas, motives and/or actions. It will be useful to explicate some of the features of this form of enhancement before turning to its advantages over direct moral enhancement.

The output for indirect moral enhancement is reliability, rather than actuality.[57] One cannot expect indirect moral enhancements to lead to moral improvements in each and every case. This is due to the indirectness of the means, examined below and in later chapters. In the very least, though, we should generally expect a successful indirect moral enhancement to lead to moral improvement.[58] More particularly, the enhancer can identify and try to improve upon particular processes that generate more moral ideas, motives and behaviors. These improved processes will not always lead to the right outcomes. The enhancer, however, may be able to identify strong connections between certain processes and moral improvement. This focus on (likely internal) processes is not the only potential form of indirect moral enhancement, but it will be the focus here.

---

[57] The sense of reliability used here is somewhat distinct from its use in epistemology, where it is used as a condition on knowledge or justification. There is some similarity – especially with the sort of process-reliabilism defended by Goldman (1979). Elements of this sort of epistemic approaches will be useful in later chapters for delineating what processes, in particular, we can expect to lead to improvement. Still, the connection between reliability and improvement is not quite so tight – having a reliable process is not partly constitutive of or strictly necessary for moral improvement, the way one might think it is for knowledge or justification.

[58] Minimally, it should lead to morally improved ideas, motives or behaviors more often than not, or improve more individuals than it denigrates. But we should hope for more than minimal improvement – and indeed, I believe the arguments in later chapters will justify relatively high credence that a given indirect moral enhancement will lead to a moral improvement.

The clause concerning the enhancer's commitments is crucial to understanding the distinction between direct and indirect moral enhancement, and hence the advantages of the later over the former. Now, the indirect enhancer may well have particular ideas about the morally proper ideas, motives and behavior. But he or she will not be using those ideas as the basis of enhancement – his or her goal will not be to bring about specific moral outcomes. This, then, could be understood as the neutrality condition on indirect moral enhancement: the goals of indirect moral enhancement will be neutral as to the morality of particular beliefs, motives or behaviors that result from enhancement. That is not to say that indirect moral enhancement is completely neutral; it will involve commitments to substantive connections between certain processes and moral outcomes. But this limited neutrality will be enough to overcome the brunt of the objections leveled against direct moral enhancement in the preceding chapters. One downside of this approach, discussed below, is that it will not allow the enhancer to directly confirm whether or not a given enhancement has led to a moral improvement. The expectation of good effects will have to rely heavily on more theoretical underpinnings of the connection between certain alterations and morally improved ideas, motives and behaviors.

**Dealing with disagreement**

Direct moral enhancement ran into trouble because its imposition of moral norms on enhancees was problematic in the face of widespread disagreement over those norms, as well as risked diminishing valuable dissent. One key reason to prefer indirect moral enhancement is that it can, for the most part, avoid these objections.

By focusing on processes and remaining neutral on the morality of outcomes, indirect moral enhancement sidesteps the great majority of substantive debates and moral disagreement. Enhancers do not claim a privileged moral position – they do not say their moral views are correct and others should get in line, or even that they are better moral reasoners than the enhancees. The enhancer does identify some processes that can be improved, but this is not justified based on the idea that the enhancer is in some way better than the enhancee. This leads to a more respectful and equal relationship between the enhancer, as there is no inherent or inevitable claim of moral superiority. This should in turn make the approach more appealing and acceptable to the population at large.

It must be admitted that complete agreement over the proper way to improve processes cannot be expected. Indeed, insofar as the enhancer is making substantive claims about certain processes (along the lines of subsequent chapters), one might worry that the same concerns over disagreement raised above still emerge, just at a different level. Nevertheless, there is a crucial difference. The disagreement that indirect moral enhancers face is not over the content of morality, but rather the relation between certain processes and good outcomes. The proper reference frame, then, is not centuries-old normative debates, but rather disputes over pedagogical techniques to improve how children think about various problems. Disputes will still emerge, but there will be more hope for convergence (insofar as intractable moral precepts are not at stake) and in any case will not threaten people's core values and mores.

More importantly, widespread indirect moral enhancements promise to promote rather than inhibit the value of disagreement. Let us consider each of the three elements of the value of disagreement (as discussed in Chapter 6) in turn – fallibility, reasoning and individuality. On fallibility, the worry was that direct moral enhancement would lead to a morally stagnant population. Indirect moral enhancement, however, carries no such risk. Indirect enhancers

are not aiming at convergence on a particular set of moral ideas, motives or behaviors, so suppression of disagreement is not an inevitable outcome of a successful indirect enhancement program. Quite the contrary, improving processes can be expected to ensure moral progress. The enhancers might well admit that their own moral ideas are fallible, and expect enhancees to be able to improve upon and develop more reliable ideas. There is still the possibility that the enhancers have the wrong ideas about the most reliable processes, and in fact make the enhancees much less morally reliable. That would be a bad outcome, but the resulting individuals would not be 'stuck' in the same way as with direct moral enhancements. Indirect enhancements, after all, would not have instilled particular ideas of the good in enhancees, leaving the door open for them to challenge and modify the enhancement program.[59]

Reasoning and deliberation also come out well from a program of indirect moral enhancement. Indeed, if one focuses on reasoning processes themselves (as in Chapter 9), there is a clear way in which indirect moral enhancement promotes the importance and value of reasoning abilities. This reflects a general strategy in the enhancement debate: if one criticizes enhancement on the ground of inhibiting something of value, then one should support enhancements to that value. If we take seriously the value of reasoning, then we should support efforts like indirect moral enhancements that will promote it. And these improvements can be expected to improve the quality of deliberation, insofar as deliberating agents will be using better and more reliable processes. But even if one does not pursue

[59] This could, admittedly, lead to a downward spiral: 'enhancees' with worse processes would choose even worse alterations for the next generation, who in turn disenhance the subsequent generations. Still, there is some reason to think this spiral will not result. We have, after all, already had a form of generational enhancement via traditional education. There was some risk that we would get the pedagogy badly wrong, and inculcate worse and worse ways of thinking and reasoning with each generation – but this does not appear to have been the case, perhaps because the overarching educational principles were sound. As the standards for indirect moral enhancement will bear some relation to those sorts of reasoning processes promoted in educational contexts, we can have some confidence that a downward moral spiral will not result. And, if one did think this spiral was a prohibitive risk, it would have the counterintuitive implication that we should cease any and all education of younger generations – lest we begin a downward spiral.

reasoning-enhancement, indirect moral enhancements preserve such reasoning because they do not put the cart before the horse – that is, they do not presume the correctness of some output and alter various processes for the sake of generating that output. Instead, they focus on improving the process itself, properly justified on the grounds that such processes have features that are conducive towards moral improvement.

For similar reasons, individuality is consistent with indirect moral enhancement. Recall that direct moral enhancement involves a problematic form of will-substitution, with the enhancer's moral views being imposed on the enhancees. The neutrality of indirect moral enhancement precludes such normative imposition. The enhancee is left free to decide for themselves what the correct moral ideas, motives and behaviors are. The enhancer is, to be sure, helping the enhancee come to the right conclusions – but the enhancer's role there is one of assistant, not master. All the enhancee's ideas still must be internally generated, and we can properly describe her ideas as her own.

It might appear that another form of will-substitution is going on, though: the enhancer is substituting his or her ideas about proper, morally reliable processes for the enhancee's. This conflict is possible, and it speaks powerfully against certain sorts of coercive programs of indirect moral enhancement. But, unlike with direct moral enhancement, the conflict is not inevitable – while most people have commitments to many moral ideas, motives and behaviors, commitment to morally reliable processes will be significantly rarer. And among those who do hold considered views about processes, as noted above, we can expect more convergence over the proper processes to inculcate, as general acceptance (even with dissent over details) of current pedagogical approaches to education indicates – making imposition of the enhancer's will unnecessary. This indicates there is more room with indirect moral enhancement for rapprochement between the enhancer

and the enhancee, mitigating potential conflicts, which can preserve enhancees' autonomy and individuality.

There will be further reasons to favor indirect moral enhancements based on the (somewhat theoretical) connection between certain processes and moral outcomes. These will be discussed in more detail below. For now, though, it is enough to note that indirect moral enhancement can avoid the problems that beset direct moral enhancements, and are to be favored for at least that reason.

**Verifiability**

The preceding has suggested a significant downside to indirect moral enhancement: the enhancer would not be in a position to verify that a given intervention has, in fact, resulted in a moral improvement. Reliability of processes are not to be judged based on their propensity to bring about certain outcomes, after all. To do so would be a form of direct moral enhancement, essentially aiming at instilling particular ideas, motives or behaviors and thus running into the problems of disagreement that indirect moral enhancement is meant to avoid. The neutrality condition essentially precludes the possibility that such outcomes could be used as a basis for evaluating indirect moral enhancement.

If we cannot appeal to the morality of particular outcomes, how then are we to evaluate a program of moral enhancement? We need a 'success condition' to determine whether such a policy was effective, or there will be no way to distinguish between true indirect moral enhancement and any number of nefarious interventions.[60] One solution might

[60] It might be thought that, on my account, the success can be determined more or less a priori: if we identify theoretical connections between certain processes and moral reliability, no further evidence is needed.

be to focus entirely on design, adhering strictly to the explication of indirect moral enhancement given at the beginning of this chapter. This would, at least, rule out non-moral projects from the category of indirect moral enhancement such as those intended purely to benefit the enhancer. And, it works to distinguish direct from indirect moral enhancement. Still, it would be useful if we had some grip on when, in particular, the enhancer had succeeded in his or her aim.

In this regard, there is some allowance for verifiability of indirect moral enhancement. Insofar as the enhancer has a sense of the sorts of processes that reliably lead to improved moral ideas, motives and behavior, he or she can check whether in fact those processes have been altered in the desirable way. Evidence for this will depend on what one takes the best processes to be, but it will likely involve engagement with the enhancee and observations concerning how he or she comes to form moral ideas, motives and behaviors. These observations wouldn't involve mind-reading (which would be infeasible), but rather would be reasonable inferences from the enhancee's self-reports and perhaps certain sorts of behaviors. Furthermore, systematic tests of various processes (especially if one is focusing on improving reasoning processes) could be developed to adequately evaluate the outcomes of indirect moral enhancement.

But it may be worrisome that the enhancer cannot use the morality of the enhancee's ideas, motives and behaviors as determinants of the success of a program indirect moral enhancement. Suppose, post-enhancement, someone becomes quite prone to theft. Can we say that the enhancement has failed? Not under indirect moral enhancement, as doing so would violate the neutrality condition. And, conversely, if that person becomes incredibly kind and generous, we would not be able to say that the enhancement has been a success.

---

However, pure a priori analysis is not sufficient because the effects of any given intervention (of the sort discussed later on in Chapter 11, on implications) will be an empirical question; we need to know, in the very least, whether an intervention actually has the effect on processes that we want.

Such would presuppose a normative framework concerning kindness and generosity, which if incorporated into evaluation would lead future iterations of indirect enhancement to impose that framework on enhancees. This may be problematic, insofar as the enhancer would not be in a position to correct for what might seem to them like obviously immoral behavior, or obvious improvements. In the extreme, perhaps enhancers would be compelled to continue with the program despite ample evidence that the program is leading people to become immensely depraved, heartless and cruel.

At a certain level, this is an outcome we should be prepared to accept. Though it may lead to counterintuitive outcomes, we will have good reason (laid out in more detail in later chapters) to suppose that the outcomes are actual improvements. The processes, after all, are taken to be reliable producers of the proper moral ideas, motives and behaviors. It may well be that our current ideas are so warped by various biases and procedural failings that we simply misjudge the morality of the resultant individuals. Indeed, if those enhancees themselves judge their new ideas, motives and behaviors to be more moral, then perhaps we should take their word for it. But, we shouldn't expect quite so much convergence among enhancees, as processes rather than the ideas, motives and behaviors are being targeted. The next generation of enhancees will have to deliberate amongst themselves whether the processes were in fact reliable, and make alterations they deem appropriate. And so long as there was some procedural improvement among that group, we should expect their judgments to lead to further procedural improvements of the following generation of enhancees.

However, there may be *some* room for indirect moral enhancers to revise their views about the most reliable processes based on the outcomes of processes. Consider a process of reflective equilibrium: an enhancer believes that process X reliably leads to moral improvements. But she then discovers that process X reliably leads to what appear, to her, to be extreme immorality. Can she reasonably alter process X to avoid these outcomes, without

running afoul of the problems of disagreement? Perhaps, under a few conditions. First, potential enhancees would have to agree, ex ante, that the outcomes of an unrevised process X would be morally worse. That way, the enhancer is not imposing her view about extreme immorality on them. Second, the enhancer would have to maintain significant theoretical justification (that is, not appealing to particular outcomes) for her belief that the altered process X will lead to good outcomes. If the justification for the reliability of process X becomes determined entirely or even dominantly reliant upon the apparent morality of particular outcomes, then the threats to disagreement would loom large. And third, this revision should only occur in truly extreme cases – cases where the change seems not just bad but immensely bad. This helps protect against potential catastrophe while allowing more limited shifts away from current generations' 'comfort zones', which may well be what morality requires.

This allowance somewhat weakens the neutrality condition. To a certain extent, then, it will make indirect moral enhancement more vulnerable to the charge that dissent is being stamped out – radical dissent, but dissent nonetheless. As the conditions limit revision to a relatively narrow set of cases, the problem will not be nearly so significant as that with direct moral enhancement. Still, it is a serious cost of this approach. It builds an element of conservatism into a project of indirect moral enhancement that could impede progress and limit how people think about certain problems. Indeed, there were periods in history when views most today would consider abhorrent (e.g., racial supremacy or slavery) were widely accepted, and dissent was considered fanatical. One might worry that these conditions shut off the possibility of radical reform of such views that we today readily accept as obvious, but are in fact morally odious.

Nevertheless, these conditions do not preclude the possibility of radical reform over time. A thought experiment would be useful: suppose an indirect moral enhancer lives in a

society where slavery is considered acceptable, and abolitionism is abhorrent (and she agrees). She discovers that instilling what she (and society at large) took to be the most morally reliable process X actually leads people to become abolitionists. She then goes about making minor alterations to process X to avoid abolitionist implications. In doing so, she fulfills the three conditions: potential enhancees agree that abolitionism is abhorrent, process X is still largely determined by theoretical underpinnings, and she takes abolitionism to be an extreme case that would bring catastrophe upon society. So, she goes ahead with implementing the altered process X. This is problematic because the enhancees' altered views come about for the wrong sorts of reasons (the enhancer and society prejudged abolition to be wrong).[61] Still, it does not preclude future generations from correcting this view in the way that direct moral enhancement does. So long as the bulk of process X is justified without appeal to particular outcomes, it will remain somewhat untainted and we can have confidence it will lead to more morally reliable ideas, motives, and behavior. Over time, this will lead to a weakening of abolitionist constraints (if abolitionism is indeed morally correct) due to the pressure from the process overall. The change might take longer than an unadulterated process X, but this time-lag may be worth the cost of avoiding catastrophe.

There is a further question of whether avoiding such catastrophe is worth the cost to reasoning[62] – at least some moral ideas would be held for the wrong sorts of reasons. I suspect that it is, but there may nevertheless be practical reason not to implement the revision process I outlined above. Such catastrophe is, it seems, unlikely to result from the sorts of approaches discussed in later chapters – they involve more subtle and variegated means that

[61] One might also think it is a bad outcome because the enhances come to deny abolitionism. However, we can admit that indirect enhancement might sometimes lead to the wrong outcomes, and in any case we shouldn't be presupposing the morality of those outcomes under the present framework.

[62] This suggests some sort of substantive cost-benefit analysis that might be seen to violate the neutrality of the indirect approach. However, the argument being presently entertained is essentially that the reasons to remain neutral are outweighed by the catastrophe. I am open to this possibility, though (for reasons stated immediately after this note) think it is unlikely.

are unlikely to cause radical change, at least in the near term. That is to say, the approaches outlined below have a natural conservatism built into them that should be sufficient to avoid catastrophe. The revision process outlined above would only serve to undermine the reasoning and deliberation of enhancees, and may slow down moral progress that one hopes to bring about through indirect moral enhancement. But this case against the revision process may not be decisive, and it is more or less compatible with the particular forms of indirect moral enhancement outlined in the chapters below.

## Sentimentalist enhancement

To fully evaluate a proposal for indirect moral enhancement, we will need to know a bit more about what sort of processes we are targeting, and how we hope to evaluate them. The next chapter will discuss a broadly rationalist approach – improving reasoning processes in the hopes that this will lead to moral improvement. But this focus on reasoning might seem to overlook the importance of emotions and sentiment in morality. Indeed, sentimentalists would claim that our moral ideas and motives (and, more derivatively, our behavior) are grounded in emotion or sentiment. A natural sentimentalist thought would be that a program of moral enhancement should therefore focus on improving emotions or sentiments. I do not wish to cast doubt on sentimentalism as a viable metaethical theory here, but it is a further question of whether in fact sentimentalist indirect moral enhancement is possible. I will argue that it is not, so long as the sentimentalist focuses on improving emotion or sentiment rather than other processes such as reasoning.

### *Humean approach*

Though he did not originate the theory, David Hume can be credited with providing the classical explication and defense of sentimentalism that is still taken seriously by many sentimentalists. According to Hume, any given moral view "depends on some internal sense or feeling which nature has made universal in the species." (Hume 1758/1995, p. 172) There are at least two important features to note about Hume's account: moral ideas originate in 'sense' or 'feeling', and that these feelings are a universal feature of all in the species. This points towards the method that Hume endorses for making moral determinations: introspection. A person "needs only to enter into his own breast for a moment" (ibid, p. 173) and consider the matter to come to a moral conclusion. The results of such introspection will be universal, and can properly ground our moral ideas.

This points towards a Humean standard of moral enhancement: the universal results of introspection inform our values, and can define the standard of moral enhancement. However, it is hard to see how this method can inform indirect moral enhancement. The process of identifying the standard for enhancement often involves making substantive moral judgments about particular ideas, motives or behaviors – the neutrality condition is impossible to satisfy. A Humean could alternatively propose that such introspection actually does not result in judgments concerning the morality of particular ideas, motives or behaviors, but instead judgments concerning the non-reasoning processes behind those ideas, motives or behaviors (an indirect claim). For instance, perhaps after introspection, one finds that disgust reactions are morally reliable. The difficulty here, however, will be the claim that such results are universal. The universal nature of sentiments Hume adverts to undermines the possibility of general moral enhancement: only universally-accepted moral claims warrant acceptance (and so use as a standard for enhancement). But if they are truly

universal, then moral enhancement would be impossible – nature has already endowed people with the appropriate sentiments.

Perhaps that brief gloss on Hume is misleading. Henning Jensen (1977), in defending Hume against the charge that he cannot deal with disagreement, has offered an alternative interpretation that might allow for a Humean moral enhancement. According to Jensen, Hume's claims of universal concurrence concern not moral thought in general, but constraints on moral language. If that is correct, then we should not read Hume as providing a normative theory at all – substantive disagreement can persist, but Hume is concerned with the convergence in moral language. Still, Jensen's understanding of these constraints provides some substantive guidance. For instance: "moral judgments must be universalizable, they must concern everyone and must concern a kind of act." (Jensen 1977, p. 502) This could serve as a basis for moral enhancement, making sure that everyone conforms in practice to the universally-accepted constraints on moral language. And, for our purposes, this would count as an indirect moral enhancement: as the constraints are grounded in language, not substance, there is little risk of presupposing particular substantive judgments concerning particular ideas, motives or actions.

However, as Jensen admits, high-level disagreement over those very constraints can persist. The vast literature in philosophy of language, including moral language, should be sufficient to show that nature has not endowed all with a universal understanding of moral terms. There is still the question, then, of how an enhancer would decide which standard of moral language to rely on. Universal sentiments cannot be used, in this case, because of the disagreement. Jensen, drawing on Frankena, suggests that we just do not (and cannot) have proper epistemic access to the full conditions on moral language. (511) If that is right, though, the Humean approach will be of little use in designing a program of moral enhancement – we cannot have confidence that any given set of constraints used as a basis of

enhancement are, in fact, proper. Alternatively, one might look for sentimentalist grounds for moral enhancement other than moral language – but that would likely have to depart from Hume to a substantial degree, to avoid the preceding issues arising from universality.

*Modern approach*

In fact, most modern sentimentalists eschew the universality condition in favor of a purer focus on the sentiments themselves. D'Arms and Jacobson provide a useful analysis of the core claim of these modern sentimentalists (including Blackburn, Gibbard, McDowell and Wiggins), called the Response Dependency Thesis. The Thesis is defined as follows: "to think that X has some evaluative property Φ is to think it appropriate to feel Φ in response to X." (D'Arms and Jacobson 2000) Of course, there will be deep disagreement among modern sentimentalists concerning what, exactly, it is to think some feeling appropriate. So, figuring out those conditions would be crucial for using this approach as the basis for a program of moral enhancement. However, in this section, I would like to highlight the problems of this approach as a form of indirect moral enhancement, irrespective of what view one ultimately takes on appropriate feelings.

In its general form, modern sentimentalism might appear to offer a viable avenue for indirect moral enhancement. The processes to target would be people's feelings in response to a given phenomenon. These psychological responses can be more or less appropriate, and the standard of appropriateness (whatever it turns out to be) can be deployed by the enhancer to evaluate different programs of indirect moral enhancement. This helps avoid the problem of verifiability that appeared to plague indirect moral enhancement. The question of whether an enhancement was a success would simply be determined by the extent to which one thinks

enhancees are indeed reacting appropriately. But because this not a normative but rather a metaethical theory – concerning moral psychology and what it is to make moral claims, rather than about morally correct ideas, motives or behaviors – it looks as if the enhancer can remain substantively neutral concerning the output.

However, this appearance of neutrality is an illusion. Though sentimentalism may not in itself have any normative implications, any application of it will. The standard of appropriateness one deploys, in particular, will inevitably involve substantive commitments. To see why this is so, we can reflect on the sentimentalist approaches an enhancer could take. The sentimentalist enhancer would first have to identify what they take it would be appropriate to feel in response to a given phenomenon. Then, the enhancer would do one of two things: 1) work to ensure that people's ideas, motives and actions end up promoting or conforming to this evaluative property, or 2) work to ensure that people have the same judgment as the enhancer concerning the appropriateness of the phenomenon. Yet, either of these options involves violation of the neutrality condition. The first presupposes some good outcome and imposes this view on the enhancee. The second presupposes some particular claim about the appropriateness of certain feelings, which according to the sentimentalist is just the same thing as making a substantive judgment. The sentimentalist enhancer, then, cannot avoid making substantive judgments in designing an enhancement program and so can only end up performing direct moral enhancement.[63]

What's more, the sentimentalist form of moral enhancement could very well be even more pernicious than other forms of direct moral enhancement. Consider Jamie Dreier's

[63] This is a problem not just for sentimentalists but anyone who would target emotions as the basis for indirect moral enhancement. Any identification of good emotional processes will most likely rely on presuppositions concerning the outputs of those emotions, which will run afoul of the neutrality condition. One might try to argue on purely theoretical grounds why certain emotional processes are more morally reliable than others, but these arguments will most likely end up relying not on the emotions themselves, but the appropriateness of various influences on the emotions – such as systematic biases. If one wants to remain substantively neutral in evaluating those influences, one will most likely have to advert to standards of reasoning, which will be discussed in the next chapter.

speaker-relativism form of sentimentalism, which attempts to avoid the objection that sentimentalists cannot account for moral error and disagreement by claiming moral judgments just refer to the speaker's own sentiments. (Dreier 1990; see also Prinz 2006) If that is right, then the enhancer would be particularly disrespectful of the enhancee's individuality: the enhancer is imposing her own sentimental tendencies on the enhancee. These tendencies are not – and cannot be – justified by appeal to anything the enhancee would have reason to accept, given that the enhancee does not ex ante share the enhancer's tendencies. At best the 'enhancement' would be arbitrary and pointless. More troublingly, the enhancer would be engaging in a form of moral imperialism: bringing everyone else into line with the enhancer's way of thinking, just because the enhancer happens to hold those views.

*The way forward*

If sentimentalism as such cannot be used as the basis of a program of indirect moral enhancement, does this mean that sentimentalists should give up on the prospect of such a program entirely? Not exactly. Though the core of sentimentalism may not be a very useful tool for moral enhancement, sentimentalism is in fact perfectly compatible with other approaches – including my own. Hume may have thought all moral judgments originate in sentiments, but he readily accepted an important role for reasoning: "But in order to pave the way for such a sentiment, and give a proper discernment of its object, it is often necessary, we find, that much reasoning should precede, that nice distinctions be made, just conclusions drawn, distant comparisons formed, complicated relations examined, and general facts fixed and ascertained." (Hume 1758/1995, p. 172) Reason's role is complex and multifaceted – it is useful to understanding exactly what one's sentiments are, draw the proper implications

from then, properly come to grips with the relations between different ideas, as well as ascertain empirical facts in the matter at hand. In a similar vein, Jesse Prinz has outlined how sentimentalists can focus on extranormative values to bring about moral progress – including consistency, implicature, and empirical accuracy. (Prinz 2007)

In this way, sentimentalism is not truly at odds with moral rationalism, at least when it comes to designing a program of indirect moral enhancement. Though they may disagree over the origins and grounding of our moral ideas, rationalists and sentimentalists can agree on the important role reasoning plays in morality. What's more, that role need not presuppose any given normative framework or judgment – one can accept the importance of certain reasoning processes for morality without making claims about the morality of particular ideas, motives or behaviors.

# Chapter 8: Moral Optimism

## The Nature for Moral Optimism

Before detailing the sorts of processes that can be the subject of indirect moral improvement, I will need to defend moral optimism. This defense is necessary because my account is intended to be as substantively neutral as possible. How, then can I argue that reasoning improvement and akrasia reduction will count as moral enhancement? If the basic inputs into reasoning and judgment are themselves unreliable, then simply tinkering with reasoning and akrasia could fairly be characterized as 'garbage in, garbage out.' But if, on the contrary, those inputs are reliable (as I will argue), then we can understand how better processing of the inputs leads to moral improvement.

Moral optimism, as I use the term here, is the view that the basic normative inputs into a system of moral reasoning are morally reliable – more likely to be correct than not.[64] I will refer to these basic inputs as intuitions, though this is meant to be an ecumenical term – it will include a priori judgments, normatively-laden emotional reactions, and apparent moral perceptions. All of these intuitions are basic insofar as they are non-inferential – not inferred from some prior or more basic moral ideas. Moral optimism can be contrasted with moral pessimism (moral intuitions are more likely to be mistaken than not) and moral ambivalence (moral intuitions are just as likely to be to be correct as incorrect). Also, moral optimism (and pessimism and ambivalence) can apply personally, to one's own moral intuitions, or globally, to humanity's moral intuitions.

[64] Preston-Roedder (2013) has argued for a normative notion of moral optimism – we should be optimistic about the morality of others' ideas as well as actions, as this is a virtuous disposition and can have good down-the-line consequences. My claim, by contrast, is primarily descriptive – people just do tend to make the right moral judgments. Still, it can be seen as having the same normative indications: we should be optimistic about people's moral judgments because that's the epistemically correct position.

A key feature of moral optimism is what goes on at the point of comparative judgment – deciding between two seemingly-incompatible moral views that boil down to competing intuitions. Assuming a resolution in favor of one intuition and against another, moral optimism says the chosen intuition will most likely be correct (or, at least, more likely correct than the competing intuition). In its general form, moral optimism will not say why this is the case. Indeed, my defense will not offer a thoroughgoing defense of such grounds for moral judgment. Instead, I will try to show that moral optimism follows from other general features of realist or quasi-realist moral thought.

I will argue for moral optimism as follows. Personal moral optimism is necessary to have any justifiable realist or quasi-realist moral view. Without moral optimism, one can have no confidence in one's own moral ideas, even relatively sophisticated ones. Next, I will argue that one has no general grounds for privileging one's own intuitions over those of others. There may be grounds in particular cases or concerning the expressions of others' intuitions, but not people's internal intuitions in general. Finally, combining personal moral optimism with non-privileging of one's own intuitions will yield general moral optimism. An objection from error theory will also be considered.

*Personal moral optimism*

Let us assume, for the moment, that some form of moral realism or quasi-realism is correct. This will enable us to talk either about the correctness of moral judgments (whether they be cognitive or non-cognitive).[65] Within these confines, there are a vast array of potential substantive moral views, both systematic and particular. But, I will submit,

[65] Quasi-realists will dispute further claims about the objective truth of these judgments, but because they still want to allow the full range of first-order substantive evaluation, the arguments in this section will by and large be untouched.

adopting any substantive moral view without constant vacillation implies commitment moral optimism, even if the view is not explicitly held.

For some approaches, this commitment is relatively easy to identify. Take the popular reflective equilibrium approach. As understood by Rawls (1999; see also Daniels 1979), this is the process of going back and forth between one's particular judgments about cases, one's general principles and ones' background theories that help adjudicate the correctness of judgments and principles. When any of these conflict, one is revised to achieve greater coherence. This process continues through a nexus of judgments, principles and theories until general coherence is obtained. But what grounds deciding to hold onto, say, a background theory at the expense of a particular judgment? Further theories of proper argument or constraints could be appealed to, but this just pushes the question back to whether to accept the further theories or the particular judgment, which evidently conflict. These determinations will ultimately just come down to which position is more acceptable or plausible. And given the wide definition of intuitions given above which can include basic theoretical moral considerations, this is really just a contest of the relative attractiveness of competing intuitions.[66]

The only way someone who engages in reflective equilibrium can have confidence in the correctness of their resultant views is if adjudication between competing intuitions is morally reliable. If pessimism were true, errors would propagate: people would choose the wrong moral ideas, which would in turn be used to improperly rule out other moral ideas that were inconsistent, leading to a set of moral views that are less correct than when one started.

[66] This analysis bears some similarity to what Campbell and Kumar (2012) have recently called moral consistency reasoning, where people adjudicate not between case-based judgment, principle and theory but conflicting judgments. Moral consistency reasoning could be interpreted as a form of narrow reflective equilibrium that, unlike Rawls' wide reflective equilibrium, does not address background theories. However, in another way, it is more general: all forms of reflective equilibrium ultimately adjudicate between conflicting intuitions (whether about cases, principles or background theories), making Rawls' wide reflective equilibrium a particular species of moral consistency reasoning.

This has been referred to as the fragility of reflective equilibrium (Harman, Mason and Sinnott-Armstrong 2010), and the only way to overcome it is to hold, for one reason or another, that people will at least tend to make the right decision.[67] Later on, I will suggest that one could have confidence in at least some improvement through reflective equilibrium by having more coherent beliefs, even if we are just as likely to choose the right intuition as the wrong one. Still, this will only help improve reliability at the margins; to get the full-blooded commitment to moral views of many who endorse reflective equilibrium, stronger optimism is needed. Just as under pessimism reflective equilibrium will propagate errors, under optimism it propagates correct judgments – typical emendations will be in favor of correct judgment, and those correct judgments will themselves be used to root out other inconsistent fallacies. The result will be a moral system in which one can have great confidence.

Reflective equilibrium is not without its critics, of course. So, for instance, Brandt (1990) thinks considered judgments are not themselves reliable. But Brandt, notably, is not without his own substantive commitments that ground his favored moral theory (utilitarianism), in particular that moral systems "be accepted if and only if they survive certain tests which keep desires in touch with the real world." (Brandt 1990, 276) Even if Brandt eschews the method of reflective equilibrium, he is nevertheless committed to a series of personal emendations: accepting intuitions concerning the validity of his 'scientific' morality versus the appeal of reflective equilibrium; accepting utilitarian intuitions at the expense of non-utilitarian intuitions; rejecting intuitions in favor of credence in case-based moral judgments, instead ruling them morally irrelevant; and so on. Some of these competing intuitions may not be basic, instead flowing from other more basic commitments, but at some point we will find Brandt has some basic commitments. And in order for Brandt

[67] I remain neutral as to why this is the case. All I am arguing here is that realists and quasi-realists are committed to some story about the reliability of intuitions; I will not defend a particular story.

to be confident that his theory is correct, he must believe that he is more likely than not to have the right basic commitments.[68]

This analysis can be generalized. Any realist or quasi-realist moral account will, at some point, appeal to basic commitments that are not, either individually or as a set, flow from prior commitments. To do otherwise would risk an infinite regress of commitments – all intuitions flowing from ever more basic levels of intuitions. This is implausible on three levels. First, it is not psychologically possible to hold an infinite regress of intuitions; practically speaking, everyone has to stop at some point. Second, the structure of most people's moral ideas does not incorporate infinite regress; moral systems typically do not rely on regressive justification, and it is quite common for people to espouse basic moral judgments that do not flow from prior commitments. And third, even if one did espouse such a regress, it would mean a groundless moral theory. Full justification would be impossible, and without the possibility of such justification it is difficult to see how one could continue to hold realist views. Put another way, on such a view there would be no actual moral truths, only inferred claims. This makes a realist position untenable. It also poses an analogous problem for a quasi-realist; insofar as quasi-realists want to be able to make all the first-order claims that realists make, they would be unable to do so because the infinite regress never settles on any pure first-order claims that realists espouse.

Realists and quasi-realists, then, must have at least some set of basic moral intuitions from which other views flow. But insofar as, like Brandt, they have confidence in those intuitions and their resultant views, they must have confidence that any given moral intuition is more right than not. This is in many ways an inevitable outcome of *any* moral

[68] An alternative critique of reflective equilibrium is that, in the face of internal conflict, we are not warranted in revising any given view – rather, we should simply suspend judgment. While this approach has the virtue (like reflective equilibrium) of ensuring consistency, it lacks another important feature – the generation of more sound moral ideas. Instead, the suspension approach would (given the frequency of internal inconsistency) would lead to widespread moral skepticism, undermining any attempt to articulate and defend a substantive ethical theory (or any particular ethical view, for that matter).

determination – one believes oneself to be correct, and so one holds that one chose the right moral judgment and threw out the incorrect one. And if this is one's general attitude, then it implies one is personally committed to generally being correct in adjudicating between moral intuitions. This still leaves room for fallibilism, of course – one may have doubts about one's determinations, leading to further reflection and revision. But in order to have overall confidence in one's views (and thus entertain realist or quasi-realist moral positions), one must think that at least one is more likely to get it right than wrong. In this way, realists and quasi-realists are generally committed to personal moral optimism.[69]

*Privileging personal intuitions*

To demonstrate that we should think our own intuitions are just as reliable as others' intuitions, I will return to the issue of peer disagreement briefly discussed in Chapter 5. The issue of peer disagreement is usually framed in the following terms: suppose someone is your epistemic peer. This means they are just as empirically informed as you are, just as rational, just as consistent, and more generally just as likely to get the right answer as you are (in the present context, on moral matters). Suppose further you disagree on some moral matter. Shouldn't you substantially revise your prior credence in the moral matter, at the extreme adopting a stance roughly halfway between your prior view and that of your epistemic peer? After all, what ground could you have for privileging your own moral ideas over that of your peers, once you accept they are in the same epistemic position?

Indeed, given that we hold a number of basic moral intuitions, it is hard to see what reason we have for thinking our own moral intuitions are more reliable than others'. The correctness of moral views, in the first place, are not verifiable – not without further

[69] Later on, I will argue that anti-realists, in particular error theorists, can also endorse optimism, and in any case have reason to accept the overall connection between moral judgment and reasoning endorsed in this chapter.

normative assumptions, anyway. So we can't use some external grounds for privileging our own moral intuitions. One could posit that one has a special moral faculty that others lack, or some special metaphysical access to the moral truths. But such special privileges seem absurd and unjustified – where are these special abilities supposed to come from, and moreover what evidence is there that others lack them?[70] And the mere fact that you disagree does not in itself provide any warrant.

Several solutions have been proposed to avoid radical skepticism in the face of disagreement and allow one to retain one's own views in the face of disagreement. While some of them are plausible, none successfully refute the present claim that one's own internal intuitions are in themselves generally as reliable as those of others. One solution is to deny that epistemic peers are all that common. King (2012) points out it's quite rare for people to have just the same evidence and the same epistemic capabilities. But that only affects how one judges the epistemic apparatus around intuitions (including the various facets of reasoning ability discussed below), rather than the intuitions themselves. My own view is that this can be pressed even further: even if you found someone with just as adequate evidence and epistemic abilities, external reports of intuitions may not adequately represent people's internal intuitions themselves. They might misstate them, misinterpret them, misapply them or otherwise miscommunicate the actual nature of the intuition. In general, then, one has good positive reason to discount the professed intuitions of one's interlocutors. By contrast, one has direct access to one's own intuitions and be much more confident in their nature, applicability, force and so on, and on those grounds can privilege one's own intuitions over the mere reports of others'. Still, this leaves open the notion that people's intuitions as internally represented are equally reliable.

[70] In particular cases, one might be able to claim one's interlocutor suffers from a specific mental defect impairing his or her judgments. However, it is unlikely that one can generalize such defects to the whole population.

Alternatively, one could conceptualize intuitions themselves as evidence. This move (indirectly suggested by Bogardus 2009) would again mean that one lacks epistemic peerhood in the case of disagreement: different parties have different evidence, putting them in different epistemic positions. While this move is metaethically controversial, it does technically mean the equal-evidence requirement typically attached to epistemic peers is not met when intuitions conflict. This technical solution might help avoid skeptical implications, but again the issue of internal intuitions' reliability remains. Why take one's own evidence, in itself, as more credible than that of others'? The mere fact that it is from oneself seems at best irrelevant, and potentially a sort of self-serving egocentric bias.

A different solution from Elga (2007) poses more problems for my equal-reliability view. According to Elga, we can discount particular moral claims of epistemic peers that conflict with our own on the basis of further disagreement over a wide nexus of moral issues. That further disagreement essentially rules out the agent in question as reliable. If correct, then one could justify taking one's own intuitions as generally credible but rule out some others on the basis of widespread disagreement; to the extent that some portion of the population disagrees over a wide range of issues, their views will not be reliable. This might not pose a huge problem for the equal-reliability view for contingent reasons; as McGrath (2008) points out, there is in fact a good amount of agreement over a wide range of moral issues like the wrongness of murder and theft or the goodness of honesty or charity. So Elga's agent would be justified in accepting general equal moral reliability of intuitions, on the basis of this general agreement, even if some individuals' reliability can be discounted.

There are some further problems for Elga's view. Simpson (2013) persuasively argues that using cluster-disagreement looks suspiciously dogmatic and self-serving. Privileging one's own cluster of moral views poses the same justificatory problems as privileging personal individual moral views. The proper response to wide disagreement

among true peers would be to change one's credence in that wide set of moral views. In response, one could interpret Elga as not actually privileging one's wide array of moral views, but instead using that as a contextually-relevant standard by which to evaluate individual claims. This would avoid general dogmatism and some of Simpson's critique, but notably leaves the equal-reliability view untouched. The equal reliability view has scope over a wide array of moral claims, indeed all of a person's moral views. The interpretation of Elga as only addressing individual claims would mean he could easily accept more interpersonal neutrality concerning the reliability of moral claims (or at least, the intuitions backing those claims) more generally.

Another challenge comes from Wedgwood. In criticizing Alan Gibbard's (1990) defense of placing fundamental trust in everyone's beliefs, Wedgwood recognizes that moral disagreement poses a serious problem for non-relativist morality: the lack of an independent check on our own intuitive moral claims indicates we should adopt a principle that gives equal weight to our own intuitions as those of others (at least when we can be confident the intuitions of others are honestly reported and not the result of irrationality or empirical error). But, he rejects that principle and claims instead we can reasonably reject other people's intuitions when they conflict with ours. This solution relies on accepting a form of internalism in which what it is rational to believe depends solely on internal mental states and processes. One's own intuitions come directly from one's own internal mental states and processes, while others' intuitions does not; that difference makes it permissible to privilege your own intuitions over others'. Interestingly, this implies we are also rationally permitted to reject others' claims in certain non-moral cases like differing memories; this might be considered a cost, but it is not nearly so great a cost as moral skepticism. If correct, this would clearly imply the rejection of the equal reliability view.

While Wedgwood's solution is elegant, it remains somewhat unclear why, even within an internalist framework, self-originating mental states and processes (like one's own moral intuitions) are to be privileged over externally-originating ones (like other people's intuitions). Both are mental states and processes, and internalism is neutral about the internal vs. external provenances of internal mental states and processes. As Roger Crisp (2011) observes, indirectness of others' intuitions does not license privileging our own intuitions; it just licenses us to make claims concerning what appears to us to be the case. Perhaps some ancillary internalist epistemic principle is at play, but it is not clear what that principle is or why we should accept it. We may well only be able to base our moral ideas on (ultimately) internal intuitions of some sort, but this will in itself give us no reason to discount the reliability of others' internal intuitions.

*General Moral Optimism*

The claims of the two preceding sections can be combined to form a sound, valid argument for general moral optimism. The argument is quite simple:

P1: One's own moral intuitions are generally reliable

P2: One's own moral intuitions are as reliable as everyone else's moral intuitions

C: Everyone's moral intuitions are generally reliable

This conclusion is just a reformulation of general moral optimism. It is a positive argument, albeit one that relies on realist or quasi-realist assumptions. Also, it does not imply general skepticism in the face of disagreement – people's internal, basic moral intuitions are all generally reliable, but how they report those or deploy them in reasoning processes may not be. Indeed, the flaws of that deployment will motivate much of the following chapter. Indirect moral enhancement can be brought about by making the reasoning apparatus

surrounding those intuitions more robust, helping ensure that the outputs (in terms of non-basic moral judgments about what to do or believe) are themselves as reliable as possible. The reliability of these outputs will, as a matter of necessity, be limited by the reliability of the intuitionistic inputs, so we cannot hope for complete certainty. Even with perfect reasoning abilities, errors will remain – as will divergent moral opinions. But as argued in Chapter 6, this is not at all a bad thing – quite the opposite, that diversity is something that should be welcomed.

General moral optimism will play a more direct role in some sections than others. In particular, the sections on logical competence and avoidance of biases will rely significantly on it. It is not strictly necessary for the general point to be valid – that improved reasoning will lead to improved moral judgment. However, general moral optimism makes that connection much tighter and stronger. Importantly, this connection is made without commitment to particular moral views. Notice how, in the preceding discussion of moral optimism, the case was not based on any particular moral claims. This is in keeping with the moral neutrality demanded by indirect moral enhancement. I will strive to maintain that neutrality as much as possible going forward.

This view, while having clear implications for the prospect of moral enhancement via reasoning improvements, should not be overstated. It might be argued that, if everyone has a better-than-50% chance of being right on any given moral matter, the best way to adjudicate moral matters is to accept whatever a majority of people think. This would be a moral form of the Condorcet jury theorem, which can justify high confidence in a group's assessment even if any individual person's chance of being right is barely above 50/50. However, the above account applies only to basic moral intuitions – it makes no guarantees that inferences or deductions from those basic intuitions (which play a substantial role in moral thought and discourse) will be equally reliable. It may turn out that people's reasoning abilities, as they

stand, are not particularly reliable and so we cannot use simple majoritarianism to resolve moral debates.[71] This still leaves room for (and indeed emphasizes the importance of) my own account of moral enhancement that focuses on improving those reasoning abilities.

Additionally, it was noted above in the context of peer disagreement that general moral optimism is a theory about internal intuitions, not the expression of those intuitions. It can often be difficult to isolate and identify the content of others' intuitions, as well as establish whether they are actually basic or in fact inferred. For instance, it is quite popular these days to conduct surveys on variants of the trolley problem. But even determining that, say, a majority think it's wrong to push a person onto the train tracks, killing the one person but saving five others from being struck by the train, tells you little about their intuitions. Was their judgment really basic, or the result of some prior principle? Are different people deploying the same concept of wrongness? Do they have the same understanding of the facts of the case? Sophisticated studies may be able to isolate some of these issues, but without a mind-reading device we can never get a full picture of other people's intuitions. [72]

This inability to adequately grasp other people's intuitions may pose some problems for fruitful dialectical engagement, but it lends ancillary support to the indirect approach to moral enhancement. Unlike direct moral enhancement, indirect enhancement does not rely on the identification of particular correct moral ideas. By focusing on processes like reasoning, we need not worry about potentially misinterpreting others' intuitions. As long as

[71] This analysis might be seen as in tension with the claims of Chapter 5, which cast doubt on the possibility of true consensus to guide moral enhancement. However, reliability does not imply consensus – it may be that people's intuitions are right 60% of the time, making them reliable, but disagreement persists to the extent that (say) 60% of the population supports one intuition and the remainder its contrary. Additionally, the concerns just adduced – the problems of comprehending and communicating intuitions – will make judgment of true consensus difficult.

[72] If one could guarantee that reasoning processes of all agents were completely adequate and we had perfect epistemic access to people's basic intuitions, then majoritarianism might be a good procedure for determining some moral truths. However, I am doubtful that we could feasibly achieve such perfection.

the processes are reliable and general moral optimism holds,[73] we can trust in people's own internal judgments without having proper knowledge of the content of those judgments. This is less of an advantage over direct moral enhancement when it comes to generating improvements in behavior (we don't have the same sort of barriers to observation of others' behaviors as we do observation of their judgments), but in the very least provides support for the indirect approach in improving moral ideas and motives (insofar as the correctness of motives is at least partly reliant on the underlying judgments).

*Error theory and pessimism*

It could be pointed out, quite fairly, that the preceding already involved substantial, non-neutral commitments. Recall that the argument for optimism assumed either realism or quasi-realism. Those positions might not be strictly first-order substantial claims, but they are certainly controversial. An anti-realist could press that, in fact, ambivalence or pessimism are correct due to the structure of their metaethical views.

One metaethical view, in particular, would appear to directly imply pessimism: error theory. According to error theory, people are systematically mistaken in their moral views – mistakenly attributing moral properties where there are none. The view is not necessarily nihilistic, but in the very least poses serious problems for any defense of moral knowledge. A well-known of error theory comes from J.L. Mackie (1977), and puts pressure on the reliance on intuitions outlined above. Mackie thinks that the existence of the sort of disagreement discussed in Chapter 4 gives us reason to be skeptical of objective moral truths; if they really were objective, we should expect more convergence than has actually occurred. Perhaps

[73] And note that the above account of general moral optimism, like indirect moral enhancement in general, did not rely on the capacity to correctly identify or interpret others' intuitions – only confidence in one's own intuitions and the recognition that one's own intuitions are just as likely to be right as others'.

more famous is the argument from queerness, which claims that moral realism (implicit in most people's moral judgments) actually implies a strange array of metaphysical qualities unlike anything else, which we are supposed to have special access to. Mackie finds both the existence of such qualities and our special access to them (should the metaphysical entities exist) implausible. The result is that our everyday moral judgments, which imply such a metaphysics, are wrong-headed.

If that is correct, a form of pessimism seems to follow. Many moral views are positive claims in one way or another – something is good or bad, right or wrong. All those views turn out to be mistaken. A slew of negative claims remain untouched – claims that things are neither good nor bad, neither right nor wrong. But even if the vast majority of people's intuitions were of the negative sort, pessimism about the positive intuitions would doom a whole network of moral ideas. Those positive intuitions would ground a network of more complex derivative judgments, all of which would be on epistemically unsound grounds. The errors would propagate, with more and more false positive moral belief cropping up. This would be a form of moral disenhancement, insofar as people would come to have a wide array of mistaken ideas concerning morality.

That is indeed a form of pessimism, but nevertheless it actually does support the connection between reasoning and morality. Earlier in Chapter 4, there was a brief discussion of what moral enhancement for the error theorist might look like: essentially making people more likely to recognize that their moral views are mistaken. This could be done directly by getting rid of moral intuitions, but also more indirectly. Such an indirect method, I submit, would look quite similar to the present proposal. Error theorists are, after all, presenting arguments that they take to be reasonable, logically sound and appealing. What's more, they will notice that not all disagreements come from people clinging to their views – disagreement may come from a variety of reasoning failures, including logical errors,

empirical mistakes, unwillingness to critically reflect, conceptual misunderstandings or self-serving biases. These, of course, are the very dimensions of reasoning that I propose to discuss. So even if one is an error theorist and rejects this notion of optimism, one should be sympathetic towards the likelihood of improvements in these various factors to lead to improvement in moral judgment (as judged by the error theorist's standards, which should accept the importance of sound reasoning). Error theorists might become more critical of this proposal if, post-enhancement, people became even more dubious of error theory. But this would be the wrong sort of reaction – instead, the fact that improved reasoning leads to more rejection of error theory would be reason to reject error theory, rather than the reliability of the reasoning process.

Additionally, while error theorists like Mackie need not be quasi-realists, error theory is strictly speaking compatible with quasi-realism. As quasi-realists do not posit the existence of moral properties or facts, there is no issue of such properties or facts tracking our evolutionarily-induced intuitions. This observation helps clarify the exact nature of error theory pessimism: it is pessimism concerning the metaethical claims of realists (and perhaps implied by folk morality), not in itself pessimism concerning the first-order substantive claims that consume most moral discourse. Technically, error theory-based moral enhancement would then only be seeking out revision in people's metaethical views about real moral facts or properties (or knowledge thereof). It need not affect people's first-order judgments and in that way maintain a quasi-realist approach that is, independent of strong realist claims, actually optimistic about first-order judgments for reasons similar to those adduced above for quasi-realists – that optimism is cashed out in terms of some sort of moral correctness not tied to moral realism.

There is some further reason for error theorists to endorse a stronger form of optimism that encompasses metaethical views as well. When it comes to their own personal moral

views, convinced error theorists will inevitably be optimistic. Taking Mackie's view, the error theorist will have some background intuitions concerning plausible moral ontology and ethical faculties. This, after Mackie's analysis of the metaphysics of morality, will lead to counter-intuitive results. Mackie evidently decides to discard his folk intuitions about moral realism in favor of the metaethical intuitions driving error theory. This is fair enough, but for the error theorist to be confident in his views, he has to believe that determination was reliable. And because his views will (if substantively revisionary) require abandonment of a wide range of moral intuitions, in each case of competing first-order and metaethical intuitions he must be confident she will make the right choice. This only makes sense if he has a general confidence that, when such intuitions collide, he will reliably choose the correct one. So, he is committed to personal moral optimism. And since error theory does not provide basis for privileging one's own intuitions over those of others, he will in fact be committed to general moral optimism. [74]

So, either by noting the commitment of error theorists to sound reasoning processes, suggesting they endorse a form of quasi-realism, or pointing out the optimism inherent in it, error theory is not so hostile to optimism itself or, more importantly for my purposes, the connection between sound moral judgment and reasoning processes. Despite earlier appearances, then, this account does not actually rest on realist/quasi-realist assumptions, maintaining overall neutrality at the metaethical level – even if the account is perhaps more appealing to a strong realist.

---

[74] A defender of moral pessimism (or ambivalence) on metaethical grounds might claim that metaethical intuitions are reliable while first-order normative intuitions are not. This would be internally consistent, and would (in concert with later chapters of this work) lend support to a purely metaethical project of moral enhancement. Yet the reliability distinction seems baseless. Why, exactly, should we take there to be moral reliability at the metaethical level but not the first-order level? What sort of basis could there be that applies to the former not the latter? The burden of proof is on the pessimist; without a convincing argument for such a distinction, we should take their reliability to stand or fall together. The fact that error theorists like Mackie implicitly accept the reliability of metaethical judgments means (absent an argument otherwise, which has not at present been provided) they should accept the reliability of their first-order judgments as well. This puts serious pressure on the soundness of error theory itself, but I shall not pursue this implication here.

# Chapter 9: Reasoning

**Preliminaries**

We are now in a position to elaborate in more detail how to engage in indirect moral enhancement. Recall that we are looking for processes that will reliably lead to better moral ideas, motives and actions, whose reliability is not determined by the enhancer's view of the correct moral ideas, motives and actions. The defense of such a process must be (for the most part) substantively neutral on the morality of particular ideas, motives or actions. This is a tall order, but as indicated at various points above, I believe it can be met by focusing on improvements in people's reasoning abilities. This does not mean improvements to reasoning processes are the only defensible means of moral enhancement, but it is a focus that will avoid the objections of earlier chapters. Moreover, even if one is unconvinced by those objections, there are strong positive reasons to believe improvements in reasoning can reliably lead to moral improvement.[75]

Reasoning ability is an admittedly vague and nebulous concept. But what follows does not depend on a particularly well-defined and unified concept. Instead, I will identify several key components or aspects of reasoning ability that are both largely substantively neutral and can reliably lead to improved moral thought, motives and action.[76] These are: logical competence, empirical competence, formulating and articulating ideas, critical analysis, and avoidance of bias. Because I want to avoid making particular substantive moral claims, the connection between each aspect of reasoning and morality will be largely conceptual – relying on what I describe below as moral optimism. In addition, the

[75] Some of what is proposed here is compatible with third-party imposition of what Douglas (2014) has called brute conformity enhancements, in particular those brute enhancements that target reasoning and deliberative processes. So, a pharmaceutical intervention that improves reasoning capacities might be directly altering mental states, but it counts as an indirect enhancement insofar as it is not trying to get people to hold particular moral ideas, but help them think through those ideas more effectively.

[76] This list was selected not on the basis of a systematic analysis of reasoning, but the extent to which each – on reflection – fulfilled the neutrality and reliability criteria.

connections will be – at this stage – directly between morality and moral thought itself (including thoughts about what are the morally right beliefs, motives or actions), and only indirectly to motives and actions themselves. The next chapter on akrasia will deal more with the bridge between moral thought and action. For now, though, I will work under the assumption that moral thought itself – especially moral thought about what is the right thing to do – can have at least some direct impact on what people do, or what they are motivated to do, such that more morally correct ideas will generally contribute towards more morally correct motives and actions. As argued previously, even sentimentalists should be able to accept this connection.

This conceptual focus, though, should not be interpreted as an exclusive focus on theoretical reasoning. Theoretical reasoning can be understood as reasoning over what to think, contrasted with practical reasoning over what to do. (Harman, Mason and Sinnott-Armstrong 2010) While the latter may be more connected to action, they are both only directly concerned with internal activity. In this way, the following discussion will be primarily addressing internal mental processes, though external influences on those processes (such as argument and deliberation, not to mention the prospect of biomedical interventions) will of course be relevant.

It might appear odd that I am focusing on reasoning ability rather than rationality, which has received an abundance of attention in the philosophical literature. In fact, one could argue that reasoning ability as I understand it just is a form of rationality. But notions of rationality are contentious. If rationality is a sort of reasons-responsiveness, the ability to recognize the normative features of the world or the reasons that there are (Raz 1999), then the connection between rationality and proper moral ideas will be tight and necessary. But if rationality involves a response to experience, rather than reason, (Audi 2001) that connection becomes looser. And a more purely procedural account of rationality, as a set of decision-

theoretic rules (Nozick 1995) or, more minimally, internal coherence (Hinchman 2013) might naturally be taken to have no connection at all to morality. (Zangwill 2012) I do not wish to ignore these issues – indeed, I will defend the connection between coherence and morality below – but as none of my arguments turn on the controversial nature of rationality, I will avoid focusing on rationality per se.

Furthermore, as noted in the previous chapter, nothing I say here is meant to indicate a strong proceduralist account of morality. That is, I am not arguing that morality simply consists in engaging in certain reasoning processes, or the results of those processes. Better reasoning will reliably lead to more moral ideas, motives and actions, but not necessarily so. It will not eradicate errors in inputs or in the process itself, and what's more this account is meant to largely avoid substantive commitments like proceduralist morality.

## Aspects of reasoning

What follows is an explication of different aspects of reasoning that, I hold, can contribute to the reliability of our moral ideas. This list is not meant to be completely exhaustive either of all aspects of reasoning or all indirect means of moral enhancement of judgments. It will, however, encompass a set of significant and appealing means by which to bring about indirect moral enhancement.

### *Logical competence*

Perhaps the easiest target for indirect moral enhancement would be improving logical competence. Logical competence concerns people's ability to make proper logical inferences and deductions, spot contradictions in their own beliefs and those of others, as well as

formulate arguments in a way that can highlight the true point of contention between interlocutors.

Identifying logical inconsistencies is the most important aspect of logical competence, at least for present purposes. Putting emphasis on consistency in a discussion of proper reasoning is hardly controversial. Though they differ substantially on the nature and standards of rationality, Nozick (1995) and Audi (2001) both hold logical coherence as a universal constraint on rational thought. Coherence is also the basis of the popular method of wide reflective equilibrium discussed above, and it is implicitly valued in almost all discourse in moral philosophy. This general acceptance makes perfects sense. Whatever the correct moral views are, we should expect those views to be consistent with one another. This is certainly an implication of cognitivist views that treat moral beliefs as structurally like non-moral beliefs and so similarly subject to the rules of logic. And quasi-realists will endorse this as well, insofar as the standards of correctness of first-order views they apply do not differ substantially from those of realists.

Because the importance of coherence is so widely recognized, the source of improvement will not be in the abstract recognition of that importance. Instead, we can get indirect moral enhancement by improving people's ability to identify inconsistent judgments. Logical competence comes into play insofar as it helps people identify the logical implications of their views. People may not realize their views are, taken together, jointly incoherent. One might hold, for instance, the following three views: all corrupt politicians should be punished no matter how mild the corruption; one's favorite politician is mildly corrupt; and one's favorite politician should not be punished for so mild a corruption, given all the good work she is doing. These are jointly inconsistent, as the first two views imply by *modus ponens* that one's favorite politician should be punished even for mild corruption. Something has to give – logically, one of the views must be given up. Better understanding,

implicit or explicit, of logical rules like modus ponens can help avoid these inconsistencies and force corrections.

But why should we expect those corrections to lead to improvements? Here, the above argument for general moral optimism comes into play. Logical inconsistencies like in the politician case will, at a certain point, bottom out at competing intuitions that double as logical premises or conclusions.[77] At that point, when people choose between competing intuitions, general moral optimism allows us to say with confidence that they will generally make the right choice. That is, they will generally discard the fallacious intuition and adopt the correct one. This is a form of moral enhancement, insofar as incorrect judgments are gradually replaced by correct ones. Errors will sometimes occur, but as long as the correct choices outweigh the incorrect ones (as implied by general moral optimism) the aggregate effect should be improved judgment.

In fact, one could even expect some improvement rejecting general moral optimism in favor of moral ambivalence (people are as likely to choose the correct moral intuition as the incorrect one). Returning to the politician example: suppose the individual in question chooses which judgment to abandon at random. This is more or less equivalent to moral ambivalence - we have no reason to believe that the choice will be the right or wrong one. Nevertheless, the person's epistemic situation has improved in at least one regard. Given the three positions are jointly inconsistent, the *conjunction* of all three propositions cannot be true. So, while we do not cannot say the individual's individual judgments have been improved, we do know that at least one incorrect judgment, that conjunction, has been

[77] This points to a further important role for logical competence: identifying the intuitions that ground one's moral judgments. This identification will be useful in ensuring one has properly identified actually inconsistent intuitions, as opposed to merely apparently inconsistent intuitions that may result from faulty logical reasoning.

abandoned. And the abandonment of a necessarily incorrect judgment will count as at least a small improvement.[78]

Moral improvement can also be achieved by improving people's ability to correctly identify the implications of their moral judgments. This is distinct from identifying inconsistencies insofar as, in drawing out entailments, one might never consider inconsistent views. Identifying implications is, then, in itself a positive project: adding additional moral views, rather than replacing existing ones. This can be accomplished in the same way as improving identification of inconsistencies – cultivating proper understanding of the logical rules of inference so as to better identify the implications of one's views. And like identifying consistencies, our confidence in the resultant improvements comes from general moral optimism about the originating intuitions. Reliable intuitions combined with sound logical inferences will result in additional sound judgments. These additional judgments will be crucial not just for the theoretical value of having more correct moral views, but in deciding what to do in particular novel circumstances on the basis of some prior views.

Further benefits of logical competence accrue in the context of moral discourse. Being able to frame ideas and arguments logically will improve the extent to which one can effectively communicate one's ideas, easing the way to propagation of one's ideas and arguments. Conversely, understanding the logical flow of others' thoughts help in understanding the structure of their thoughts. While proper knowledge of those ideas may be elusive, we can at least gain some better comprehension by grasping the premises that others

[78] Admittedly, if general moral ambivalence is correct, this improvement is only in the realm of moral judgments and not moral actions. Even if there is a tight connection between judgment and action, correcting conjunctions of judgments will have no impact over and above improvements in the individual judgments. This is because, while logically inconsistent judgments are possible, logically inconsistent actions are not. So, while it is possible to eradicate conjunctions of logically inconsistent judgments, there are no corresponding logically inconsistent actions to prevent and thus improve upon. And as improvement under ambivalence only applies to conjunctions and not individual judgments, it cannot lead to improvement of actions.

are working from. This helps identify actual sources of disagreement and bring out potential logical sources of error. In case that discussion does persuade one party or the other, the persuasion will then be on more solid grounds and less likely based on a misunderstanding or logical mistake.

Some moral views do not easily admit of logical analysis (for example, basic, non-inferred intuitions). Still, more complex views will require at least some logical reasoning to ensure that people make proper inferences and do not descend into moral incoherence. And insofar as we would expect people's motives and behavior to be informed by their moral judgments (more on this issue in the next chapter), logical coherence should improve the reliability of those outcomes as well.

*Empirical competence*

Logical ability is an important aspect of reasoning, but in some ways it is relatively narrow – concerning a small set of logical rules applied to evaluate the validity, not the soundness, of argument. Much education, by contrast, is focused on more general abilities – improving knowledge of science, history, society, mathematics, etc. and the skills to evaluate various claims therein. These more general skills can be referred to as empirical competence – the ability to reliably generate and evaluate empirical, non-moral claims.

At first blush, it might seem such empirical claims cannot play much role in moral judgment. Hume's famous dictum that one cannot derive an 'ought' from an 'is' severely constrains the role of empirical claims. The mere empirical fact that, say, a politician takes bribes does not on its own have any moral implications. Any moral implications come from independent moral judgments (such as, accepting bribes is wrong). And again, those

independent moral judgments will themselves ultimately be derived from basic, non-empirical intuitions.

But brief reflection reveals important ways that empirical understanding informs non-basic moral judgments. Consider the following valid moral argument:

P1: Senator Barney accepts bribes

P2: Anyone accepting bribes should be punished

C: Senator Barney should be punished

P2 and the conclusion are moral claims, and so without further elaboration are untouched by empirical concerns. However, P1 is an empirical, non-moral claim. The moral conclusion only follows if it is correct. Anyone endorsing the conclusion that Senator Barney should be punished on the basis of the above reasoning needs to have good grounds for the claim that Senator Barney accepts bribes. Some sort of evidence such as a witness of the bribery will be needed. And those evaluating such evidence will need to assess a number of factors. Is the witness reliable? How do we know what was witnessed was really a bribe? What did the briber procure? Those who are generally more competent at evaluating empirical claims will more reliably ascertain the truth of P1, and in turn make more reliable evaluations of the moral question of whether Senator Barney should be punished.

This point can be generalized. Non-basic moral judgments will often rest on arguments (or something approximating arguments) with empirical premises. General moral optimism ensures the reliability of basic moral intuitions that make up many moral premises. In combination with those basic moral intuitions that ultimately ground the moral premises of an argument, empirical competence can then bring about moral improvement through improving people's ability to effectively evaluate those premises, and in that way improve the

reliability of the moral conclusions that rely in part on such premises. This will make improvements in people's empirical competence a route for indirect moral enhancement.[79]

Empirical competence is a vaguer notion than logical competence, so some explication of what it involves will be useful. Like reasoning itself, empirical competence is an umbrella concept encompassing a number of different sub-capacities, and I will try to delineate a few (this discussion is meant to be indicative of the nature of empirical competence, not exhaustive of all ways it might be improved). One aspect is memory. Properly remembering prior personal observations will assist in judgments concerning personally-experienced events. For instance, if one personally witnessed Senator Barney taking what may be a bribe, accurate recollection of what actually occurred will be crucial in evaluating his culpability. Relatedly, remembering related facts such as whether Senator Barney gave the briber any favors or the content of others' witness statement will also aid in evaluating whether a bribe actually took place. And improving memory is relatively straightforward – it is easily testable, and already has a significant body of research supporting various means of improvement.

Another relevant capacity is knowledge of an array of facts potentially relevant to moral judgment. These might be general like laws of physics or specific like the occurrence of various historical events. The range of knowledge should be wide so it can be deployed in diverse and unexpected circumstances. In the case of Senator Barney, it may involve knowledge of what constitutes bribery. This is closely related to conceptual understanding discussed in the next subsection, though here I mean knowledge of non-moral facts, as opposed to understanding of moral concepts. It also has some relation to memory, insofar as part of having knowledge of some subject involves the ability to bring to mind previously-entertained beliefs. Still, it goes beyond mere memory by requiring further conditions of

[79] A similar suggestion has been made by Harris (2011), though his analysis is relatively vague and brief.

understanding that allow people to properly appreciate and deploy the relevant facts. Imparting such empirical knowledge is arguably the primary goal of much education. Education, then, will have the welcome side-effect of being an indirect moral enhancement by equipping students with an adequate knowledge base with which to evaluate a wide range of different empirical claims that play a constitutive role in some moral arguments.

There are limits to the amount of knowledge we can reasonably expect people to personally accrue. For that reason, an important aspect of empirical competence is the identification of experts who do possess the relevant knowledge and whose conclusions can be used as a basis for evaluating empirical premises. By identifying reliable experts, one can make more sound assessments of the empirical premises of moral arguments. This identification process is to be sure not straightforward, but Alvin Goldman (2001) has proposed a few ways it can be (imperfectly) managed by non-experts: evaluate experts' track records, detect deception, observe biases, assessing (to the best of one's ability) the arguments of competing experts and taking into account consensuses among experts. This is most obviously relevant in the domain of science (modern debates over the ethics of climate change crucially rely on evaluation of the empirical claims of climate scientists), but is also important in assessing witness testimony, legal advice, political punditry and a host of other potentially morally relevant empirical domains. To the extent that reliance on experts is pervasive in an age of intellectual specialization, improving on the ability to identify reliable experts will have significant dividends in improving moral reasoning.

This brief discussion will hopefully give an indication of the shape of empirical competence and how a project of indirect moral enhancement might go about improving it. Chapter 11 will address in more detail the practical programs that might be developed and promoted in order to achieve these improvements, but some implications – such as the importance of general education – should already be clear. Indirect moral enhancement need

not be its own independent project, but part of a larger system of cognitive improvement that society already engages in for pragmatic reasons.

*Conceptual moral understanding*

Another morally relevant reasoning process is conceptual understanding, more specifically understanding of moral concepts. Having what Descartes referred to as clear and distinct ideas is morally relevant in a number of ways. At the most basic level, simply possessing moral intuitions is not adequate – one needs to understand their content, strength and scope. Without a proper grasp of intuitions' content, one cannot be said to have made a proper judgment at all. Introspection gives one a leg up in adequately discerning the content of an intuition, but people could become confused or even self-deceived which will significantly interfere with the reliability of those judgments. Strength will be crucial in helping determine which of two competing intuitions to abandon, or whether a moral consideration outweighs a non-moral one. And identifying the scope of an intuition – what it applies to – is necessary to ensure it is correctly deployed.[80]

Furthermore, in order to understand the implications of a particular moral idea (say, killing is wrong), one must have some grasp of the notions involved (in this case, not just wrongness but what exactly killing constitutes). Vague and distorted ideas will lead to unreliable inferences, inducing behaviors that are not in line with someone's considered judgments. By contrast, proper understanding of an idea will clarify and make salient the proper inferences to make. In this way, conceptual understanding aids in logical competence.

[80] Some of this understanding will intersect with logical analysis (e.g., identification of inconsistent concepts). The emphasis here, though, is not on the logic per se but the structure of the concepts themselves, which can be combined with logical analysis to yield fruitful insights.

Conceptual understanding is not just important for internal thought but also external deliberation. When discussing moral ideas with others, evaluating competing moral ideas it is crucial to fully understand the various claims that are being made. Misapprehension of the concepts can easily lead to misidentification of which argument is stronger or which position more compelling. This at best makes proper discourse and engagement with others' ideas difficult; at worse people will make unreliable revisions to their moral ideas based on such misunderstandings. General moral optimism won't matter much if basic intuitions are corrupted by such interpersonal misunderstandings.

If people, by contrast, have better understandings of each other's ideas, robust discourse and debate can flourish and lead to more improvements. By gaining some understanding of the ideas of others, one can map those ideas onto one's own. This can reveal insights one had not previously considered such as implications of one's views not previously adopted, contradictions not previously noticed or questions not previously entertained. In essence, conceptual understanding allows one to at least partly externalize the reasoning process. This should not be too problematic so long as general moral optimism holds, in particular the notion that others' intuitions are just as reliable as one's own. There are of course limitations on the extent to which people can so externalize due to the gap between one's own mind and others' minds, as previously noted. But at least some progress can be made in the direction of better grasp of what others have in mind, and to the extent that such understanding can be improved, one's own reasoning process and resultant moral ideas will benefit.

*Critical Analysis*

Oftentimes, when pressed to justify its very existence in the curriculum, philosophers emphasize how their discipline can enhance critical thinking. This may seem like a self-serving defense, but it does highlight a central aspect of the reasoning process. Critical analysis, understood here, involves both the ability and the willingness to critically evaluate various competing judgments, and even more importantly to discard those judgments that are found wanting.

The ability to critically analyze differing judgments involves at least some of the capacities noted above. Logical competence helps identify logical errors in moral arguments, and conceptual understanding allows one to fully grasp the concepts to be critiqued. But critical thinking involves more than just competence and understanding. It also requires a certain degree of creativity. This creativity comes into play in a number of ways. When entertaining an idea, it allows one to come up with potential counterarguments – either to press abandonment of the idea, or strengthen it by refutation of the counterarguments. Similarly, in debate, such creativity allows one to put pressure on others' ideas with critiques that had not been previously addressed.

More positively, creativity is important for generating moral judgments more complex than mere intuitions. Ascertaining the sorts of cases that one might encounter will help one generate more general moral judgments in advance, so one is prepared when the time comes. And the ability to come up with compelling accounts to justify various claims will indubitably strengthen the appeal of those claims. So long as the intuitions feeding into those judgments are reliable, as held by general moral optimism, and one is adequately open to potential critiques, that process of forming justifications for more complex judgments should serve to ensure one's judgments are as reliable as possible.

The motivational component of critical analysis may be where most people stumble. This motivational issue is not the sort discussed in the next chapter, the motivation to act on one's considered judgments. Rather, it is a more theoretical motivation to revise one's moral ideas in the face of compelling reason to do so. To be sure, it is hard to give up one one's ideas. One becomes attached to them, personally invested in their truth. And, perhaps, some conservatism can be justified – constantly changing one's ideas can lead to interpersonal unreliability and a fragmented sense of self. But without openness to revision in the face of what one takes to be devastating flaws in one's judgments, moral enhancement through improvements in reasoning would be almost impossible.

This necessity of openness to revision for improvement can be easily illustrated. Suppose someone identifies an inconsistency between two moral intuitions. Previously, I have been assuming that something has to give – one will be abandoned, the other retained. This revision is meant to constitute a moral enhancement (given that people are reliable in making such choices, as entailed by general moral optimism). But someone could instead simply choose to live in logical contradiction. There is nothing physically stopping them from doing so (in contrast with logical contradictions of actions), and perhaps pride or personal attachment to one's own ideas makes the option of living in contraction appealing. This decision, though, comes with severe costs: it shuts off a golden opportunity for the person to undergo a moral improvement, by their own lights. Insofar as someone cares about being moral, they should be willing to make changes in such circumstances.[81]

Generally, any cases where a person refused to change (adding a new moral judgment or altering/abandoning a previously held one) after undergoing a proper reasoning process with input from one's intuitions would be missed opportunities for moral enhancement. If

[81] Openness will have the ancillary benefit of preserving dissent, whose value was defended in Chapter 6. When people are more open to critique and revision of their views, we can expect vigorous disagreement to persist in society – helping avoid stagnation, facilitate reasoning and respect individuality.

someone flatly refused to ever change their judgments, any moral improvement of those judgments would be impossible. What could possibly justify such intransigence? Perhaps if the person had reached the pinnacle of human moral thought, there would be no need for further change because improvement is impossible. But reaching such a pinnacle is not plausible, and it is much less so if someone in such a position nevertheless faces a critique of their views that they recognize as devastating.

More plausibly, someone might have a form of ambivalence or pessimism about reasoning processes. While the moral ambivalence and pessimism critiqued above pertained to basic intuitions, this ambivalence or pessimism would instead pertain to the reasoning processes. They might think reasoning processes are simply tools to rationalize pre-determined basic intuitions. This is not a straw man position, but one advocated most prominently in recent years by Haidt (2012). Perhaps, on Haidt's skeptical view, we can trust those basic intuitions, but there's no reason to think reasoning should improve on them. One might even think that any such reasoning just leads to a corruption of the basic intuitions – insofar as the intuitions are reliable, and reasoning processes are not, any alteration of basic intuitions will lead to inferior moral judgments. There may be external reasons to engage in reasoning (e.g., achieve social cohesion), but it would not be a process that could reliably generate improved moral judgments.

This pessimistic critique from the likes of Haidt is a serious challenge for my approach to moral enhancement, but it cannot be sustained without abandoning the very same notions that thinkers use to justify the view. If we give up on the reliability logical competence, it is hard to see how pessimists could make compelling arguments for their view. They could not out of hand discard inconsistencies or press implications of their views. Without reliability in empirical competence, there is no reason to trust the experimental evidence sometimes adduced for such pessimism. Without conceptual understanding, we

could not even have an adequate grasp of notions like rationalization or pessimism itself that underpin the critique. And without openness to revision in the case of devastating critiques, no one would ever abandon their folk intuitions about the reasoning process itself in favor of Haidt's rationalization view.

Haidt could reply subtle revisions to his account to avoid it becoming self-undermining. Whatever those revisions are, I submit, they will of necessity rely to some extent on preserving various aspects of the reasoning process as central to generating improved moral judgments. This is because it is impossible to make compelling arguments, as Haidt seeks to do in his own account, without them. And to the extent they rely on such reasoning processes, they can be improved by strengthening the processes. Haidt and others might still think that proper reasoning will ultimately give support to their own skepticism about reasoning over morality. Fair enough. But as long as they accept the importance of the reasoning process to generate reliable judgments, the present account of moral enhancement is untouched. After all, it is meant to remain neutral concerning substantive positions, including (to whatever extent possible) skeptical positions like Haidt's. If enhancement of moral reasoning leads to more moral skepticism, so be it – that is just reason to have more credence in moral skepticism, and in any case a project of moral enhancement should not (for reasons outlined in earlier chapters) be evaluated on the substantive content of such outputs.

*Bias*

The above facets of reasoning could be critiqued as overly abstract and theoretical. In practice, most people do not spend much time deliberating over moral judgments or formulating complex arguments. Even if people don't rely on basic intuitions, they think and form judgments quickly – too quickly for improvements in logical reasoning or critical

analysis to have much effect. This could be addressed by focusing on the moments that people do engage in considered reasoning, and trust those moments to inform less considered moral judgments later on. But this does point to another aspect of proper reasoning that plays a role even in relatively unconsidered judgments, albeit negatively: the avoidance of biases. Avoiding bias in moral judgments and decision-making may well play a larger role in indirect moral enhancement than the preceding factors, insofar as such biases can operate at all levels of thinking.

In order to defend the importance of such bias avoidance, it will be necessary to define bias. This is surprisingly tricky, as the concept has not received a great deal of focused philosophical attention. What's more, for present purposes we need a substantively neutral definition of bias – one that does not presuppose certain substantive moral ideas or standards. This neutrality is first and foremost a necessity of the indirect approach to moral enhancement. Yet even putting aside the constraints of my particular approach, a robust notion of bias will have to be substantively neutral if it is to be a concept unique from just being generally mistaken.

Consider, for example, the non-neutral understanding of bias proffered by Sunstein (2005): "error in a predictable direction." This might match standardly understood cases of bias, such as racism: people will predictably favor their own race over others (which is a moral mistake). And it seems marginally more specific than error in general. But it is still far too broad. All predictable errors become biases, which includes errors that are predictable because the flow from some background theory. So, if utilitarianism is an incorrect theory, then utilitarians could be criticized as thoroughly biased in their views, in addition to being incorrect. This is not only an implausible application of the notion of bias, but seems like a form of double-counting of errors. Both the moral error and the bias appear to count against the utilitarian, when in fact the bias just reduces to that very error.

Indeed, any attempt to define bias in terms of actual error will run into this problem. One could try define bias as a more narrow form of error so as to avoid improperly deploying it, but inevitably the concept will just collapse into a substantive claim over whatever the source of error is. Yet bias plays a more neutral role in debate. After being accused of bias, the interlocutor does not typically dispute that the bias in question is acceptable; rather, they deny possessing that bias at all. For example, it is not typical – on being accused of racism – to defend racism, but to deny that one is in fact being racist. This indicates that bias is a relatively neutral concept, or at least does not rely on substantively contentious issues. Yet if bias just reduces to a form of substantive error, we would expect many more such disputes than actually occur.

A more attractive understanding of bias along these more neutral lines comes from Nozick. Nozick, who thinks bias avoidance is an important component of rationality, suggests that first-order biases consist in the uneven application of standards. (1995, p. 103) This is neutral insofar as Nozick does not define explicitly what those standards are. Bias, on this account, is really a form of internal inconsistency. Racism, for instance, is a bias because people on the one hand accept that race should not be taken into account in some context like employment but, on the other hand, those very same people do sometimes take race into account in hiring. This advantage may be somewhat subjectivism, but it captures why bias is so universally accepted as problematic – bias occurs when people violate their own considered norms.[82]

[82] Nozick, oddly enough, applies a more narrow and substantive definition to second-order biases, that is, biases in selecting the standards for first-order bias. Such second-order biases occur when "these very standards and weights would work to the exclusion or detriment of particular groups and this motivated them to put forward these particular standards." (ibid., p. 103) It is not at all clear why Nozick only applies this motivational understanding to second-order biases and not first-order biases as well. Moreover, the second-order definition is too narrow – it excludes epistemic (rather than motivational) explanations of second-order bias, like salience, or any implicit second-order bias. It also abandons the advantages of neutrality, suggesting quite controversially that any standard motivated by exclusion or detriment of a group is biased. Correcting racism might be motivated by the detriment of whites, because in the racist mindset whites are unfairly privileged. On Nozick's

Nozick's definition, though, is still too broad. It implies that all forms of inconsistency are forms of bias, whereas the notion is generally used more narrowly. I will offer, then, a narrower definition of bias: taking factors into account in a moral judgment that, by one's own lights, are not relevant to that moral judgment. This captures the essence of what goes wrong in biases – racists are taking race into account when they should not. And, by relying not on objective standards but one's own, it maintains substantive neutrality. It does not include too much; utilitarians are not biased in applying their standards, as their standards are morally relevant. Even non-utilitarians should be able to accept this; one's considered moral theory is surely relevant to how to act in particular cases.

Arguably, this definition is overly permissive. It prevents us from criticizing as biased people who really consider some factor relevant. For instance, the thoroughgoing racist who has an internal view that whites just are morally superior to other races would not be biased in taking race into account in various moral judgments. For this reason, the account admittedly may not serve as an adequate analysis of the notion of bias as it is typically deployed. Nevertheless, it is useful because it a) is substantively neutral, thus fitting comfortably into the indirect approach to moral enhancement and b) in practice would not generate overly-permissive results very often. Such thoroughgoing racists are relatively rare in modern society; actual racism much more manifests itself as people unintentionally taking race into account even when they accept that they shouldn't.[83] Other sorts of biases are similarly uncontroversial in their standards: how you frame a question shouldn't matter to one's opinion of it; one should not hold oneself to different moral standards as that of others; one shouldn't privilege one's relations over others in the public sphere; and so on. Given

---

definition, then, correcting such racism would count as biased, which seems problematic. These factors make his second-order definition of bias unattractive, and so I will reject it.

[83] This approach still has some resources to address thoroughgoing racists. Better appreciation of empirical facts concerning races as well as more thorough understanding of the concept of race itself along with moral notions like fairness and equality could all be leveraged in an effort to morally improve the thoroughgoing racist.

general acceptance of such standards, attribution of bias will be acceptable in such cases. In any case, this approach allows for true substantive neutrality while also holding people accountable for their actions.

With this definition in mind, the link between avoiding bias and moral enhancement should be clear. With a background assumption of general moral optimism, we can be confident that when, say, an intuition over what to do in a particular case conflicts with an intuition over what factors are morally relevant (as occurs in cases of bias), people will generally make the right choice. Promoting bias avoidance can in part consist in helping people recognize such conflicts (such as by making their standards more personally salient or explicitly pointing out such standard-violations when they occur), as well as techniques that might reliably reduce instances of erroneously taking various factors into account. So, for one who takes racism to be problematic, a program of sensitization to other races may count as an indirect moral enhancement insofar as it helps people conform their specific judgments to their standards over when race can be taken into account. Developing diverse and effective strategies for bias mitigation will then be an important part of a program of indirect moral enhancement. What shape those strategies will take is outside the scope of the present dissertation, though later on I will discuss some potentially promising avenues.

These different aspects of reasoning offer a robust, if not exhaustive, account of how improving reasoning ability can bring about indirect moral enhancement. It requires some commitment to general moral optimism, but as I have argued, this follows from realist or quasi-realist views of morality that underpin most projects of moral enhancement. And the account has managed to maintain, for the most part, neutrality concerning particular substantive moral issues. The connections, moreover, are not minor but quite strong – we can

generate substantial moral improvements because of the close relation between the reliability of the reasoning process and the soundness of the moral judgments that result.

# Chapter 10: Akrasia

**The concept of akrasia**

In the preceding chapter, I argued that indirect moral enhancement could be brought about by improving people's reasoning processes. In the course of making that argument, I assumed a relatively robust connection between forming moral ideas and behaving in certain ways. That assumption, however, is somewhat problematic. Simply improving people's moral ideas is not sufficient to bring about improved motives, much less actions, because of the phenomenon known as akrasia.

Akrasia refers to intentionally acting against one's all-things-considered (ATC) judgment.[84] 'ATC judgment', in turn, refers to consciously-entertained opinions about what one has the strongest or most reason to do, taking into account all factors that are (by the agent's own lights) relevant. Often, reducing akrasia would be in someone's interest; say, a person recognizes that they need to lose weight but frequently succumbs to the temptation to eat sweets. In this chapter, it will be argued that akrasia reduction can also reliably make people more moral, at least in their behavior and possibly motives.[85] Akrasia reduction would be a significant form of indirect moral enhancement.[86]

The philosophical literature on akrasia has typically focused on two questions: whether akrasia is actually possible, and whether it is necessarily irrational. The

[84] Sometimes, the term 'better judgments' is used in the definition of akrasia. I prefer not to use that phrase, as it deceptively implies that the judgment is by definition superior in some way. I do not want to assume that superiority to avoid begging the question, though I will argue for the general superiority of ATC judgments later in this chapter.

[85] One could even argue that similar means could be used to improve moral ideas. Rationalization, as discussed below, sometimes leads people's inclinations to affect and alter their ATC judgments; to the extent that this is a corruption of reliable judgments, reducing the effect of such inclinations on judgments could make those judgments more reliable.

[86] I will not outline in this chapter the means by which akrasia can be reduced; that is addressed in the next chapter, along with applications of the arguments in other chapters.

impossibility of akrasia was famously defended by Socrates in Plato's *Protagoras* and *Meno*, and later by R.M. Hare (1952) on the grounds that ATC judgments inevitably lead people to be sufficiently motivated to action (displaying a strong internalism). All apparent akratic actions were either unfree, actually in line with an (ignorant) ATC judgment, or merely against what others would judge best, not the agent. Davidson (1969/2001) famously argued in favor of the possibility of a form of akrasia on the basis that an ATC judgment could merely concern what is pro tanto best; this leaves room for other considerations to outweigh the pro tanto judgments. However, this account is still strongly internalist and thus denies the possibility that one could intentionally act against one's ATC judgment concerning what is overall best. More simply and persuasively, others (including Aristotle in the *Nicomachean Ethics*, 1145b25-30) note that cases of akrasia are obvious based on observation of others. If internalist approaches cannot account for the phenomenon, so much the worse for internalist approaches. In any event, I will side-step this debate by conceding that akratic actions may well be unfree and therefore possible on strong internalist frameworks. The question at hand concerns the morality of actions and motives, not people's agency, so the issue of freedom should be irrelevant.

Similarly, I will not claim that akrasia is necessarily irrational, or even that it necessarily involves a failure to be fully rational. There has been a significant amount of debate in this area recently; many (such as Davidson 1969/2001, Bratman 1979, and Hinchman 2013) defend the intuitive notion that acting against one's ATC judgment is necessarily contrary to rationality, as it involves a failure of practical reason, while a number of authors (e.g., Audi 1979 and 1990, Arpaly 2000, Jones 2003 and McIntyre 2006) have put pressure on this view. I put these issues to one side because the present question is not whether akrasia is *always* rational, but whether reducing akrasia is *typically* a moral improvement. Still, much of the following discussion will intersect with the content and

cases deployed in the rationality debate. While the irrationality of akrasia is not strictly relevant, the arguments and cases that are put forward on either side of the debate can also inform the question of whether akrasia reduction will count as an indirect moral enhancement.

Importantly, the conception of akrasia being deployed here is substantively neutral. ATC judgments do not necessarily align with correct judgments, moral or otherwise. Not all share this conception – Aristotle and (more recently) Alison McIntyre (2006) take normative failure (in the form of an immoral decision) to be constitutive of akratic action. Nevertheless, a substantively neutral concept of akrasia is crucial for my approach. I am attempting to show that akrasia reduction counts as an indirect moral enhancement. To do so, I cannot rely on particular substantive claims. If akrasia were normativized, what counts as akrasia (and hence akrasia reduction) would depend on one's substantive commitments. In what follows, I will argue – without relying on particular substantive claims – that akrasia reduction is indeed a form of indirect moral enhancement.

**From thought to action**

Akrasia affects morality in a very straightforward way. Someone recognizes that some course of action is all-things-considered morally ideal or morally required,[87] but nevertheless carries out that action. For instance, someone might recognize the moral imperative to donate significant sums of money to charity because that money could save a number of lives, yet remain selfishly tight-fisted. This is a failure of someone's consciously-held moral judgments to sufficiently motivate them to action. The person's ATC moral judgments are apparently overridden by an inclination that is not itself sensitive to all the

[87] Cases of non-moral ATC judgments will be considered later on.

relevant factors (self-interest may be morally relevant, but the ultimate disinclination to donate fails to take into account the overriding altruistic reasoning that are factored into the moral judgment that one should donate). By reducing akrasia, we could help ensure that those ATC moral judgments more often motivate action.

Again, I am not claiming that avoiding akrasia is necessarily a moral improvement, just that avoiding akrasia can reliably lead to moral improvement. Still, the question remains: why would ensuring people act in accordance with their ATC judgments make them more moral? The simple answer is that those ATC moral judgments are themselves generally reliable, at least compared to contrary inclinations.[88] ATC judgments are more likely to lead to morally correct action than the inclinations that lead people to act against ATC moral judgments. I will offer three reasons to accept this claim: intuitions; the fact that ATC judgments result from the reasoning process outlined in the previous chapter; and the all-encompassing nature of ATC judgments.

*Intuitions*

There is significant intuitive force behind the claim that ATC judgments can generally be relied on.[89] When we attend to the general prospect of acting in accordance with one's ATC moral judgments and going against them, it is clear that the ATC judgments usually win out. And, for the reasons explicated in Chapter 8, we can take these intuitions seriously if they are indeed basic. Showing they are basic is difficult, especially given that I will offer two non-basic independent reasons for this claim below. But having a basic intuition for a

[88] For brevity, I will throughout this chapter refer to ATC judgments as generally reliable full stop; but at every such point, I am actually referring to the moderate claim that they are more reliable than whatever motivates someone to violate an ATC judgment. There will be more discussion on aberrant cases of immoral ATC judgments (as with the fictional character Huckleberry Finn) below.

[89] Jones (2003), who presses the point that one *sometimes* has most reason (including, sometimes, moral reason) to act akratically, concedes that usually, one does not have most reason to go against ATC judgments.

claim is perfectly compatible with having independent, non-basic reasons supporting that claim. More positively, I would simply urge readers to attend carefully to their immediate thoughts and the (apparent) thoughts of others concerning the case of akrasia to appreciate the intuitive force of the claim that ATC moral judgments are generally reliable.[90]

If introspection and reflection is unconvincing, an argument analogous to that provided in Chapter 8 can be deployed instead – though it admittedly only warrants a narrower intuition concerning one's own ATC judgments. While basic intuitions are the building blocks of any moral view, ATC judgments are the inverse – one's total view, considering all the evidence. If one doubts one's ATC judgments, then it is unclear how one could coherently hold that one's view is correct, given all the evidence. Perhaps one could deny that one's moral views are ATC judgments at all, but this severely weakens any view. It means that one holds the view *without regard* to all relevant factors. One has left something relevant out, left it unconsidered when it could undermine one's view. And even if one has good reason to leave something out (perhaps due to time or resource constraints), it would be odd to claim that if everything had been factored in, the view would somehow become unjustified. To be sure, unlike in the case of basic intuitions, one is not committed to the reliability of *other people's* ATC judgments on these grounds – perhaps most people employ faulty reasoning that impairs their ATC judgments. But, in the very least people are committed to personal ATC judgment reliability, to the extent that the ATC judgment itself by definition involves a commitment to a particular judgment.[91] Because of this, akrasia reduction would be a personal moral improvement.

---

[90] The intuitive idea that akrasia is bad has motivated McIntyre (2006) to claim that akrasia is *essentially* bad. But that is too extreme. The intuition merely warrants the weaker claim (which is all I need) that akrasia is generally bad – that is, it generally leads to morally worse action.

[91] It may be that one has higher-order reasons to doubt one's judgment. Perhaps one is aware of biases that corrupt it. In such a case, however, one is really forming a new ATC judgment – one that takes such biases into

One issue with this approach is that ATC moral judgments are, intuitively, not always right. Jonathan Bennett (1974) has astutely observed a case of akrasia leading to right action from the Mark Twain novel *Huckleberry Finn*. Huckleberry Finn is traveling with runaway slave Jim, contemplating whether or not to turn Jim in. On consideration, Huck believes that helping Jim escape is morally wrong (specifically, a wrong to Jim's owner). However, when the moment comes to turn Jim in, Huck has a moment of what he thinks of as weakness and relents. It is arguable that Huck did indeed fail to turn in Jim for the right sort of reasons (his sympathies towards Jim), but it is fairly clear that at least Huck considers this to be a moral failing. This case illustrates a broader objection that reducing akrasia would not reliably make people more moral, as akrasia reduction is content-neutral and so just as likely to lead to the right action as the wrong one.

However, the Huckelberry Finn case is so striking because it is so unusual – that is, it is unusual that akrasia would lead to the morally right action. And for some intervention to count as a reliable moral improvement, it need not always improve someone's morality – just do with relatively high frequency. It is noteworthy that the Huckleberry Finn case is indeed a work of fiction and so somewhat contrived. When contemplating instances where someone doesn't do what they think they should do, the most frequent cases that come to mind are much closer to the donation case than the Huckleberry Finn case. To the extent that this particular case militates against the claim that ATC judgments are reliable, a plethora of more common and plausible cases (failing to give what we think we should to charity, failing to be as honest as we think we should, being less considerate to friends than we think we should, and so on) outweigh them.

---

account and amends the judgment accordingly. One would then be committed to the reliability of the new (but not old) ATC judgment.

Moreover, advancing intuitions about the morality of particular cases (whether about Huck Finn or more mundane examples) violates the neutrality of the indirect moral enhancement framework. It presupposes the morality of an outcome (it is wrong to turn Jim in) in order to judge whether ATC judgments are reliable and, in turn, whether akrasia reduction is a form of indirect moral enhancement. That would mean what counts as moral enhancement is whatever brings about the particular outcomes one thinks best (rather than the reliable processes used here), which is problematic for the reasons laid out in earlier chapters. Intuitions concerning Huck are not, then, admissible in the present argument. But what about the more general intuition that ATC moral judgments are reliable? This is not so clearly violating the neutrality of indirect moral enhancement because it does not concern particular outcomes, but rather more general moral truths. The claim that ATC moral judgments are reliable (supported by intuition as well as process and scope considerations discussed next) does not prejudge the content of those judgments, leaving the morality of various outcomes open. It does not lead us to become locked into one set of judgments, leaves substantial room for people's personal reasoning capacities to determine what is best, and respects individuality by putting faith in people's judgments, rather than imposing the judgments of others on people.

*Reasoning processes*

A second reason to accept the general reliability of people's ATC moral judgments flows from the arguments of the previous chapter. ATC judgments, as a matter of course, involve the sorts of reasoning processes discussed previously that are conducive to coming to correct moral judgments. The 'things' in an ATC judgment are ultimately composed to basic intuitions, or more complex ideas (including other ATC judgments) built up from those

intuitions. Those basic intuitions have been shown to be reliable in Chapter 8. What's more, ATC judgments involve consideration of each of those factors, including how they hang together. That is to say, ATC judgments are the product of a reasoning process, which will display certain features conducive to reliable judgments. These include analogues to all the facets of good reasoning discussed in Chapter 9: logical assessment, empirical assessment, attention to the content and strength of moral ideas in play, critical analysis of various factors, and (perhaps less commonly) identification of potential sources of bias. The presence of these processes alone does not ensure that ATC judgments are usually right. However, it does warrant the weaker claim (which is all I need here) that the ATC judgments are generally more reliable than the contrary inclinations.

Some writers (e.g., Arpaly 2000 and Jones 2003) have pointed out that inclinations conflicting with ATC judgments can have an important role to play in moral reasoning, and sometimes lead to better or more reasonable action than the ATC judgments. That may be true, but the presence of various reasoning processes helps ensure that the ATC judgments will generally be superior to contrary inclinations – cases where inclinations are superior to ATC judgments, like Huckleberry Finn discussed above, will be relatively uncommon. The ATC judgments are simply more sensitive to the relevant factors, and sensitive in the right way. ATC judgment relies, in the first instance, on a set of basic intuitions we can accept as reliable. Then, judgment can involve assessesing the various inputs, checking for applicability and potential flaws. It weighs up different considerations against each other, and can exclude prima facie appealing evidence if it is deemed irrelevant. It is not always the case that this assessment will occur, much less occur properly – but it is a significant advantage that it occurs at all. Inclinations, on the other hand, display much less of this sensitivity. While it may be that, at times, inclinations align better with what one should do,

that alignment is not as reliable because there is not the same level of consideration constitutive of ATC judgments.

One could argue that certain contrary inclinations are more complex and involve some of the same processes as ATC judgments. This may well be the case, and such complex inclinations would be more reliable than basic inclinations. Yet, these inclinations will always be at a disadvantage next to ATC judgments to the extent that the inclinations are not as thoroughly sensitive as ATC judgments. Consideration and evaluation, which provide the reliability, are essential to the ATC judgment, and central in the way it operates. Even in complex inclinations, these are not so central – evaluation may play a role, but it is more ancillary and so less reliable.

The arguments from Chapter 8 that we are generally reliable in deciding which intuitions to hold onto might be thought to pose a challenge here. An inclination could be seen as a sort of basic intuition, insofar as it is a basic, non-inferential input into one's moral reasoning (broadly construed). But if that is the case, then the arguments from that chapter would indicate that we're more likely than not to be right when deciding whether or not to act in accordance with the inclination. Choosing to go along with our intuitions instead of our ATC judgments is just a case of competing intuitions, because we're generally reliable in choosing one intuition over the other, we're reliable in deciding when to act akratically and when to follow our ATC judgment. Akrasia reduction would be pointless, as people are perfectly reliable as is.

This objection, however, is mistaken in its application of the arguments in Chapter 8. ATC judgments, while based on a set of intuitions, are not themselves intuitions, so the claim that we can trust our choices concerning which intuitions to hold does not logically imply we can trust ourselves to adjudicate properly between an ATC judgment and an inclination.

Moreover, the spirit of the arguments in Chapter 8 do not apply here. My argument there was that anyone who holds a moral view is committed to the reliability of the basic intuitions underlying that view. It was, in that way, a claim based on theoretical coherence. Akrasia, however, concerns the practical matter of deciding what to do. Choosing to act against one's ATC judgment in no way commits one to the view that the ATC judgment was mistaken. In fact, the opposite is true – holding the ATC judgment commits one to the idea that acting in accordance with one's inclination is mistaken. If anything, then, the considerations of Chapter 8 provide more reason to support the claim that ATC judgments are reliable.

*Wide scope*

The previous section emphasized the 'things' and the 'considered' in all-things-considered judgments, and now I will turn to the just-as-important first clause, 'all'. 'All' establishes the wide scope of ATC judgments, which lends them particular reliability when compared with mere inclinations. The wide scope is not universal, of course – it generally refers to (what one takes to be) relevant considerations. Determining what is relevant and what is not is itself an important part of the reasoning process. Once that determination has been made, the wide scope of ATC judgments establishes just how thorough they are compared with inclinations. An inclination just takes itself into account; ATC judgments take *everything one thinks is relevant* into account. All those additional factors count as evidence that, on balance, supports the ATC judgment. Just as important, the contrary inclination ignores that very evidence, making its claim to authority over the ATC judgment dubious at best.

We can go one step further and note that ATC judgments have the advantage of being able to take inclinations into account. Inclinations are, after all, a potentially sound input into

one's judgment. Proper ATC judgment would have done do one of two things with a contrary inclination: (1) exclude the inclination from consideration because it is not relevant (e.g., excluding a racist inclination because it is not relevant to a judgment about the best candidate) or (2) weigh the contrary inclination against other considerations, and find that the inclination is outweighed. A 'proper' contrary inclination that serves a good purpose (say, the inclination to avoid causing suffering), by contrast, will not by take the ATC judgment into account. Choosing the ATC judgment in such a case would be the dominant option, since it includes all the 'evidence' contained in the inclination and a wide range of factors besides, whereas the inclination does not by nature take anything other than itself into account.[92]

ATC judgments will not, of course, always be 'proper'. They could improperly exclude some inclinations and improperly weigh others. Yet they still hold an advantage over inclinations because they at least attempt, in their structure, to take inclinations and all other relevant factors into appropriate account. Indeed, by forming an ATC judgment, one is more or less committing oneself to the claim that potential contrary inclinations are either irrelevant or outweighed. That claim is not groundless – it comes from the fact that one has taken into account a variety of relevant factors and assessed them to determine the appropriate judgment. The contrary inclination involves no such essential claim to taking other things into account, and will be less reliable because it lacks the totality and reasoning processes of the ATC judgment.

One might still worry that the inclusion of a greater number of considerations increases the risk of error. Each new factor put under consideration is subject to biased distortion that might introduce greater error into one's judgment. However, it is important to

[92] Here, my purpose is to defend the notion that ATC judgments are generally reliable. However, the greater reliability of taking more relevant factors into account suggests a further possible means of indirect moral enhancement: encourage people to take more relevant factors into account when deliberating. This suggests, for example, an education program that presented various moral considerations can count as a moral enhancement.

recognize that errors of omission are also likely: by failing to take some (relevant) factor into account, one risks distorting the factors that remain. In fact, there is greater reason to worry about errors due to omission of relevant factors than errors due to inclusion of relevant factors. The very judgment that a factor is relevant implies it actually has bearing on one's judgment, and its inclusion would make that judgment more accurate. Distortion is certainly a risk, but there is generally more reason to worry about distortion due to failure to include some piece of information. However, if one has access to some relevant piece of information but knows one would grossly distort it due to some bias, the proper reaction would be to downgrade how much one weighs that piece of information. In such a case, one has taken an *additional* factor, one's bias, into account to ground the downgrading. Indeed, the fact that wider-scope judgments are more able to take such biases into account and amend other considerations accordingly is a further advantage in producing reliable judgments.

## Alternative frameworks

The preceding has presupposed a certain framework for reducing akrasia by bringing one's actions in line with one's ATC judgments. However, my claims can be expanded to include both (a) non-moral ATC judgments and (b) certain classes of non-ATC judgments. I can also accommodate the position that ATC judgments could be revised based on akrasia.

### *Non-moral ATC judgments*

Up until now, I have for the most part been addressing ATC judgments concerning what is morally best and evaluating them in terms of whether they can be expected to lead to more moral outcomes. However, entertaining an ATC judgment does not imply that one thinks the option is the most moral. While some theorists (such as Kant (2002) and Hare

(1981)) think that morality is indeed generally overriding, this is a substantive claim that others (such as Williams (1985)) deny, making it an unsuitable assumption in a framework for indirect moral enhancement. Moreover, some people might not themselves believe that morality is overriding – perhaps, in their internal calculations, moral considerations are outweighed by non-moral, prudential ones. Even if morality is in fact overriding, dissenting opinions must be accounted for. My present claim is that reducing akrasia in general will lead to moral improvement, not just akrasia in the case of an ATC moral judgment.

Expanding the scope of the present argument is warranted because the content of even non-moral ATC judgments comes close enough to the goal of moral enhancement. Suppose Jane weighs the moral reasons to donate a sum of money to charity against the prudential (stipulated here, for the sake of argument, as non-moral) reasons to buy herself a nice painting, and finds that all things considered it would be best to buy the painting. Suppose further that this is correct[93] – it is actually best, all things considered, for her to buy the painting. Would it be a moral improvement to reduce Jane's likelihood of her acting akratically and, despite her ATC judgment, donating the money to charity? Yes, so long as we allow a somewhat ecumenical view of what can count as 'moral improvement'. In the present case, the objective truth of the matter (we are assuming) is that it is best for Jane *not* to donate her money to charity. Morality, on this picture, falls along a scale – one can be too immoral, but also too moral (when other considerations outweigh moral ones), and there is some appropriate mean of morality we should strive for.[94] Enhancement of morality, in this context, does not refer to maximizing how moral someone is, but rather getting them

[93] Thus, for the sake of argument, assume morality is not overriding. If morality is in fact overriding, then Jane is mistaken and akrasia reduction no longer leads to a moral improvement – it leads her to do the wrong thing. This would then become a case of a fallacious ATC judgment, discussed above.

[94] Something similar is, I take it, the force behind Susan Wolf's arguments against moral perfectionism in *Moral Saints* (1982) – it is better that people be less than saintly. Alternatively, one could (as suggested by Parfit 2011) think that we can potentially have sufficient reason to act against moral requirements. In this case, moral enhancement would consist (in part) in ensuring one only acted in accordance with moral requirements when one has sufficient reason to do so (which will naturally involve improving one's reasoning processes and acting in accordance with one's reason-sensitive judgment, perfectly in line with the present arguments).

to the appropriate level of morality. So, reducing akrasia and helping ensure that Jane acts on her ATC judgment is indeed a moral improvement.

If that is correct, the next step is to show that ATC judgments in general – like ATC moral judgments – are reliable. This task is relatively straightforward, because all three arguments above are applicable to the non-moral ATC judgments as well. Our intuitions will indeed match up that (say) ATC judgments concerning what is in one's personal interest can generally be trusted, and our commitments to a wide variety of non-moral projects will ensure this. Non-moral ATC judgments involve the same sort of reasoning processes as moral ATC judgments. And, non-moral ATC judgments are total in the way that inclinations are not – they take into account a variety of evidence, the totality of what the agent considers relevant.

It is important to clarify that 'non-moral ATC judgments' does not refer to ATC judgments that ignore moral considerations; rather, it refers to judgments whose final claim is not necessarily moral in content. Non-moral ATC judgments will typically admit of moral considerations, as they will almost always be relevant. Yet, one might worry that this could make non-moral ATC judgments unreliable, on the grounds that many will (improperly) ignore moral considerations entirely. It may be that the contrary inclination is in fact a moral one, and people's (perhaps selfish) blindness to that consideration undermine the reliability of the ATC judgment. This is in essence a stronger version of the worry discussed earlier that ATC judgments might leave out important considerations captured by contrary inclinations – perhaps people systematically leave out a whole class of crucial contrary inclinations. Two things can be said in response. One, this is party an empirical claim that is – on its face – quite dubious. It would imply that a wide segment of the population is not committed to

morality at all, that they consciously think it *completely* irrelevant to what they should do. This would be to impart a sort of psychopathy on a huge number of people, and without convincing evidence that such pathology is widespread there is not much reason to think there are so many who think morality completely irrelevant. Two, this problem looms much larger for contrary inclinations, again due to their lack of totality. ATC judgments may have blind spots, ignoring relevant considerations, but contrary inclinations will always have more blind spots – due to their structure, they fail to account for a much wider swath of relevant considerations, both moral and non-moral, beyond the scope of the narrow content of the inclination itself. So, even if ATC judgments did have such systematic blind spots, contrary inclinations have even more and the ATC judgments are to be preferred.

This expansion to non-moral ATC judgments suggests a slightly altered mechanism by which akrasia reduction operates as a moral improvement: people form ATC judgments concerning what is best, forming a judgment that at least partly concerns what would lead to the *best* amount of morality. The ATC judgments are generally reliable, so ensuring people act in accordance with them helps make them as moral as they should be. For the remainder of the thesis, I will return to the simpler language of simply making people more moral; however, readers who seriously doubt that morality is indeed overriding can substitute 'making people as moral as they should be'. All my arguments before and after this point apply equally well in either case.

*Pseudo-akrasia*

Akrasia is typically understood in terms of intentionally acting against one's ATC judgments, and I have relied on that conception. The completeness of ATC judgments allows people to make strong claims about its rationality, and allows me to claim they are reliable.

However, the actual phenomena associated with akrasia (for instance, eating a piece of cake when one knows one shouldn't), or the related notions of incontinence and weakness of will, need not be so complete. Oftentimes, agents will form a judgment but make no claim that it has taken everything they consider relevant into account (perhaps because time or resource constraints prevent that). Here, I will argue that the preceding account can be expanded to what I will call pseudo-akrasia: acting against one's considered and decisive (CD) judgment. I will argue that CD judgments are more reliable than contrary inclinations, and so reducing pseudo-akrasia is a (perhaps weaker) form of indirect moral enhancement. This expansion not only increases the means by which moral enhancement may be brought about, but helps ensure the means employed to reduce akrasia – which may not be sensitive to the distinction between akrasia and pseudo-akrasia – are indeed reliable moral enhancers.

CD judgments share some features of ATC judgments, but crucially lack others. Like ATC judgments, CD are formed after reflection and consideration of various relevant factors. Unlike ATC judgments, CD judgments do not take into account – by the agent's own lights – *all* potentially relevant factors. Consequently, the third argument for the reliability of ATC judgments – their totality – is inapplicable to CD judgments. Still, the two other arguments (from intuition and reasoning) remain and, I submit, are sufficient to ground the reliability of CD judgments. I have argued that we can trust our intuitions in these sorts of cases, and the actual intuitions concerning ATC judgments are not actually reliant on the totality condition (or 'A') of ATC judgments. We attend to everyday cases of thinking that we should exercise more, but fail to do so. While sometimes the judgment that we should exercise more involves a conscious assessment of all the evidence, more often it is not so ambitions – a weaker judgment based on a few relevant considerations of time, effort and health. And we find it quite appealing that these sorts of judgments are to be trusted over the contrary inclination, say, to stay in bed. As important, CD judgments are reliable because they are considered –

they are the product of a process of reasoning and reflection that can weigh up different considerations in a sensible manner, attend to relevant details, discount irrelevant factors, and so on. Contrary inclinations are not so reflective and therefore reliable. They will be more susceptible to bias, disproportionate emphasis, inattention to potentially crucial factors, and so on. As before, there is no compensating advantage possessed by contrary inclinations that could outweigh those flaws, and so CD judgments can be accepted over contrary inclinations.

The lack of totality might be thought to raise a problem for CD judgments. If not all factors are taken into account, one could have two opposing CD judgments – one based on consideration of one set of factors, another based on consideration of a different set of factors. There would seem to be no way to distinguish between the two, and so no way to pick which judgment should determine action while remaining neutral about substantive content. This would be an issue for mere considered judgments, but the 'D' in CD judgments is crucial here: it would not be coherent to entertain two contrary decisive judgments, as 'decisive' is an exclusive notion. It may be psychologically possible to think two incompatible judgments are both decisive, but any self-aware agent will be aware of the apparent conflict. So, the agent will have to pick one as decisive over the other. What's more, because the agent must actively choose in this manner, she will as a matter of course go through the process of taking both considered judgments into account and, in essence, weighing them up. The decision over which judgment is decisive then, in effect, merges the two sets of considerations into one new, unified CD judgment. It may still lack the totality of ATC judgments (some third set of considerations ignored), but it will nevertheless be a good guide to action.

Certainly, CD judgments will be less reliable than ATC judgments and the latter should take precedence if there is ever a conflict. But absent ATC judgments (as will often

be the case), we can reliably bring about moral improvement by ensuring people act in accordance with their CD judgments.

*Upstream reasoning*

So far, I have suggested that akrasia should be reduced by submitting to one's ATC judgments. There is an alternative response to contrary inclinations, however: revise one's ATC judgment. Hinchman (2013) refers to this as 'upstream reasoning', contrasted with 'downstream reasoning' when one brings one's inclinations in line with one's ATC judgments. This is, in a way, a form of rationalization, but Hinchmann suggests it may not always be such a bad thing. The claim is not that it is always best to revise ATC judgments, but merely sometimes – when that revision is the result of reasonable self-mistrust.

Hinchman characterizes upstream reasoning as follows: "I judge that the facts of my situation give me conclusive reason to φ. But I can't bring myself to trust that judgment by forming a practical commitment to φ. Thus, [here you abandon the judgment] it is not the case that the facts of my situation give me conclusive reason to φ." (Hinchman 2013, pp. 539-40) At that point, the agent revisits and revises the ATC judgment based on the mistrust. This is too strong an implication – while it is true the agent mistrusts her judgment, the fact that she holds that judgment evidently means she also mistrusts whatever inclination is leading her to doubt her judgment. At most, then, we should be neutral between revising our ATC judgment and overcoming the contrary inclination in order to act in accordance with it. But even that more moderate reading of Hinchman's view undermines the ATC judgment-conformity I have been advocating.

Still, I would argue that what is going on in self-mistrust cases is that one's confidence in the ATC judgment has shifted. First, one forms an ATC judgment, thinking

one has taken everything relevant into account. The ATC judgment recommends action, but a contrary inclination (in this case, mistrust) militates against that. The mistrust is not an ordinary inclination, however; it manifests itself specifically by reducing one's confidence in the ATC judgment. One no longer believes the ATC judgment has taken everything into account, or weighed up everything properly. In this way, the mistrust points out a new piece of evidence that informs the ATC judgment. It is analogous to a person who decides she should buy a new sweater, then gets a suspicious feeling immediately after making that decision, looks closer, notices the sweater has holes in it and so shouldn't be bought. Hinchman denies the mistrust is itself evidence, but it is nevertheless something that points out or flags evidence that was previously not taken into account properly. Notice that at no point while the ATC judgment was confidently maintained was revision (and acting against the judgment) warranted. Once the doubt manifested itself, it is fair to say that the judgment was, at that moment, no longer all-things-considered at all – there was a piece of evidence that, by the agent's own lights, should have be accounted for and wasn't. The principle that it is generally best to follow one's ATC judgments (so long as they remain all-things-considered), then, remains intact.

This line of thinking does lead to some practical revision, however, because it emphasizes the importance of having good ATC judgments in the first place. Part of akrasia reduction, then, is being attentive to all relevant factors (including those which motivate mistrust) when deciding what to do. Thus, the reasoning process improvements outlined in the last chapter can actually operate as means to reduce akrasia down the line – by ensuring potential contrary inclinations are taken into account.

**Ancillary benefits**

If the preceding is correct, then we have good reason to help people to reduce akrasia on the grounds that it will generally lead to moral improvement. But it should be briefly noted that akrasia reduction, like reasoning improvement, has ancillary benefits that would make it especially acceptable and worth promoting.

Just as often as akrasia interferes with morality, it interferes with prudence. The intuitively compelling cases are generally of this sort – failing to do something (eat well, study more, get enough sleep) that one knows would be in one's best interest because of some contrary inclination like laziness. Reducing akrasia helps promote people's interests in a wide range of such cases. There may be cases where prudence and morality conflict and ATC judgments must pick between them. Akrasia reduction may not maximize prudential benefit, in this way. Still, it improves people's interest in the most acceptable way – helping them become better off when doing so is not outweighed by relevant moral considerations. This compromise will naturally be acceptable, since it is people's own internal judgments that will determine when personal interests are so outweighed.

Akrasia reduction also has the benefit of potentially improving someone's autonomy, assisting them in controlling the course of their own lives.[95] Insofar as autonomy consists in self-governance, akrasia is a paradigmatic case when one fails to govern oneself. ATC judgments are regulative in a way that allow for proper adjudication of all of one's values, while contrary inclinations do not allow anything comparable. Even someone thought, on the contrary, that acting in accordance with inclination was the best way to govern oneself, that principle itself would count as an ATC judgment serving to regulate and govern future behavior. Akrasia reduction even more clearly promotes autonomy on the picture advanced by Frankfurt (1971) and others. On this view, autonomy involves a certain coherence between one's higher-order desires or judgments (what one thinks one should desire/wants to

[95] This is contrasted with the more heavy-handed, individuality-suppressing direct moral enhancements.

desire) and lower-order desires or judgments. ATC judgments are typically higher-order in this way, to the extent that they concern the intentions one thinks one should form, while contrary inclinations are lower-order. Bringing them into coherence (which is accomplished by akrasia reduction) would be a straightforward way to improve autonomy, on this popular view.

Finally, and most provisionally, akrasia reduction has the *potential* to make people comport more with the norms of rationality. I say potential because, as noted above, there is a live debate over whether akrasia is necessarily irrational and I have not offered any particular arguments for or against that claim. Still, the openness of the question means that it is quite possible that akrasia is indeed necessarily irrational. To the extent that people are committed to rationality, they would then have good reason to support significant efforts to reduce the occurrence of akrasia in the population. And even if the critics are wrong and akrasia is not necessarily irrational, that still leaves open the possibility that it is typically rational, or its avoidance usually brings one closer to the norms of rationality – similar to my present claim that akrasia reduction will generally lead to more moral outcomes. While not every case of akrasia reduction would be an improvement in rationality, the general effect on a population subject to akrasia reduction would be that they become more rational than they were before. That is an outcome most, I expect, would welcome.

# Part IV: Upshots

We have now seen how focusing on an indirect approach to moral enhancement can avoid the problems of disagreement while reliably bringing about moral improvement. This final section, substantially briefer than the preceding, will explore the upshots of this work. Chapter 11 addresses in some detail the various practical implications of all that has been said. Among other things, these arguments imply that (in agreement with liberal neutralists) public education should be morally neutral and manipulation of morality through propaganda and psychological tricks be kept to a minimum. It also suggests that state subsidy or promotion of particular novel biological interventions aimed to make people conform to

particular moral standards would be inappropriate (to say nothing of coercive imposition). However, this does not mean that direct moral enhancement should be banned; on the contrary, allowing *self*-enhancement would be a good way to reduce akrasia in the general population. My final thoughts are given in Chapter 12.

# Chapter 11: Implications

My analysis so far has remained relatively abstract, addressing relatively theoretical problems with direct moral enhancement and again relatively theoretical benefits of direct moral enhancement. But moral enhancement is, at its core, a practical project. The aim is to try to determine the best, most acceptable means by which to foster moral improvement. With this in mind, I will now turn to the more practical implications of what I have been discussing.[96] I will discuss what this means for both traditional and more novel forms of moral enhancement, both negatively (what we have reason to avoid) and positively (what we have reason to pursue). The result will be a series of policy recommendations heavily informed by the theoretical discussion that has dominated the preceding chapters.

## Traditional, Non-Biological Moral Enhancement

Traditional approaches are those that have been in use for millennia by various governments and individuals to bring about moral improvement. Mirroring the discussion in Chapter 2, I will discuss implications in several domains: moral education, argumentation and punishment, propaganda, social influence, and psychological manipulations.

### *Moral education*

[96] That having been said, I will not go into great detail concerning particular interventions or the way institutions should be designed. Such details are outside my area of competence. Instead, my aim is to specify the contours and direction that policy should take, with details to be filled in depending on what are the most efficient means to bring about those policies.

In a system of traditional education, there is a teacher and a pupil – the former attempts to improve the latter's knowledge, skills and other attributes through a variety of pedagogical techniques. In the case of moral education, that goal is more or less to improve moral ideas, motives and actions. But the arguments in Part III were intended to put serious pressure on a direct approach to this moral education – an approach that presupposes the teacher has the right moral ideas, motives or behaviors and attempts to bring the same about in the student. These problems involve both the feasibility of determining who the morally astute teachers are as well as the desirability of imposing one teacher's moral views on her students (which problematically suppresses dissent).

If those arguments are successful, then a certain Aristotelian model of moral education is ruled out: identify the virtuous individuals, and use them as a model for students to imitate (over time, habituation leads them to act in accordance with the right reasons). The problem is not with imitation and habituation, but with the prior identification of the virtuous individuals. What exactly allows us to reliably identify such individuals? Aristotle's answer would presumably be the same sort of natural perception of virtue that allows individuals to discern the appropriate levels of various traits (see, e.g., *Nicomachean Ethics* 1109b21-27). However, this obscures what to do in the face of disagreement – when two different agents have differing perceptions concerning who are the virtuous individuals and what standards should be employed in moral education. There may well be a fact of the matter that one is more virtuous than the other, but there is no clear, reliable way to distinguish them. What's more (related to the problem of peer disagreement previously discussed), it is not clear how one can privilege one's own internal moral perception of virtue over that of others.[97]

[97] One could try to pick out minimally virtuous people by overlapping consensus and use them as paragons. However, the arguments at the end of Chapter 5 undermine this prospect; for example, while we might agree on the general contours of justice, agreeing that one person's particular conception is correct will be disputed. Indeed, I would challenge anyone to find a generally-agreed-upon moral paragon. One could instead advert to very specific features of paragons most everyone would agree are virtuous – for instance, someone who thinks

More straightforward approaches to moral education are also ruled out. One might assume a particular moral view or theory, and deploy various pedagogical techniques in order to bring students in line with that view or theory. This might involve simply informing them of the correct moral view (appealing to the authority of teachers), having them read (only) proponents of that view, or utilizing more nuanced techniques like interactive exercises intended to elicit intuitions in favor of the preferred view or theory. These means are certainly questionable (I offer some discussion of them below), but the approach goes wrong from the outset by presuming that the teacher's position is the correct one. All the problems of direct moral enhancement previously discussed make it untenable.[98]

Moral education is not a lost cause, however. Building on Chapters 9 and 10, we have good reason to think that fostering better reasoning is a powerful means of indirect moral enhancement and fit naturally within an educational context. Indeed, traditional non-moral education is already committed to improving reasoning. Problem-solving, creative thinking, careful reflection and other aspects of reasoning are well within the remit of current practices. Empirical competence, especially, is central to science and historical education. But this analysis implies some revision to current practice: logic (which helps weed out moral inconsistencies, both in formal and informal iterations) should play a more central role in education. While currently it is mostly only introduced at the undergraduate level, the central concepts of what constitutes valid arguments and helping explicate fallacies should be

---

causing immense, wanton suffering is inherently wrong. But beyond the problems of spelling those features' implications out (does that mean causing immense suffering can never be justified for the sake of the greater good?), such extreme cases will actually be grounded in near-universal intuitions. One does not need virtuous paragons; one can instead leverage people's common intuitions through argumentation to get them on board with, say, opposing wanton torture.

[98] Interestingly, despite great theoretical interest in direct moral education, it is presently not a very strong part of public education programs in liberal democracies. Core curriculums are defined by subject matter (English, science, history, math, etc.) and typically do not have strong explicit moral goal. The negative portion of this approach, then, might not be substantially revisionary. The one notable exception is (in my experience) civics education; students are often taught that their own government's system is just and fair. On my view, this sort of approach to civics should be excluded (it is really a form of propaganda). It would be replaced with a more open discussion of why proponents of the government's system think it is justified, as well as engagement with the system's critics and a fair hearing concerning alternative systems – essentially a form of structured argumentation.

introduced at a younger age, when moral ideas are still being formed. And imparting a greater understanding of the role of bias in people's thinking – as well as helping students mitigate those biases – could go a long way towards reducing bias in moral thinking.[99]

This approach, while not wholly Aristotelian, actually comports well with Aristotle's influential account. A key part of Aristotle's account are the virtues of thought, to which he devotes the entirety of Book VI of the *Nicomachean Ethics*. These virtues of thought work together to (partly) constitute correct reasoning in various contexts; they include wisdom (concerning universals), prudence (concerning particular actions), deliberation (concerning careful inquiry towards some end) and comprehension (concerning particular judgments). Each of these areas is plausibly the target of moral education *qua* reasoning improvement: helping people understand universal laws; making them aware of the consequences of their actions; revealing unbiased means of deliberation; and fostering more acute comprehension of various issues, to name a few examples. And one need not pledge oneself to any particular substantive moral view in order to promote virtue in this way, avoiding the problems of substantive disagreement.[100]

It should be clear at this point that such indirect approaches to moral enhancement align well with various neutral approaches to moral education outlined in Chapter 2. However, it does so for substantially different reasons. Importantly, it does not presuppose moral relativism, non-cognitivism or Rawlsian liberalism (which have in the past motivated a neutral approach to moral education). In this way, it is substantially *more* neutral than any of those three approaches. This has some practical import; those three views justify neutrality in educational curriculums *except with regard to the view itself*. A Rawlsian liberal might argue that the state should avoid endorsing a particular comprehensive view, but it should (if

[99] Reducing akrasia, on the other hand, falls rather far outside the scope of traditional education. Any traditional attempts to reduce akrasia will likely have to fall within the more psychological approaches discussed below.
[100] See Fröding (2010) for a more developed account of this Aristotelian approach to enhancement.

consistent) nevertheless endorse imparting Rawlsian liberal values on its students. A relativist approach may appear neutral by not endorsing specific moral norms, but it takes a strong substantive stance that moral truth varies from individual to individual; a part of the curriculum would involve imparting this controversial view. And a non-cognitivist approach similarly would have to advocate for its own controversial approach.[101] Indirect moral enhancement, by contrast, does not require commitment to any such controversial views; it is committed to much more minimal and uncontroversial claims about the connection between reasoning/akrasia and morality, as outlined in the previous chapters.[102]

*Argumentation and Punishment*

Towards the end of Chapter 6, I outlined how argumentation and punishment need not run afoul of the objections to direct moral enhancement outlined therein. The idea was roughly this: proper argumentation is unlikely to stifle dissent and can be quite respectful of people's reasoning processes and individuality, insofar as it is left entirely up to them whether to accept a view. Punishment's purely external character helps avoid the worry that it would stifle dissent, and its coercive assault on the individual's freedom is evidently already justified (if the punishment is thought permissible independent of the present arguments) by either retributive or consequentialist considerations. Yet certain approaches to argument or punishment may still be problematic. Here, I will sketch in a little more detail how one can ensure that argumentation and punishment conforms to the ideal of indirect moral enhancement.

---

[101] Unless, of course, the non-cognitivist is a quasi-realist like Blackburn who denies the view has implications for first-order views – in which case, it won't recommend any particular approach to moral education at all.

[102] As has been noted, indirect moral enhancement is not *completely* neutral; however, the minimality of its core commitments make it sufficiently neutral to avoid the problems discussed in earlier chapters. At most, the view implies it is acceptable to teach students that better reasoning and less akrasia can make you more moral – but this seems generally acceptable and would do little to lock people into one way of thinking, subvert reasoning or impair individuality.

Ideal arguments work to improve reasoning; they bring about moral improvement by exposing people to (among other things) factors they had not previously taken into account, overlooked flaws in one's own thoughts or ill-understood alternate ways of thinking. Dialectics can take a variety of forms, but in order to count as enhancement they should remain consistent with the notion that either side is open to improvement based primarily on the content of the dialectic itself. Not all arguments are productive, however, and there are a few systematic ways it can actually subvert improvement.

A discussion that omits crucially relevant facts or positions may lead to a distorted view of the field. This omission can be brought about most perniciously as censorship: certain ideas are not permitted for discussion. This makes censorship anathema to a project of indirect moral enhancement. But omission need not be the result of a conscious decision to exclude material; it may instead arise from unconscious bias (discounting disconfirming evidence) or mere ignorance. Positive efforts on the part of interlocutors may then be required, including educating themselves and attempting to counteract potential biases. At a certain point, though, time and other resource constraints will prevent inclusion of all the relevant details in a dialectic. Good faith efforts may be needed at that point to determine the most relevant factors to consider and bring them to the table.

Interlocutors may also end up distorting the relevant facts or positions. At its worst this would involve outright deception, which naturally would subvert moral improvement and cannot be tolerated. But again, it is quite possible for people to innocently distort the relevant issues. They may overstate the strength of their own claims or misrepresent the details because of poor understanding. And again, self-education is a plausible way to avoid this. In addition, a better sense of self-awareness and intellectual honesty may be needed. Careful introspection can help agents discern the strength of their own ideas more fairly and communicate them more accurately. Intellectual honesty, in turn, involves presenting one's

credence fairly and accurately; if one has a dim recollection of the facts or is unsure, one should say so and one's interlocutor judge accordingly.

Finally, a mutual sense of openness is needed for there to be any possibility of improvement during a debate. Each side must be willing to revise their own views in light of critiques. This was identified in Chapter 9 as a component of critical analysis, and it plays an especially important role in argumentation. Too often in such contexts, people remain trenchant in their views and only wish to convince others to change. But that attitude, universalized, would make any improvement through argument impossible – shutting off what may be the most reasonable and reliable way to shape moral ideas and behavior. One might think that only those who are *wrong* need to be open, and those in the right need not be subject to revision. But that presupposes one could properly identify who is right and wrong, in violation of the neutrality of the indirect moral enhancement approach. In addition, since most people think they are in the right in an argument, the rule 'be open to revision if and only if you are wrong' would in practice make revision impossible.

As for punishment, care must be given to ensure that it does not fall into the trap of direct moral enhancement. Punishment, as previously stated, is external and so could avoid the problems of suppressing dissent by simply not aiming to affect people's reasoning and individuality in determining right and wrong. Clearly, this means that punishment for possessing particular moral ideas is prohibited. It also suggests, in concert with the concerns surrounding argumentation, that one should not be punished for advocating particular ideas. This justifies a strong freedom of speech, insofar as free speech is crucial for open debate and moral progress.

State punishment also needs to be justified very carefully. It is problematic to justify any particular punishment based on the idea that it will lead to a moral improvement in the

subject. That is to presuppose a particular substantive view one wishes to inculcate in the populace through coercion. Unlike reasoning and akrasia reduction, there is little reason to think that coercive practices will generally lead to moral improvement. Instead, punishments need to be justified based on alternate goods. This may include giving people what they deserve in the case of a retributivists, or preventing harms to third parties in the case of consequentialists. Other goals – such as the rehabilitation of criminals – must similarly be designed with an eye towards goods other than moral improvement based on a presupposed standard of behavior.

It might be wondered whether punishment can actually act as an indirect moral enhancement by improving reasoning or reducing akrasia. Aristotle, for his part, thought legislation could lead to improved virtue by habituating people to doing the right thing, and over time they come to appreciate the right reasons for doing the right thing. (*Nicomachean Ethics* 1180a-b) However, this approach still problematically presupposes what the right ideas are. People's reasoning counts as improved only insofar as they reason in accordance with how the punisher thinks they should. One could instead try a form of punishment that is substantively neutral – instead, punishing people when they reason improperly or act akratically. With regard to reasoning, this is how grading in the classroom works. This may well be an acceptable form of punishment, but it is important to note that the context is substantially less coercive than standard criminal circumstances. Grading can comport with individuality and allow dissent insofar as it leaves reasonable room for people to value the grade (and its long-term benefits) as they see fit.[103]

*Propaganda*

[103] This also suggests that parents should have significant latitude to opt their children out of standard educational curriculums, in line with the proposal of liberal neutralist Kyla Ebels-Duggan (2013).

Propaganda typically involves at least two salient features: it operates via non-rational or sub-rational processes to influence public attitudes, and it is designed to bring about particular attitudes in the public. These features make it a problematic form of direct moral enhancement, and it is unlikely that any forms of propaganda could be classified as acceptable, indirect moral enhancement.

The appeal to emotion typical of propagandist efforts might not be in itself objectionable. While I have identified somewhat rationalist routes of indirect moral enhancement (improving reasoning and avoiding akrasia), it is in principle possible to avoid the perils of direct moral enhancement through other emotionally-laden means. This sort of approach would have to be carefully constructed, however, so as to avoid limiting the possibility of moral progress, infringing on people's reasoning processes, and disrespecting their individuality (key flaws of direct moral enhancement identified in Chapter 6).

Political propaganda fails on all these counts. By design, political propaganda efforts presuppose some moral view rather than being earnestly open to dissent and the possibility of revision (as in ideal argumentation). This by itself is limiting, insofar as one particular view is pushed on the public; if successful, it narrows their moral outlook and reduces the likelihood of critical engagement with the propagandist's positions. The appeals to emotion are not designed in a neutral way that might be used to augment or complement people's reasoning; rather, they are meant to supplant it. And it is hard to respect individuality when there is a unidirectional attempt from the propagandist to bring wide swathes of the population into line with their view.

Not all forms of propaganda are political, of course. One might think that certain educational campaigns could count as both propaganda and acceptable indirect moral enhancement. A commercial, for instance, might stigmatize dropping out of school by

portraying the dropouts as losers. The ultimate aim would be to reduce dropout rates, making the population better educated. This could be expected (if the arguments of Chapter 9 are compelling) to generally lead to moral improvement in society. The problem, however, is that the means used to bring about that indirect moral enhancement is to inculcate a particular morally-laden view – in this case, dropouts are losers. The ultimate purpose of the propaganda might not be to make that belief widespread, but it is a constitutive part of the effort. That constitutive part, in turn, problematically presupposes a controversial moral view and attempts to impose that view on society without respect for their reasoning and individuality.[104]

This isn't to say that public service announcements are generally unacceptable. Their design, however, must avoid propagandist tendencies. One way to do so is to have a purely informative campaign – informing individuals of the relevant facts. This may run into the problem of incomplete information noted in the subsection on argumentation above; the campaign might include some pertinent information but omit others. One solution would be to encourage multiple parties to put forward their own (accurate) informational campaigns, and gaps in one group's presentation would be offset by its inclusion in another's. More thoroughly, the campaigns could strive to include all the relevant information themselves and leave it up to the audience to make an informed decision. This is more or less the (ideal) function of news organizations, and such organizations can certainly play an important role in bringing about moral improvement in society by making them better able to engage with a wide variety of issues in society. Of course, given how susceptible people are to various biases, this picture of information transmission may be overly optimistic – any campaign

[104] One could respond that the stigmatization is in fact aimed at improving reasoning and individuality (the output of a better education). This is then an instance of the paradox of autonomy – inhibiting autonomy in order to promote it. I would concede that, in theory, there may be times that such inhibition is justified. However, there would need to be clear and convincing evidence of those gains – that they are large enough to outweigh the harms of the ad itself. In absence of that evidence – but good reason to think the ads in themselves inhibit individuality and proper reasoning – we should reject this approach.

might inadvertently play on people's prejudices and fail to be fully balanced. Nevertheless, purveyors of that information have a duty to at least attempt to make the campaign as fair and forthright as possible.

*Social Influence*

While propaganda is a fairly blunt and direct tool for would-be moral enhancers, social influence is much more subtle. It is not necessarily the product of a particular group or concerted effort; it can emerge organically without any direction. Such entirely emergent phenomena might result in moral improvement, but their lack of direction means they could not count as moral enhancements. Still, other forms of social influence might me more intentionally designed. Individuals within a society, for instance, can exert peer pressure in order to get others to conform. In addition, organizational campaigns might try to reinforce a group's already-existent commitments and encourage them in turn to influence others.

These influences and campaigns could be pernicious if designed with a particular moral view in mind. Like with propaganda, an individual or group would be attempting to impose its substantive moral views on others by subverting rational processes. The processes it does use (such as bestowing social benefits on those who conform and alienating those who dissent) do not seem particularly reliable, compared to the reasoning processes it is supplanting. Excessive conformity is indeed one of the central problems of direct moral enhancement, insofar as the group develops an overly-unified mindset and cannot develop or meaningfully critique its prior views. It is also inimical to people's individuality (as suggested in Chapter 2) because the will of the individual is, to a certain extent, substituted by the will of the group.

In light of my arguments in Chapter 8 for moral optimism, it might be thought that there actually is some reason to think that social influence is a morally reliable process. If it moral optimism is correct, we can expect people's basic moral intuitions to be reliable. Perhaps one should then trust a large group's professed intuitions in various matters and allow oneself to be swayed by them. However, as previously noted, the moral optimism view is primarily internal and basic; it does not imply that people's expressed moral ideas are reliable – the intuitions might be misrepresented or poorly communicated, and moreover might be corrupted by unreliable reasoning processes. Moreover, even if one thinks that the group's opinion is relevant to one's moral views, it is important that these be taken into account in the right way – via a process that can reliably evaluate that evidence and weigh it up against other relevant considerations. That process is, of course, reasoning. To the extent that one is influenced by society, then, it should not be in the way that 'social influence' is typically understood (subconsciously or through pressure), but instead via a reliable reasoning process.

Social influence could also be an acceptable form of indirect moral enhancement if it aims not at moral outcomes but reliable processes. Cultivating an environment where critical reflection, openness, clear-headed evaluation of the evidence, and so on are actively encouraged could work to improve reasoning processes (and thereby morality) without commitment to any particular worldview. This is still somewhat problematic, to the extent that encouraging conformity with the group is the proximate mechanism by which reasoning is improved. This conformity is still in tension with people's individuality. This tension can be mitigated, however, by ensuring that the attempts at conformity are, to a certain extent, self-undermining. If part of proper reasoning is learning to think for oneself, then peer pressure might be used to encourage the idea that one shouldn't simply conform to the group's opinion. Once the process is over, the individual would no longer be as susceptible

to social influence. This may be for the best, given that social influence could impede one's moral reasoning.

*Psychological Manipulations*

In Chapter 2, I foreshadowed how certain psychological manipulations are problematic examples of direct moral enhancement. Allegedly indirect psychological manipulations typically presuppose certain substantive views, for example, concerning moral development and what constitutes improvement. That presupposition runs the risk that any intervention will be just another (more complex) case of the enhancer imposing his or her values on the enhancee. And manipulation of motives themselves might be able to avoid specifically interfering with the reasoning process per se, but it strongly runs afoul of individuality. One's personal motives are close to one's sense of self, and altering those motives to better conform with the enhancer is another problematic form of will-substitution that subsumes the individual. In addition, motives themselves could be subject to reasoning and revision; motive manipulation alters them through an unreliable process (conformity to the enhancer's view of what motives are best), rather than through a more reliable process like reasoning.

The way forward, then, is to look for psychological techniques that reliably improve morality without threatening individuality. In Chapters 9 and 10, I argued that reasoning and akrasia reduction are just those sorts of processes and it should indeed be possible to use recent developments in psychology to improve these areas. One promising area is bias reduction. For example, Gaertner et al. (1989) found that intergroup bias among undergraduate subjects could be reduced by framing groups as sets of individuals rather than unified wholes. In another experiment, racial bias was found to be reduced when subjects

were under less cognitive load and could more thoroughly focus on the evaluation itself. (Wyer, Sherman and Stroessner 2000) On the present neutral model of indirect enhancement, these only count as enhancements if the bias is rejected by the participants' own lights. Still, arbitrary groupings and race are not overly controversial cases of bias and so would most often be fairly classified as improvements.

These sorts of interventions respect individuality by relying, first and foremost, on standards accepted by the subjects themselves. The enhancer is not imposing his or her idea of bias on the individuality; rather, he or she is helping the individual actualize his or her own goals. But it is still a moral enhancement (in a non-relativist sense) because we can expect, for reasons outlined above, most of those judgments to be reliable and so correcting perceived internal faults to generally lead to moral improvement. Insofar as reasoning is truth-conducive and bias interferes with reasoning, reducing bias will lead to more reliable moral judgments and – if we avoid akrasia – more moral behavior.

There is not the space here to go into greater detail concerning the whole range of other psychological interventions that might improve reasoning or reduce akrasia. Indeed, such an investigation would be outside my area of competence. My aim here is instead primarily schematic – to outline what features an intervention would have to possess (substantive neutrality and making idea/motive/behavior-formation more reliable) in order to count as acceptable moral enhancement. Researchers hoping for psychological interventions that can improve morality should look for these sorts of indirect interventions, rather than undergo interventions that presuppose disputed moral claims.

**Young Children**

Before moving on to more novel approaches to moral enhancement, a potential objection to the foregoing discussion must be addressed. The foregoing has emphasized that various forms of direct moral enhancement that impose the enhancer's view of morality on the enhancee are illegitimate. However, there is one class of cases where this perspective might seem untenable: young children. It is commonplace for parents to insist to their children that, e.g., sharing is nice and stealing is wrong. We correct children's mistakes, argue from authority ('because I say so') when they dissent and moreover generally think this sort of behavior will help in kids' moral developments. Parents who don't provide direction, the worry is, will have morally stunted children who don't develop the proper ideas later on.

This objection has some force because the view I have been defending implies, against all conventional wisdom, that one should not simply insist to one's children that such-and-such is right and such-and-such is wrong. The recommendation might appear to be that one should simply be a bad parent, overly permissive and providing little structured guidance. This risks the failure to adhere to the most basic societal norms, delinquency and criminality later in life. It suggests that the most powerful tool we have to ensure the next generation is moral – structured guidance of one's malleable, easily influenced children – is shut off.

In response, I would note that the situation for the indirect moral enhancer is not so bleak. In the first place, the acceptance of some forms of punishment provides a possible (partial) paradigm for how to deal with young children. To the extent that one's goal is not the moral improvement of one's children per se, but preventing harms to others, one can accept limited punishments (time out, being sent to one's room, detention, etc.) at the expense of the child's individuality.[105] One could also justify these punishments (and normative instruction in general) not on moral grounds, but prudential – children need to conform to

[105] This would be a substantive moral view. However, the Millian arguments outlined in Chapter 6 apply primarily to those interventions regulating thought and discourse. Whether or not we classify punishment as a direct moral enhancement, it is (for reasons defended in Chapter 6) in its consequentialist, harm-prevention form exempt from the arguments against direct moral enhancements.

social norms if they are to get on in society, and parents have strong reason to promote the interests of their children.[106]

Moreover, there is a positive side to child-rearing suggested by my account that, while still somewhat counter-intuitive, will help ensure one's children don't end up being moral monsters. Recent research has shown that – contra to the work of Piaget and others – young children do actually have something like inclinations that line up with the moral intuitions of adults. In one famous (and striking) study, it was found that infants as young as six months prefer individuals who help others to those who hinder. (Hamlin, Wynn and Bloom 2007) That might be explained purely by self-preservation, but follow-up studies on slightly older babies showed a tendency to want to exact greater harm on those who hindered. (Bloom 2010) The point here is not to defend the controversial claims that morality is innate. Rather, the point is that even young children have intuitions that align at least somewhat with generally-accepted moral views.

These intuitions can be leveraged in the rearing of children to bring about moral improvement without a substantial amount of substantive commitment. Parents need not insist absolutely based on their authority that stealing is wrong. However, they could appeal to their children's own intuitions and engage in a sort of Socratic dialogue – 'Don't you think it's bad when one kid takes another's?' 'How do you think that makes the other child feel?' 'Do you feel sorry?' and so on. The parent would be, to a certain extent, pushing his or her own views – but doing so in a way that is compatible with the model of argumentation defended above.[107]

---

[106] Although this is in itself somewhat weak – it licenses a parenting style akin to Plato's Thrasymachus, where one would violate even the strongest of norms against murder if one could benefit and get away with it.

[107] This is in some ways compatible with Habermas's discourse ethics approach (see, e.g., Habermas 1996 and 2003), though I argue in the next section (contra Habermas) that, among other things, genetic enhancements to reasoning are appropriate and worth promoting.

This dialectical approach can, admittedly, only get the parent so far. Young children are not merely naïve adults; they are in the process of developing the reasoning processes that adults often take for granted. The indirect approach to moral enhancement can easily accommodate this difficulty, however. Helping children's reasoning capacities develop and grow will be a crucial part of moral development, more so than simply laying out rules. Standard classroom education plays a role here, but parental interaction is important as well. The usual practices of reading to one's children, engaging with them, and providing them with a rich environment in which to learn and grow will contribute. Other good practices would include encouraging children to think things through, helping them be attentive to the consequences of their actions, being receptive and honest when (sometimes incessantly) quizzed can all play a role. Inculcating impulse control is also perfectly acceptable on this indirect approach, as that will be a crucial part of reducing akrasia.

The indirect moral enhancement I have been advocating may well lead to substantial revision in parenting practices. Top-down, authoritative models would be replaced by a more reciprocal approach that engages with and encourages children's developing capacities. These revisions may make behavior management a bit more troublesome, but they have the advantage of being focused on the long-term (where appeals to authority may not be appropriate) and relatively respectful of children's autonomy. Children are people too, after all. Their reasoning capacities and individuality are of just as much value as adults, and it is especially important to ensure that they grow up being open to moral revision and refinement so that progress can be made.

**Novel Biological Moral Enhancement**

The question of how to revise our traditional practices may have more current relevant, but much of the present debate over moral enhancement has focused on novel biological interventions of the sort discussed in Chapter 3. These include pharmaceuticals (for which there is presently the most evidence of an impact on moral life, as previously discussed), genetic selection and therapy, as well as neurological therapy (such as with transcranial magnetic stimulation). The implications of the indirect moral enhancement are more or less uniform across these different types of interventions, however, and so I will instead divide this subsection into negative implications, positive implications for reasoning and positive implications for akrasia.

*Negative Implications*

Under a model of direct moral enhancement, non-biological interventions would be classified as moral enhancers based on the extent to which the enhancers (or proponents of enhancement) judge people's resultant beliefs, motives or behaviors. So, in this vein, Persson and Savulescu (2008 & 2010) advocate for an improvement in moral traits such as justice and altruism – traits which are assumed to be good, and whose appropriate levels could apparently be determined by the enhancers. But as argued in Section II, disagreement over those moral issues makes this approach problematic – it is not feasible to get widespread agreement on which traits to enhance and how (particularly when one attempts to spell out the content of such traits, as discussed at the end of Chapter 5), and even if such agreement were achievable it would inhibit moral progress, subvert people's reasoning processes and infringe on their individuality.

State-based coercive imposition of direct moral enhancers is obviously objectionable, as would semi-coercive interventions such as putting a chemical in the water. But these

arguments apply to the softer forms briefly mentioned in Chapter 6, including selective subsidy, propagandistic encouragement of particular interventions and contractual enforcement. Selective subsidy would involve the state paying the cost for some individuals to voluntarily undergo interventions (or provide those interventions to their children). Propaganda would involve public campaigns encouraging voluntary uptake, perhaps based on patriotic duty or more mundane concerns for one's fellow citizen. And contractual enforcement would involve some group paying others to undergo moral enhancement (or making the intervention a compulsory part of employment) in order to bring about morally better ideas, motives or behaviors. All these interventions run into the problems discussed because there is still a presupposition that the enhancers have the correct moral ideas, motives or behaviors while the enhancee does not, with the intention to bring the enhancee into line via a process that is not obviously morally reliable.[108]

A more difficult case involves the subsidy of research into direct moral enhancement. Robert Sparrow (2014) has recently argued that much of the theoretical research into moral enhancement is objectionable is a waste of time, money and intellectual resources and encourages problematic essentialist thinking about morality.[109] Similarly, we might worry that subsidy of empirical research on direct biological moral enhancement is objectionable. If direct biological moral enhancement is to be avoided, then there may not be much point in spending money and human capital on its discoveries. This is only partly right – such research should not be conducted based on the belief by researchers that a particular intervention will lead to moral improvement. However, such research is important insofar as it can identify the salient moral side effects of an intervention. If SSRIs make people less

[108] None of this should be taken to imply that moral self-enhancement should be banned; indeed, I will endorse a form of self-enhancement in the section on akrasia-centric implications.
[109] Sparrow also worries that the research encourages an overly-perfectionist attitude in the population. I do not share these anti-perfectionist concerns; I am not certain that proponents of enhancement need to be perfectionist any more than proponents of a good education need to be.

willing to sacrifice and punish individuals, this is an important effect that patients being prescribed the SSRIs for (say) depression should be made aware of. Whether the effects are a plus or minus (or irrelevant) can be left up to the individual, but the information should be made available. What's more, research can help us better understand the psychological, neurological and genetic underpinnings of our various moral ideas, motives and behavior – topics of enormous interest in themselves.

As for Sparrow's concern that the (more theoretical) proponents of direct moral enhancement are making a mistake even publishing their views: while believe that proponents' arguments are mistaken and – if taken seriously by the population – would be deleterious to society, it is mistaken to think their views should not be aired. My argument rests centrally on the value of disagreement; this includes disagreement over moral enhancement itself. Even mistaken and overall harmful views should be discussed, to the extent that such discussion helps clarify the correct views and revise them in light of previously-unnoticed flaws. For Sparrow, this debate would serve to highlight the political issues (such as inequality) he finds particularly pressing in comparison to moral enhancement; for myself, the literature on direct biological moral enhancement has inspired a refined and (I submit) preferable alternative, indirect moral enhancement.

*Positive implications for reasoning*

Indirect biological moral enhancement is certainly feasible and – based on my arguments – quite desirable. There are a number of ways that such enhancement would work, but the most obvious is simply cognitive enhancement. Cognitive enhancement can take a variety of forms, many of them intersecting with the reasoning processes outlined in Chapter 9. Most promisingly, working memory can be manipulated via pharmaceuticals.

Methylphenidate (a.k.a. Ritalin) has been associated with such improvements in such domains as spatial (Mehta et al. 2004) and digital (Agay et al. 2010) memory. Notably, those memory benefits did not transfer over to other cognitive tasks such as pattern-recognition and risk/reward calculation. Still, improved memory can itself lend improvements to moral reasoning by bolstering empirical competence. Being better able to recall the relevant details in a situation is important for making a proper decision. In addition, if memory is improved during a class that is inculcating other components of reasoning (such as critical analysis or bias), the individual might be in a better position to deploy those lessons later on and thereby improve his or her moral reasoning.

Another potential area for improvement is attention. Originally designed to treat narcolepsy, Modafinil has been shown (in addition to somewhat similar memory benefits as Ritalin) to have beneficial effects on healthy subjects' attentiveness and alertness, especially in individuals who are sleep-deprived. (Repantis et al. 2010) Attention is relevant to reasoning insofar as it allows one to be more mindful of various relevant considerations at play – whether empirical or theoretical. In a discussion, more attentiveness allows one to better track the points of one's interlocutor and thereby (potentially) revise one's views accordingly. And, again, in the classroom, greater attentiveness could assist in internalizing lessons that could improve one's reasoning processes.

Research into other types of interventions such as genetic manipulation and neural therapy are too far off to have strong demonstrable effects on reasoning in humans. However, as our understandings of the genetic and neural correlates of various reasoning processes improve, genetic and neural interventions will become feasible. We should remain open to their use as moral enhancers, to the extent that we find they can reliably improve people's understanding of moral ideas, logical competence, critical thinking and other

capacities. What's more, we should be encouraging research at present into these areas in order to realize the moral benefits more quickly.

If we can manage to identify biological interventions that qualify as indirect moral enhancements, we should – in contrast to the case of direct moral enhancement – be open to soft forms of implementation. It would, for instance, be appropriate and even desirable for the state to subsidize particular interventions that improve reasoning processes. Not only would this help equalize disparities in access to the interventions, but it would also help foster a society that is more reasonable and thus better able to think through the relevant moral considerations. This is perfectly in line with the generally-accepted prerogative of the state to subsidize education for its citizens. It may even be justified to pay certain people to undergo these interventions, as one could expect positive social externalities resulting from better moral reasoning.

But what about more coercive measures? We do, after all, require children to attend school; could there be an analogous requirement to undergo certain biological interventions? There is at least a pro tanto case in favor of this proposal, but it is strongly mitigated in the case of biological interventions because – unlike with, say, schooling – biological interventions invade the body. A better analogy would be required vaccination of children, but even there the analogy breaks down; vaccinations are justified in part because of the direct public health risk from failing to get vaccinated. While I have argued that improving reasoning capacities will be morally beneficial (and can thus be expected to have positive externalities), it cannot be said with any confidence that the magnitude of that benefit is comparable to the risk of being unvaccinated and spreading disease to one's schoolmates.

The pseudo-coercive measure of putting reasoning-enhancing pharmaceuticals in the drinking water is somewhat more palatable. This is not quite as absurd as it sounds; many

countries practice fluoridation of their water supply in order to prevent tooth decay. While a substance is ingested without express consent, the lack of actual force used in that ingestion significantly assuages autonomy-based worries. However, unlike fluoride, reasoning-enhancing pharmaceuticals would of necessity be psychoactive – that is, they would affect the way people think. The way they would do so are not especially pernicious; they are not designed to induce particular thoughts or ideas, but help people think better according to standards few would deny. Still, the fact remains that the process involves a third party imposing their ideas (directly, rather than through a reasoning process) on the individual. A superior option is simply to offer the pharmaceutical for free, and let people take it as they choose – or, to get greater uptake, a policy could require universal uptake unless one opts out (similar to vaccine distribution for schoolchildren in some jurisdictions). This respects autonomy and avoids the problem of third-party imposition of values. It may be less effective than coercive imposition, but such is the cost of adhering to the strictures of morality.

*Positive implications for akrasia*

As with reasoning, there are a number of potentially fruitful pharmacological interventions that might help deal with akrasia. The most well-researched area of this is in treating addiction, which involves akrasia in the case of unwilling addicts who think, all things considered, it would be best to not to succumb to the addiction but do so anyway. For example, in the treatment of opioid addiction (including cocaine and heroin), methodone is used as a relatively-benign substitute for more harmful substances; in combination with therapy, patients can be weaned off their drug habit. (Lobmaier et al. 2010) An alternative approach (more successful for alcohol addiction than opiate addiction) involves reducing the

incentive to consume substance by blocking the euphoric effect of consumption via a drug like naltrexone. (Latt et al. 2002)

This sort of addiction treatment can, in a variety of ways, be expected to lead to moral improvement. Unwilling addicts' all-things-considered judgment that they should not succumb to the addiction could be motivated by entirely prudential considerations, but we can expect moral considerations to play a role as well. Addiction can cause a number of negative externalities; the addicts' desire to succumb can override more pressing moral considerations like obligations to friends and family or compliance with norms against theft. Unwilling addicts will often recognize that – by their own lights – the morally superior option is not to abuse, and seek treatment (in part) to realize that option.[110] In addition, many substances (especially alcohol) will impair judgment and reasoning itself, impairing further moral deliberation and acting as a moral disenhancement.

Alternatively, there is some promise for using pharmacological interventions to promote impulse control. Impulse control is relevant to akrasia insofar as a common cause of akratic action are impulses that override better judgments. Most interest in pharmacological impulse control has focused on treatment of impulse control disorders such as ADHD, but such treatments have had some success in reducing the impulsivity of healthy subjects as well. (Chamberlain et al. 2009) We must be cautious here, however, as such interventions may be too blunt – while impulses may sometimes lead to akrasia, perhaps at other times they support morally reliable procedures by reinforcing the urgency of basic moral intuitions. Some all things considered judgments may well be bolstered by the right sort of impulses. More study in this area would be needed before general impulsivity control could be mobilized as a method for reducing akrasia.

[110] These externalities may well justify coercive imposition of addiction treatment along the same line as consequentialist punishment (harms to others), though such coercion relies on the dubious proposition that addiction treatment can be effective when it is forced.

Indirect moral enhancement via akrasia reduction could also take a more passive form: allowing a diverse range of direct moral self-enhancements. Individuals would identify what they take to be the most morally astute ideas, motives or behaviors and seek out interventions (including pharmaceuticals, neural therapy or genetic manipulation) that bring themselves more in line with that ideal. This proposal may seem to fly in the face of the objections from Section II, but in fact it avoids those pitfalls because it lacks the crucial feature of a third party imposing its moral views on the agent. If instead the state simply permits individuals to directly morally enhance themselves – with no presupposition of what constitutes improvement – we can expect general reduction in akrasia. The sort of person who would voluntarily undergo direct moral enhancement is indeed most likely in an akratic state; they recognize they possess a particular moral failing that leads to bad behavior and seek novel remedies to correct that behavior. The difficulty of obtaining moral agreement is avoided because the laissez-faire policy is not committed to any particular moral view, instead relying on individuals to determine that for themselves. Because of the diversity of moral views in the population, we don't have to worry that allowing such self-enhancement will impede progress. And because people are deciding for themselves, the direct moral enhancement does respect people's individuality as well as their reasoning – the decision to self-enhance would, after all, be the result of a conscious, deliberative decision on the part of the subject to seek a particular improvement.

This argument, then, is in the final analysis compatible with the sort of pro-enhancement argument put forward by Tom Douglas (2008). Douglas's argument, after all, is entirely concerned with how individuals can improve themselves. I agree, though not because of confidence in any given plan of direct moral enhancement. Rather, passively allowing (and perhaps providing research funding to provide the empirical basis for) people to choose which (if any) direct moral enhancements to undergo is a form of akrasia reduction.

Because we can have general confidence in people's basic intuitions as well as the superiority of all-things-considered judgments over mere inclinations that induce akrasia, we can expect allowing direct self-enhancement will generally lead to moral improvement. It may turn out that not many people will be willing to undergo such interventions, and we should not be pushing particular direct enhancements on people, but those who do decide to self-enhance would indeed be undertaking a reliable form of indirect moral enhancement.

\

# Chapter 12: Conclusions

My primary aims in this work have been twofold. First, I attempted to show how the existence and value of disagreement problematize direct moral enhancement, which is grounded in assumptions about the correct moral ideas, motives and/or behaviors. Chapters 5 and 6 contain the main content for this critique, which is both practical and normative. Practically speaking, the existence of disagreement over many fundamental moral issues will make a broad project of moral enhancement infeasible in a democratic society. Even if we can identify general norms everyone accepts (such as justice or altruism), securing agreement on the details of those norms – their content, scope, strength, and so on – would be quite difficult. More normatively, I outlined a Millian case against widespread direct moral enhancement – it could lead to premature convergence and inhibit moral growth, it typically interferes with people's reasoning processes, and it is disrespectful of people's individuality. Taken together, this leads me to skepticism that there is a reasonable path forward for large-scale projects of moral enhancement.

Second, I provided an alternate indirect framework that avoids those problems in Chapters 7-10. Indirect moral enhancement is distinct because it is not grounded in a set of particular substantive claims. Instead, it posits much more minimally and theoretically that certain processes are more morally reliable than others. The reliability of the processes is not defined in terms of the moral ideas, motives or behaviors they produce but our theoretical reasons for thinking they will generally lead to more moral ideas, motives or behaviors. In this work, I have identified reasoning and akrasia as prime targets for indirect moral enhancement. Starting from the idea that our intuitions are generally reliable (defended in Chapter 8), we can see how more logically coherent, empirically informed, critical and unbiased people will generally be more moral (Chapter 9). And if our reasoned judgments

can be trusted over base inclinations, we should also think that avoiding acting against those all-things-considered judgments will generally lead to improvement. (Chapter 10)

These two aims do not stand or fall together; even if I am mistaken about the flaws of direct moral enhancement, we would still have reason to pursue indirect enhancement. Conversely, it might be that – despite what I have argued – indirect moral enhancement cannot avoid the difficulties of disagreement. This would likely mean that any form of moral enhancement has substantial, unavoidable flaws that provide significant reason against its pursuit. However, taken together, the two aims paint a more optimistic picture: though the direct approach to moral enhancement is flawed, there is a way forward that is worth pursuing.

Along the way, I have tried to elaborate on ancillary issues relevant to the debate over moral enhancement. Chapters 2 and 3 highlighted the distinction between traditional forms of moral enhancement like education and more novel, biological interventions. The two approaches do make for distinct implications (as explicated in the previous chapter), but we should be careful not to overstate how far apart they are. My arguments were meant to apply similarly to both traditional and biological forms of moral enhancement. This is reflective of the general position in the enhancement debate that arguments for and against biological interventions can be similarly applied to traditional interventions, education in particular. Opponents of various forms of enhancement should be prepared to criticize traditional practices that may enjoy widespread support. I have endorsed this approach in the previous chapter, arguing for an indirect approach to moral education that may go against popular wisdom – but such an approach would not necessarily require as radical reforms as it might appear, and is in many ways quite attractive.

In addition, some chapters will be of interest even if both the central arguments fail. Chapter 4 emphasizes the amount of disagreement that is inevitable over the standards for moral enhancement, but it also serves to illustrate the various paths direct moral enhancement would take depending on one's metaethical, normative and practical views. If moral enhancement becomes widespread but not centrally regulated, we might expect that sort of divergence to crop up in various subcultures. And Chapter 8 lays out a position – moral optimism – that will be of interest beyond the moral enhancement debate. If we all tacitly accept that intuitions (broadly construed) are generally reliable, this may provide some support for intuitionist approaches to moral philosophy. This will be unsatisfying to some (no particular explanation for why they are reliable has been given), but it at least suggests a way forward in the debate over intuitionism.

Despite my frequent appeals to neutrality, the upshot of this work might appear to be quite rationalist. I have defended the importance of reasoning, argued that better reasoning processes will generally lead to moral improvement, and suggested a model for moral education that is more or less about teaching people to reason better. Sentimentalist approaches were addressed but set aside in Chapter 7. Still, I believe it would be fairer to characterize my approach as *philosophical*, rather than rationalist. That is, it takes very seriously the ability of people to reflect on, analyze and revise their opinions and behavior. This is not to say that philosophy should dominate students' curriculum. Better thinking can be brought about in a wide variety of contexts – history, literature, biology, mathematics and so on. But ultimately, I have proposed in this dissertation that we can become better people by, essentially, becoming better philosophers.

# Bibliography

Agay, N., Yechiam, E., Carmel, Z., & Levkovitz, Y. (2010). Non-specific effects of methylphenidate (Ritalin) on cognitive ability and decision-making of ADHD and healthy adults. *Psychopharmacology*, *210*(4), 511–519. doi:10.1007/s00213-010-1853-4

Aristotle. (1999). *Nicomachean Ethics*. (T. Irwin, Trans.). Indianapolis/Cambridge: Hackett Publishing Company.

Arpaly, N. (2000). On acting rationally against one's best judgment. *Ethics*, *110*(3), 488–513.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership and men*. Pittsburgh, PA: Carnegie Press.

Audi, R. (1979). Weakness of will and practical judgment. *Nous*, *13*(2), 173–196.

Audi, R. (1990). Weakness of will and rational action. Australasian Journal of Philosophy, 68(3), 270-281.

Audi, R. (2001). *The architecture of reason: the structure and substance of rationality*. Oxford; New York: Oxford University Press.

Audi, R. (2004). *The good in the right: a theory of intuition and intrinsic value*. Princeton, N.J: Princeton University Press.

Ayer, A. J. (1936). *Language, truth and logic*. London: Victor Gollancz.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin Shapes the Neural Circuitry of Trust and Trust Adaptation in Humans. *Neuron*, *58*(4), 639–650. doi:10.1016/j.neuron.2008.04.009

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics*. New York: Oxford University Press.

Bennett, J. (1974). The Conscience of Huckleberry Finn. *Philosophy*, *49*(188), 123–134.

Blackburn, S. (1984). *Spreading the word : groundings in the philosophy of language*. Oxford: Clarendon Press.

Blackburn, S. (1993). *Essays in quasi-realism*. New York: Oxford University Press.

Bloom, P. (2010, May 5). The Moral Life of Babies. *The New York Times*. Retrieved from http://www.nytimes.com/2010/05/09/magazine/09babies-t.html?pagewanted=all&_r=0

Bloomfield, P. (2009). Archimedeanism and Why Metaethics Matters. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics. Volume 4* (pp. 283–302). Oxford: Oxford University Press. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=472377

Bogardus, T. (2009). A Vindication of the Equal-Weight View. *Episteme*, *6*(03), 324–335. doi:10.3366/E1742360009000744

Brandt, R. B. (1990). The Science of Man and Wide Reflective Equilibrium. *Ethics*, *100*(2), 259–278.

Bratman, M. (1979). Practical reasoning and weakness of the will. *Nous*, *13*(2), 153–171.

Campbell, R., & Kumar, V. (2012). Moral Reasoning on the Ground. *Ethics*, *122*(2), 273–312.

Cave, E. M. (2007). What's wrong with motive manipulation? *Ethical Theory and Moral Practice*, *10*(2), 129–144.

Chamberlain, S. R., Hampshire, A., Müller, U., Rubia, K., del Campo, N., Craig, K., … Sahakian, B. J. (2009). Atomoxetine Modulates Right Inferior Frontal Activation During Inhibitory Control: A Pharmacological Functional Magnetic Resonance Imaging Study. *Biological Psychiatry*, *65*(7), 550–555. doi:10.1016/j.biopsych.2008.10.014

Chen, F. S., Kumsta, R., & Heinrichs, M. (2011). Oxytocin and intergroup relations: goodwill is not a fixed pie. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(13), E45; author reply E46. doi:10.1073/pnas.1101633108

Chow, Y. W., Pietranico, R., & Mukerji, A. (1975). Studies of oxygen binding energy to hemoglobin molecule. *Biochemical and biophysical research communications*, *66*(4), 1424–1431.

Clouser, K. Danner, & Gert, Bernard. (1990). A critique of principalism. *Journal of Medicine and Philosophy*, *15*(2), 219–236.

Cohen, B. (1983). Ethical Objectivity and Moral Education. *Journal of Moral Education*, *12*(2), 131–136. doi:10.1080/0305724830120210

Crisp, Roger. (2011). Reasonable Disagreement: Sidgwick's Principle and Audi's Intuitionism. In Hernandez, Jill Graper (Ed.), *The New Intuitionism* (pp. 240–265). London: Continuum Books.

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. doi:10.1073/pnas.1009396107

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. doi:10.1126/science.1155577

Crutchfield, R. S. (1955). Conformity and character. *American Psychologist*, *10*(5), 191–198. doi:10.1037/h0040237

D'Arms, J., & Jacobson, D. (2000). Sentiment and Value. *Ethics*, *110*(4), 722–748.

Daniels, N. (1979). Wide Reflective Equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy*, *76*(5), 252–282.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, *108*(17), 6889–6892. doi:10.1073/pnas.1018033108

Davidson, D. (2001). How is weakness of the will possible? In *Essays on Actions and Events* (pp. 21–42). Oxford: Oxford University Press.

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. doi:10.1073/pnas.1015316108

DeGrazia, D. (2003). Common morality, coherence, and the principles of biomedical ethics. *Kennedy Institute of Ethics journal*, *13*(3), 219–230.

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

Douglas, T. (2011). Moral enhancement via direct emotional modulation: a reply to John Harris. *Bioethics*, *Online first*, no–no. doi:10.1111/j.1467-8519.2011.01919.x

Douglas, T. (2013). Enhancing Moral Conformity and Enhancing Moral Worth. *Neuroethics*. doi:10.1007/s12152-013-9183-y

Douglas, T. (2014). Enhancing Moral Conformity and Enhancing Moral Worth. *Neuroethics*, *7*(1), 75–91. doi:10.1007/s12152-013-9183-y

Dreier, J. (2002). Meta-Ethics and Normative Commitment. *Nous*, *36*(s1), 241–263. doi:10.1111/1468-0068.36.s1-1.11

Dreier, J. (1990). Internalism and Speaker Relativism. *Ethics*, *101*, 6–26.

Dworkin, R. (1996). Objectivity and Truth: You'd Better Believe It. *Philosophy and Public Affairs*, *25*(2), 87–139. doi:10.1111/j.1088-4963.1996.tb00036.x

Ebels-Duggan, K. (2013). Moral Education in the Liberal State. *Journal of Practical Ethics*, *1*(2), 34–63.

Ehrenberg, K. M. (2008). Archimedean metaethics defended. *Metaphilosophy*, *39*(4-5), 508–529. doi:10.1111/j.1467-9973.2008.00558.x

Ehrlich, I. (1975). The Deterrent Effect of Capital Punishment: A Question of Life and Death. *The American Economic Review*, *65*(3), 397–417.

Elga, A. (2007). Reflection and disagreement. *Nous*, *41*(3), 478–502.

Emanuel, E. (1997, March). Whose right to die? *The Atlantic Monthly*.

Emanuele, E., Brondino, N., Bertona, M., Re, S., & Geroldi, D. (2008). Relationship between platelet serotonin content and rejections of unfair offers in the ultimatum game. *Neuroscience Letters*, *437*(2), 158–161. doi:10.1016/j.neulet.2008.04.006

Enoch, D. (2010). Not Just a Truthometer: Taking Oneself Seriously (but not Too Seriously) in Cases of Peer Disagreement. *Mind*, *119*(476), 953–997. doi:10.1093/mind/fzq070

Fabre, C. (2006). *Whose body is it anyway? : justice and the integrity of the person*. Oxford: Clarendon Press.

Fantl, J. (2006). Is metaethics morally neutral? *Pacific Philosophical Quarterly*, *87*(1), 24–44. doi:10.1111/j.1468-0114.2006.00246.x

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5–20.

Fröding, B. E. E. (2011). Cognitive Enhancement, Virtue Ethics and the Good Life. *Neuroethics*, *4*(3), 223–234. doi:10.1007/s12152-010-9092-2

Gaertner, S. L., Dovidio, J. F., Anastasio, P. A., Bachman, B. A., & Rust, M. C. (1993). The Common Ingroup Identity Model: Recategorization and the Reduction of Intergroup Bias. *European Review of Social Psychology*, *4*(1), 1–26. doi:10.1080/14792779343000004

Gehler, J., Cantz, M., O'Brien, J. F., Tolksdorf, M., & Spranger, J. (1975). Mannosidosis: clinical and biochemical findings. *Birth defects original article series*, *11*(6), 269–272.

George, R. P. (1993). *Making men moral: civil liberties and public morality*. Oxford: Clarendon Press.

Gibbard, A. (1990). *Wise Choices, Apt Feelings*. Oxford: Clarendon Press.

Gilead, T. (2009). Progress or stability? An historical approach to a central question for moral education. *Journal of Moral Education*, *38*(1), 93–107. doi:10.1080/03057240802399483

Goldman, A. I. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and knowledge: new studies in epistemology* (pp. 1–23). Dordrecht, Holland ; Boston: D. Reidel Pub. Co.

Goldman, A. I. (2001). Experts: Which Ones Should You Trust? *Philosophy and Phenomenological Research*, *63*(1), 85–110. doi:10.1111/j.1933-1592.2001.tb00093.x

Habermas, J. (1996). *Between facts and norms: contributions to a discourse theory of law and democracy*. Cambridge, UK: Polity Press.

Habermas, J. (2003). *The future of human nature*. Cambridge, UK: Polity.

Haidt, J. (2013). *The righteous mind: why good people are divided by politics and religion*. London: Penguin.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559. doi:10.1038/nature06288

Hare, R. M. (1963). *Freedom and Reason*. Oxford: Oxford University Press.

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method and Point*. Oxford: Clarendon Press.

Hare, R. M. (1952). *The Language of Morals*. Oxford: Clarendon Press.

Harman, G., Sinott-Armstrong, W., & Mason, K. (2010). Moral Reasoning. In J. M. Doris, F. Cushman, & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 205–244). Oxford; New York: Oxford University Press.

Harris, J. (2011). Moral Enhancement and Freedom. *Bioethics*, *25*(2), 102–111. doi:10.1111/j.1467-8519.2010.01854.x

Hays, S., Miller, C., & Cobb, M. (2011). 2008 National Nanotechnology Survey on Enhancement Results Overview. In *The Yearbook of Nanotechnology, Volume III: Nanotechnology, the Brain, and the Future*. Springer.

Hill, T. (1989). The Kantian conception of autonomy. In Christman, John (Ed.), *The Inner Citadel: Essays on Individual Autonomy* (pp. 91–108). New York and Oxford: Oxford University Press.

Hills, A. (2009). Moral testimony and moral epistemology. *Ethics*, *120*(1), 94–127.

Hinchman, E. S. (2013). Rational requirements and "rational" akrasia. *Philosophical Studies*, *166*(3), 529–552. doi:10.1007/s11098-012-9993-5

Holm, S. (1995). Not just autonomy--the principles of American biomedical ethics. *Journal of Medical Ethics*, *21*(6), 332–338. doi:10.1136/jme.21.6.332

Huang, Y. (2011). Can virtue be taught and how? Confucius on the paradox of moral education. *Journal of Moral Education*, *40*(2), 141–159. doi:10.1080/03057240.2011.568096

Hume, D. (1742). The Sceptic. In *Essays, Moral, Political, and Literary*. Indiannapolis: Liberty Fund, inc.

Hume, D. (1758). *The Complete Works and Correspondence of David Hume. Electronic edition. New Letters of David Hume*. (T. H. Green, T. H. Grose, & N. K. Smith, Eds.). Charlottesville, Va.: InteLex Corporation.

Hunter, J. F. M. (1974). The Possibility of a Rational Strategy of Moral Persuasion. *Ethics*, *84*(3), 185–200.

Israel, S., Lerer, E., Shalev, I., Uzefovsky, F., Riebold, M., Laiba, E., … Ebstein, R. P. (2009). The Oxytocin Receptor (OXTR) Contributes to Prosocial Fund Allocations in the Dictator Game and the Social Value Orientations Task. *PLoS ONE*, *4*(5), e5535. doi:10.1371/journal.pone.0005535

Jensen, H. (1977). Hume on moral agreement. *Mind*, *86*(344), 497–513.

Jones, K. (2003). Emotion, Weakness of Will, and the Normative Conception of Agency. In *Philosophy and the Emotions* (pp. 181–200). Cambridge: Cambridge University Press.

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

Kant, I. (2002). *Groundwork for the metaphysics of morals*. (A. W. Wood & J. B. Schneewind, Eds.). New Haven: Yale University Press.

Karlsen, J. R., & Solbakk, J. H. (2011). A waste of time: the problem of common morality in Principles of Biomedical Ethics. *Journal of Medical Ethics*, *37*(10), 588–591. doi:10.1136/medethics-2011-100106

Kimelberg, H. K. (1975). Alterations in phospholipid-dependent (Na+ +K+)-ATPase activity due to lipid fluidity. Effects of cholesterol and Mg2+. *Biochimica et biophysica acta*, *413*(1), 143–156.

King, N. L. (2012). Disagreement: What’s the Problem? or A Good Peer is Hard to Find: *Philosophy and Phenomenological Research*, *85*(2), 249–272. doi:10.1111/j.1933-1592.2010.00441.x

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*(7042), 673–676. doi:10.1038/nature03701

Krakowski, M. (2003). Violence and Serotonin: Influence of Impulse Control, Affect Regulation, and Social Functioning. *Journal of Neuropsychiatry*, *15*(3), 294–305. doi:10.1176/appi.neuropsych.15.3.294

Kravitz, David A., & Gunto, Samuel. (1992). Decisions and Perceptions of Recipients in Ultimatum Bargaining Games. *The Journal of Socio-Economics*, *21*(1), 65–84.

Latt, N. C., Jurd, S., Houseman, J., & Wutzke, S. E. (2002). Naltrexone in alcohol dependence: a randomised controlled trial of effectiveness in a standard clinical setting. *The Medical Journal of Australia*, *176*(11), 530–534.

Lee, P., & George, R. (2005). The wrong of abortion. In A. I. Cohen & C. H. Wellman (Eds.), *Contemporary debates in applied ethics* (pp. 13–26). Malden, MA: Blackwell Pub.

Lickona, T. (1996). Eleven Principles of Effective Character Education. *Journal of Moral Education*, *25*(1), 93–100. doi:10.1080/0305724960250110

Lobmaier, P., Gossop, M., Waal, H., & Bramness, J. (2010). The pharmacological treatment of opioid addiction—a clinical perspective. *European Journal of Clinical Pharmacology*, *66*(6), 537–545. doi:10.1007/s00228-010-0793-6

López, B. G., & López, R. G. (1998). The Improvement of Moral Development Through an Increase in Reflection. A Training Programme. *Journal of Moral Education*, *27*(2), 225–241. doi:10.1080/0305724980270207

MacDonald, K., & MacDonald, T. M. (2010). The Peptide That Binds: A Systematic Review of Oxytocin and its Prosocial Effects in Humans. *Harvard Review of Psychiatry*, *18*(1), 1–21. doi:10.3109/10673220903523615

Mackie, J. L. (1977). *Ethics inventing right and wrong*. Harmondsworth: Penguin Books.

Marquis, D. (1989). Why abortion is immoral. *The journal of philosophy*, *86*(4), 183–202.

McGrath, S. (2008). Moral disagreement and moral expertise. In Shafer-Landau, Russ (Ed.), *Oxford studies in metaethics: volume 3* (pp. 87–108). Oxford: Oxford University Press.

McIntyre, A. (2006). What is wrong with weakness of will? *The Journal of Philosophy*, *103*(6), 284–311.

McMahan, J. (2002). *The ethics of killing : killing at the margins of life*. Oxford: Oxford University Press.

Mehta, M. A., Goodyer, I. M., & Sahakian, B. J. (2004). Methylphenidate improves working memory and set-shifting in AD/HD: relationships to baseline memory capacity. *Journal of Child Psychology and Psychiatry*, *45*(2), 293–305. doi:10.1111/j.1469-7610.2004.00221.x

Mellström, C., & Johannesson, M. (2008). Crowding Out in Blood Donation: Was Titmuss Right? *Journal of the European Economic Association*, *6*(4), 845–863. doi:10.1162/JEEA.2008.6.4.845

Mill, J. S. (1999). *On liberty*. (E. Alexander, Ed.). Peterborough, Ont.: Broadview Press. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=34092

Moore, G. E. (1903). *Principia ethica*. Cambridge: Cambridge University Press.

Moore, M. S. (1987). The Moral Worth of Retribution. In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press.

Morgan, S. (2009). Can there be a Kantian consequentialism? In J. Suikkanen & J. Cottingham (Eds.), *Essays on Derek Parfit's On what matters* (pp. 39–60). Chichester, West Sussex, U.K.; Malden, MA: Wiley-Blackwell. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=480443

Nozick, R. (1974). *Anarchy, state, and Utopia*. Oxford: Blackwell.

Nozick, R. (1993). *The nature of rationality*. Princeton, N.J: Princeton University Press.

Otsuka, M. (2009). The Kantian argument for consequentialism. In J. Suikkanen & J. Cottingham (Eds.), *Essays on Derek Parfit's On what matters*. Chichester, West Sussex, U.K.; Malden, MA: Wiley-Blackwell. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=480443

Parfit, D. (2011). *On what matters*. Oxford; New York: Oxford University Press.

Passell, P. (1975). The deterrent effect of the death penalty: A statistical test. *Stanford Law Review, 28*(1), 61–80.

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

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

Persson, I., & Savulescu, J. (2012). *Unfit for the future: the need for moral enhancement*. Oxford: Oxford University Press.

Plato. (1997). *Plato: Apology of Socrates*. (M. C. Stokes, Trans.). Warminster: Aris & Phillips.

Porcellati, G. (1976). Membrane lipids and metabolic processes. *Biochimie*, *58*(8), 981–987.

Preston-Roedder, R. (2013). Faith in Humanity. *Philosophy and Phenomenological Research*, *87*(3), 664–687. doi:10.1111/phpr.12024

Prinz, J. (2006). The Emotional Basis of Moral Judgments. *Philosophical Explorations*, *9*(1), 29–43.

Prinz, J. J. (2007). *The emotional construction of morals*. Oxford: Oxford University Press.

Rachels, James. (1975). Active and passive euthanasia. *New England Journal of Medicine*, *292*, 78–80.

Rawls, J. (1988). Classical utilitarianism. In Scheffler, Samuel (Ed.), *Consequentialismand its Critics*. Oxford: Oxford University Press.

Rawls, J. (1999). *A theory of justice*. Cambridge, Mass.: Belknap Press of Harvard University Press. Retrieved from http://site.ebrary.com/id/10318468

Raz, J. (1999). *Engaging reason: on the theory of value and action*. Oxford: Oxford University Press.

Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological research: the official journal of the Italian Pharmacological Society*, *62*(3), 187–206. doi:10.1016/j.phrs.2010.04.002

Ross, J. (2009). Should Kantians be consequentialists? In J. Suikkanen & J. Cottingham (Eds.), *Essays on Derek Parfit's On what matters* (pp. 144–153). Chichester, West Sussex, U.K.; Malden, MA: Wiley-Blackwell. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=480443

Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Scanlon, T. M. (2011). How I Am Not a Kantian. In *On what matters*. Oxford: Oxford University Press.

Scheffler, S. (1988). Introduction. In S. Scheffler (Ed.), *Consequentialismand its Critics*. Oxford: Oxford University Press.

Scheffler, S. (1994). *The rejection of consequentialism : a philosophical investigation of the considerations underlying rival moral conceptions*. Oxford: Clarendon Press.

Schleifer, M., & Douglas, V. I. (1973). Moral judgments, behaviour, and cognitive style in young children. *Canadian Journal of Behavioural Science*, *5*, 133–144.

Sherman, N. (1999). Character development and Aristotelian virtue. In D. Carr & J. W. Steutel (Eds.), *Virtue Ethics and Moral Education* (pp. 35–48). London: Routledge.

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

Sidgwick, H. (1907). *The Methods of Ethics, 7th ed.* London: Macmillan.

Simpson, R. M. (2013). Epistemic peerhood and the epistemology of disagreement. *Philosophical Studies*, *164*(2), 561–577. doi:10.1007/s11098-012-9869-8

Singer, Peter. (1972). Famine, affluence and morality. *Philosophy and Public Affairs*, *1*(3), 229–243.

Smith, M. (1994). *The moral problem*. Oxford, UK; Cambridge, Mass., USA: Blackwell.

Sparrow, R. (2-14). Egalitarianism and Moral Bioenhancement. *The American Journal of Bioethics, Forthcoming*.

Street, Sharon. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies*, *127*(1), 109–166.

Strike, K. A. (1999). Trust, Traditions and Pluralism: Human Flourishing and Liberal Polity. In D. Carr & J. Steutel (Eds.), *Virtue Ethics and Moral Education*. New York: Routledge.

Sturgeon, N. (1986). What difference does it make whether moral realism is true? *The Southern Journal of Philosophy*, *24*(S1), 115–141. doi:10.1111/j.2041-6962.1986.tb01600.x

Sullivan, Thomas. (1977). Active and passive euthanasia: an impertinent distinction? *Human Life Review*, *3*(3), 40–46.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and brain sciences*, *28*, 531–573.

Thomson, J. J. (n.d.). A Defense of Abortion. *Philosohy and Public Affairs*, *1*(1), 47–66.

Tsai, D. F. (1999). Ancient Chinese medical ethics and the four principles of biomedical ethics. *Journal of medical ethics*, *25*(4), 315–321.

Waldren, M. S. (2013). Why Liberal Neutralists Should Accept Educational Neutrality. *Ethical Theory and Moral Practice*, *16*(1), 71–83. doi:10.1007/s10677-011-9329-0

Walker, M. (2009). Enhancing genetic virtue. *Politics and the Life Sciences*, *28*(2), 27–47. doi:10.2990/28_2_27

Wall, G. (1975). Moral Authority and Moral Education. *Journal of Moral Education*, *4*(2), 95–99. doi:10.1080/0305724750040201

Wedgwood, Ralph. (2010). The moral evil demons. In Feldman, Richard & Warfield, Ted A. (Eds.), *Disagreement* (pp. 216–246). Oxford: Oxford University Press.

Williams, B. (1985). *Ethics and the limits of philosophy*. Cambridge, Mass: Harvard University Press.

Wolf, S. (1982). Moral saints. *The Journal of Philosophy*, *79*(8), 419–439.

Wyer, N. A., Sherman, J. W., & Stroessner, S. J. (2000). The Roles of Motivation and Ability in Controlling the Consequences of Stereotype Suppression. *Personality and Social Psychology Bulletin*, *26*(1), 13–25. doi:10.1177/0146167200261002

Zak, P. J., Kurzban, R., Ahmadi, S., Swerdloff, R. S., Park, J., Efremidze, L., … Matzner, W. (2009). Testosterone Administration Decreases Generosity in the Ultimatum Game. *PLoS ONE*, *4*(12), e8330. doi:10.1371/journal.pone.0008330

Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin Increases Generosity in Humans. *PLoS ONE*, *2*(11), e1128. doi:10.1371/journal.pone.0001128

Zangwill, N. (2012). Rationality and moral realism. *Ratio*, *25*(3), 345–364. doi:10.1111/j.1467-9329.2012.00546.x