ETHICAL ISSUES IN MORAL AND SOCIAL ENHANCEMENT

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy

in the Faculty of Humanities and Social Sciences

2015

Anna Pacholczyk

School of Law

Contents

ABSTRACT

University of Manchester
Anna Pacholczyk
Doctor of Philosophy in Law
Ethical issues in moral and social enhancement
30.10.2014

Recent developments in social neuroscience have stirred up increased interest within the bioethical debate (for a review see: Specker et al. 2014). Moral enhancement is a concept that directly embodies the idea of making brain science work for the social and moral good. In recent ethical discussions about biomedical means of moral enhancement, scholars have focused on so called ‘direct means of moral enhancement,’ discussing the ethical permissibility of modifying the emotional underpinnings of moral behaviour (Douglas, 2008; 2013; Persson and Savulescu, 2008; Savulescu and Persson, 2012a; 2012b). However, critics have argued that such modification only seems like moral enhancement, that behavioural modification is not ‘true’ moral enhancement, for the reason that it changes behaviours without making agents better moral agents. Critics have also noted that it can undermine freedom (e.g. Harris, 2011; see also: Douglas, 2014). This thesis addresses the ethical issues relating to enhancement. In the first part of this work I consider conceptual issues surrounding the concept of moral enhancement and argue that moral enhancement is plausible if we adjust our expectations to match those we have of cognitive enhancement. I examine the difference between pro-sociality and morality, and argue that an increase in empathy and reduction in anger cannot be seen as straightforward moral enhancements. The second part examines the objections related to moral disagreement, medicalization and narrative identity. The third part of this work focuses of the issues related to freedom and agency. I argue that voluntary direct emotion modulation, if embedded in appropriate reflection, is a *prima facie* desirable way of moral enhancement.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT STATEMENT

Introduction

Moral and social enhancement has recently emerged as a topic of public discussion and applied ethical examination. In response to global problems, and inspired by the increased focus on social neuroscience, many authors put forward proposal about how to address the worlds' pressing problems and make us better people.

Jeremy Rifkin in *Empathetic Civilization, Rethinking Human Nature in the Biosphere Era* (2010), argued that we need to extend 'empathic consciousness' (p. 168) and moving away from the Enlightenment idea of humans as rational, self-interested, and utilitarian, is what is needed to face the challenges the humanity is now facing. Simon Baron-Cohen (2011), a psychologist famous for his research on Autism and empathy, proposes that we explain evil in terms of low levels of empathetic ability. Further, in his *Science of Evil: On Empathy and the Origins of Cruelty,* Baron-Cohen proposes that more empathy would make better people and better society. Patricia Churchland, a philosopher known for her eliminativism[1] in the philosophy of mind, in *Braintrust: What Neuroscience Tells Us about Morality* (2011) argues that morality originates in the biology of the brain. Moral values, Churchland argues, are rooted in a behaviour common to all mammals the caring for offspring.

In this context, the applied ethical discussion of enhancement, typically focused on cognitive, mood and memory enhancement, was extended to considering the prospects of moral and social enhancement. In 2008, Douglas published a paper on MB in which he argued that enhancement of moral motives would be an enhancement that escapes the objections of the so called 'bioconservative' critics of

[1] Eliminative approaches assume that beliefs, desires, hopes, and fears are theoretical constructs of a theory of mind called 'folk psychology.' This theory and all its constructs are seen as having little value and candidates for elimination. According to eliminativism, the theory should be replaced by a better theory of 'neuroscience,' which would replace the obsolete categories by a better system of neurobiological categories. See for example, Churchland (1986).

enhancement. Further, Persson and Savulescu (2008) argued that since cognitive enhancement and the increased access to technology means that the society is increasingly exposed to the risk of large scale harm perpetuated by a small minority, we have a reason to complement cognitive enhancement with moral improvement aimed at preventing such harm. They argue that this gives us a reason to pursue moral bioenhancement (MB) and propose enhancing social sentiments and capacities such as empathy and sense of fairness. In their 2012 book (Savulescu and Persson, 2012b), they also propose that MB could help to solve the great problems facing humanity such as climate change and poverty.

The ethical examination of moral and social enhancement has been further fuelled by the development in social neuroscience over the past two decades, newly budding interest of neuroscientists in moral judgement and decision-making, the boom in brain imaging research, the emergence of empirical philosophy and the increasing popularity of the diagnosis of asocial personality disorder in the context of public safety and criminal justice.

The goal of this thesis is to examine the arguments put forward by the proponents of MB as well as those who raise ethical objections and doubts about the plausibility of MB aimed at enhancing moral agency. I will focus on MB that targets the underpinnings of social action, such as emotions – what some scholars have called direct emotion modulation (Douglas, 2013). Although I will at points examine ethical criticism related to the possibility of compulsory use of MB, in this thesis I will focus mainly on examining the ethical dimension of voluntary MB. The reason for this is that although the compulsory use adds another layer of ethical challenge, the examination of issues raised by voluntary use is foundational. More discussion is needed is in this area and such discussion will likely enrich further considering issues related to compulsory use. Moreover, the discussion of compulsory use of neurotechnologies would incorporate the arguments already put forward in the area of compulsory measures in mental health, crime prevention and rehabilitation of offenders and require additional reference to political theory ideas that underpin such arguments. In the service of the depth of the argument, this thesis focuses mainly on the voluntary aspects of MB, with reference to issues raised by compulsory use where appropriate.

Additionally, the current work's leading emphasis on the voluntary agent-led MB (see especially Chapter 8) is a result of a choice made to address issues from the perspective of a potential user when she is faced with a decision of whether or not to use MB (see Chapter 4 in which I argue that despite moral disagreement such decisions will have to be made). I have attempted to focus on the areas that might be especially problematic for the potential user, at the same time adding to the existing scholarship on the topic. I have aimed to use philosophical and applied ethical methods in the service of elucidating, analysing and responding to the doubts potential users of MB may be faced with.

There are two predominant methods proposed for MB via direct emotion modulation: pharmacological enhancement (via administration of oxytocin and selective serotonin reuptake inhibitors [SSRIs] for example) and enhancement using brain stimulation (e.g. modulating amygdala activation). Pharmacological agents that are potential moral enhancers are already here shall we wish to use them (e.g. SSRIs and oxytocin) and others will likely emerge as our *pharmacopeia* toolkit widens.

Brain stimulation is a more speculative mode of MB. In this work I will focus on deep brain stimulation (DBS) and not on non-invasive brain stimulation methods, because emotion modulation is more likely to involve stimulation of structures deep in the brain. If developments in technology allow non-invasive stimulation of those brain structures and circuits, this may change the balance of costs and benefits, yet the more general arguments presented in this thesis will also apply to those methods. The current deep brain stimulation technology is mostly used for movement disorders and Parkinson's and the move towards application in neuropsychiatric disorders has met with considerable controversy (see: Pacholczyk, 2015). However, the behavioural changes observed in treatment of other disorders as well as the research on stimulation-induced behaviour and mood change, indicate that attempts at MB are within reasonable reach. The application of brain stimulation for such purpose in the near future is likely to be slow, if any. This is partially due to high surgical and post-surgical risks related to the necessity for surgically implanting electrodes, partially due to the history of brain surgery and related professional and social resistance and partially due to high cost of the procedure and subsequent check-ups and maintenance of the stimulation device.

Ablative procedures to decrease aggressive behaviour (procedures in which a part of the brain is destroyed) has been largely stopped by the 1980s due to social pressure and ethical concerns. However, rarely appearing single cases and small follow-up studies indicate that a small number of patients, predominantly those with mental retardation and for whom no pharmacological treatment was effective, receive such treatment in a mental health setting (See Jiminez-Ponce et al. 2011). This opens the door for using brain stimulation in similar cases – a technique more expensive and inconvenient but also more flexible in its effects. Moreover, brain stimulation offers to be more specific and flexible in the future than generally acting pharmacological neuromodulators. Adjustable and changeable settings and the ability target a specific brain region/circuit make this technique potentially promising biomedical method.

Although the use of pharmacology and brain stimulation for moral modification and modulating pro-sociality is to some extent speculative, biomedicine's potential for emotion modulation has been shown in the treatments of mood disorders. However, using biomedical tools to change our moral and social functioning raises ethical questions and it is the aim of this thesis to examine some of them and to contribute to the current debate on this topic.

The leading question of this thesis is 'What ethical issues should I consider when deciding whether and how to use biomedical means of moral modification?' The common tread throughout this work is the importance of engaging with moral reasons in choosing, guiding the application and assessing the effects of MB. Chapters 1-4 focus on the conceptual issues and the plausibility of MB. Chapters 5-7 address objections and doubts about the ethical desirability of MB. Chapters 7 and 8 focus on the arguments related to the potential impact of MB on freedom and moral agency.

In Chapter 1 I raise the question of possible meanings of the phrase 'moral enhancement' in order to outline the possible applications of moral modification. This discussion also allows distinguishing different levels and aspects of ethical assessment that a MB can be subject to. Moral enhancement' is a potentially ambiguous term. Section 1.2 examines what 'moral' in the phrase 'moral enhancement' means and section 1.3 (especially section 1.3.3.) focuses on drawing an interpretation what ‘enhancement’ amounts in order to show the range of potential uses MB can be put to and discuss what matters ethically with those different uses in mind. Further, section 1.3 briefly discusses the meaning of the term ‘enhancement’ in the context of the treatment and enhancement distinction (1.3.2) and propose the understanding of enhancement as improvement (s. 1.3.3) with an aim of drawing the scope of the current discussion. Understanding enhancement as improvement provides a frame for the scope of the discussion in the current work, and underlies the consideration of both therapeutic and non-therapeutic uses of MB and emotion modulation in this work (see also Chapter 5).

Chapter 2 addresses doubts about the plausibility of MB. It considers whether specifically biomedical means of modifying the moral sphere are likely to be effective, asks what kinds of effects can be expected after MB and thus what goals are in this context reasonable. It also considers what should be taken into consideration when making ethical assessment of costs and benefits (see also Chapter 7). I argue that the goals for MB set out by some of the proponents are too ambitious and should be revisited. However, if we set similar expectations to the expectations we have of cognitive enhancement, meaningful moral or social enhancement is plausible.

Chapter 3 focuses on the discussion of modification of pro-social emotions and attitudes proposed by some to be a target for MB focused on enhancing moral agency. It explores the conflation between the moral and the pro-social present in the literature as well as the discourse in which some, but not all, emotions as constructed as being themselves ‘moral’. The aim of the Chapter is to ask whether or not biomedically modulating pro-social emotions and attitudes such as empathy would likely better moral agency. I argue that empathy and anger might have both pro-social and anti-social consequences, and that even pro-social sentiments are not sufficient for morality. Instead I propose that engagement with moral reasons is necessary for biomedical modification of emotion to result in better moral agency.

Chapter 4 raises the question of the extent to which the presence of moral disagreement affects the application of MB? Section 4.1 explores the limitation of the scope of the argument that MB may be implausible in the presence of moral disagreement. In section 4.3 I examine the implications of fundamental moral disagreement for MB and argue that although moral disagreement may pose a challenge for evaluation of MB applications, there is no reason to favour the *status quo* in the outcome of this deliberation.

Chapter 5 asks whether using specifically biomedical means of moral modification gives rise to a strong ethical reason to forgo using MB. After examining arguments brought forward by critics of medicalization, I argue the process of medicalization is in itself neutral, and only acquires meaning on the basis of what medicalization allows us to do and what costs it brings with it.

Chapter 6 explores the concerns raised in relation to identity. I ask to what extent MB may threaten the narrative identity of the agent and whether such threats can give raise to strong moral reasons to forego the use of MB. After examining attempts at base strong ethical objections to the use of MB on Schechtman's and Ricoeur's accounts of narrative identity, I argue that narrative identity theories face serious problems in providing strong ethical action-guiding reasons.

The last two chapters discuss the impact of MB on freedom and agency. Chapter 7 asks to what extent issues raised in relation to freedom in the discussion of Savulescu and Persson's (2012a) thought experiment called the God Machine call the desirability of MB into doubt. I argue that although the discussion of the God Machine allows for teasing out what we find important in agency (and I suggest that the God Machine would threaten moral agency by affecting the ability of the agent to engage with moral reasons in affected action thus undermining the creation of one's own free will), the conclusions that can be applied from this discussion to real-world MB are limited.

Chapter 8 asks in what way could real-world agent-led MB endanger and facilitate moral agency. I critically examine Harris' (2011) objection that MB would be beyond moral review. Further, I consider MB in the context of Aristotelian framework and limitations of self-control and argue that we should aim for virtue.

CHAPTER 1. What is moral enhancement

1.1. Introduction

This introductory chapter addresses the question of what ‘moral enhancement’ can be interpreted to be as well as outline possible uses MB can be put to. To explore the various ways in which we can understand what ‘moral enhancement’ mean, the first section address three senses in which ‘moral’ can be understood, and suggest that although those three meanings can be overlapping, they can also diverge. The second section addressed the concept of enhancement in the context of the treatment enhancement distinction. Given the arguments against the ethical force and utility of the treatment-enhancement distinction, the further arguments will proceed with understanding ‘enhancement’ to refer to ‘improvement’. Further, I argue that the group of beneficial interventions in the morally-relevant sphere is ethically interesting and needs to be examined further. This is because the prospect of modifying the underpinnings of morality carries with it a possibility of prudentially beneficial interventions (‘enhancements’) that would make agents worse moral agents (moral dis-enhancements). Finally, I argue that moral bioenhancment might not be ‘moral’ in the sense of being all-things-considered morally desirable or ethically permissible and thus warrants careful ethical attention.

1.2. Three senses of ‘moral’

In this section I explore what we mean by ‘moral enhancement’. I suggest some distinctions that might help us to avoid confusion when talking about the matter and propose that ‘moral’ can have three meanings in this context. As such, moral enhancement can be understood as an ethically desirable enhancement of any capacity, an ethically desirable enhancement in the moral sphere, or an enhancing intervention affecting the moral sphere that brings an overall benefit to the subject of enhancement.

1.2.1. Moral Enhancement as Enhancement that is Morally Desirable

When we say ‘moral enhancement’, we could be referring to an enhancement of any kind that is morally desirable. Here we may be thinking about enhancement that will result – other things being equal – in a better world. Vaccinations for smallpox resulted in the eradication of this disease (Eyler, 2003), and most would agree that a world without suffering and deaths brought about by smallpox is better than an otherwise identical world with this disease. We often think that the increase in average life expectancy over the past century is a good thing, and that promoting longevity is, at the very least, an ethically permissible goal of the state, especially if it is accompanied by a good quality of life (Harris, 2007). Some have proposed that cognitive enhancement is not only permissible, but that there may be a duty to enhance (Harris, 2007). Enhancements can therefore be said to be moral in the sense of being morally permissible or even morally obligatory. Thus, 'moral' in the first sense refers solely to such ethical appraisal of a given enhancement.

Although this is not usually the only way in which proponents of moral enhancement (e.g. Douglas, 2008; Persson and Savulescu, 2008) use the concept of 'moral,' it is important to clearly distinguish the concept of the ultimate desirability and ethical permissibility of enhancement from other meanings of the term 'moral.' The conflation of several senses of 'moral' might add to the opacity of the debate on moral enhancement, while such distinction is more obvious when discussing cognitive, mood or body enhancement.

It is important to emphasise the different senses of 'moral,' especially given the note on which the recent debate on moral enhancement started. In his 2008 paper *Moral enhancement*, Douglas argues that enhancement of moral motives might be the kind of moral enhancement that is not susceptible to some of the critiques from opponents of other kinds of enhancement – thus suggesting that enhancement of moral motives is an enhancement that is *prima facie* morally desirable. This and similar positions have been criticized by Harris (2011), who argues that at least the bioenhencement of motives proposed by Persson and Savulescu (2008) would not be morally desirable, if it is possible at all. However, when we say 'moral enhancement' we might mean something very different.

1.2.2. Moral Enhancement as a change in some aspect of morality that results in a morally better agent

Moral enhancement might also refer to making people more moral, thus making them morally better in some sense. This is what Persson and Savulescu (2008; Savulescu and Persson, 2012b) have in mind when they propose that moral enhancement may, theoretically, be an answer to the alleged increased risk posed by the cognitively enhanced and morally corrupt minority.

Being moral is a complex ability and there is a wide range of potentially enhancing interventions. Making morally better people could include making people more likely to act on their moral beliefs (see: Douglas, 2008), improving their reflective and reasoning abilities as applied to moral issues (see: Harris, 2011), increasing their ability to be compassionate (see: Bloom, 2014), and so on. We could also focus on a number of aspects of being moral—acting in a moral way, being more virtuous, or having better moral motivations. The assessment of what would make for morally better agents might somewhat differ depending on the general moral theory one finds convincing (e.g. virtue ethics, deontology, consequentialism), as well as the exact account of moral action and agency.

There are two parts to the idea of moral enhancement understood as making people morally better: the factual claim[2] that the enhancement in question in some way affects the moral sphere, and a normative claim about whether that intervention makes for a morally better person. Those two components may at first seem necessarily coexistent. But the distinction between the factual and the normative claims about moral enhancement is important, as our discussion will be constructed differently depending on whether we take the combined factual-normative claims as the basis for our discussion, or only the factual one.[3]

[2] I do not suggest that those kinds of claims are purely factual, and accept that *prima facie* factual claims might have elements of normativity embedded. However, the proposed crude distinction is sufficient for the current purposes.

[3] I also understand that drawing the scope for what is to be considered moral might be dependent on the moral theory one subscribes to. However, I propose that we draw that scope widely, to accommodate various possible conceptions of the good and morality and argue whether a given intervention is indeed morally enhancing all-things-considered and including one's *exact* account of morality later.

The first reason for this is pragmatic—keeping the distinction in mind will make the discussion clearer. Secondly, considerations of moral enhancement based solely on the factual claim are interesting in their own right, and we would be missing an important part of the ethical enquiry if we focused only on enhancement understood as making people more moral. In later parts of this thesis, for example, I suggest that we have good reasons to ethically consider the effects of SSRIs on our moral capacities.

1.2.3. Moral Enhancement as a beneficial change in the sphere of morality

As mentioned above, the word ‘moral’ in ‘moral enhancement’ can have a descriptive function and refer, for example, to a certain aspect of our cognition. A cognitive approach to moral enhancement would therefore be based on an assessment of cognitive functions and regions implicated in moral reasoning, decision-making, acting and so forth. Further, such an approach would include an investigation of how these functions are affected by, for example, mood, emotion, sleep deprivation, risk assessment, being in hurry, etc. (Blumental, 2005), and how they can be biomedically modulated. On that view, whether an intervention is a moral enhancement depends, first, on whether it affects relevant cognitive processes and behaviours. Secondly, it depends on whether the modification of function counts as an enhancement of the kind we are after.

The question of how narrowly or widely to draw the ‘moral sphere’ should be addressed. Harris (2011, see also subsequent articles) has noted that what makes morality specifically moral is an all-things-considered judgement from a moral perspective. He argues that, as a result, pursuing cognitive enhancement by biomedical or traditional means is the best way of pursuing moral enhancement. Others might talk about the importance of empathy (e.g. Hume 1751;[4] Persson and Savulescu, 2008), and/or emotions such as disgust and elevation (e.g. Haidt, 2012) for morality.

[4] Although Hume (1751) refers to ‘sympathy,’ he refers to a concept we currently label ‘empathy’ (Smith, 2011).

In this section it is not necessary to take a stance on what is sufficient and necessary for morality or moral judgement. Even if we concede to Harris that moral judgement properly so called necessarily involves deliberation from a moral stance, we can still be concerned about the effects of interventions that affect cognitive and affective capacities that feed into and are related to moral agency. For example, indiscriminately decreasing the capacity to react with anger might decrease the disposition for unjustified violent action as well the ability to be enraged at injustice and act on it. Whether or not injustice angers us can still be relevant for our moral agency, even if we accept that what makes the act moral or immoral is not that it results from anger, but rather relates to the appropriate justification based on moral reasons and reasoning.

This approach is consistent with the psychological and neuroscientific view of what psychologists call ‘moral emotions,’ without committing ourselves to meta-ethical positions such as emotivism (Stevenson, 1937; Ayer, 1937) or to intuitionism (Haidt, 2001). The downside of such a wide and loose use of the ‘moral’ is that it risks clouding the debate. But there is also an advantage to drawing the ‘moral sphere’ widely. This wider use of ‘moral’ allows us to look at a range of interventions that affect the underpinnings of moral agency and reflect on whether the intervention is beneficial for moral agency specifically.

The proposed way of using ‘moral’ is perhaps not the most intuitive. We might assume that moral enhancement makes people better moral agents. However, it is a way analogical to the use of the term ‘cognitive enhancement.’ Cognitive enhancement is often understood as a beneficial change to our cognitive capacities (e.g. Harris, 2007). This does not mean that every cognitive enhancer would make for wiser, more knowledgeable or better thinking agents. To a large extent, whether something is a cognitive enhancement is contingent on what kind of skill or cognitive capacity the agent needs enhanced and to what end the enhanced capacity is used. Increasing sustained and focused attention may be a great cognitive enhancement for those who need to focus on similar tasks for hours. But the same drug may ‘inspire’ hours of cleaning, as is sometimes reported by students using Ritalin or Adderall (Vrecko 2013). If the enhanced cognitive capacities are used to acquire false beliefs, we may end up with agents who are no wiser than before. The same intervention can also bring more harm than good if what is needed is flexible, wider attention, as is the case at some stages of a creative process or when reflecting on wider implications of a thought or idea. Thus, where 'cognitive enhancement' is usually used in a sense that implies a possible benefit of a certain kind, it is open to questioning about whether this particular cognitive enhancer is truly resulting in better deliberative capacities, knowledge or prudential benefit. It makes sense to apply an analogical understanding to moral enhancement.

The assessment of whether or not an enhancement is 'beneficial' can be done from the perspective of morality, but may also be based on a prudential evaluation. The next section further explores the interplay of various ways in which an enhancing intervention might be beneficial.

1.3. Enhancement as improvement

1.3.1. Treatment and enhancement

This section aims to explore what can be referred to under the concept of enhancement, and proposes a wide understanding of enhancement as improvement. I will be discussing 'enhancement' in relation to concepts explicated in section 1.2.2 and 1.2.3, that is both enhancement understood as an intervention that aids moral agency and an intervention in the moral sphere that is prudentially beneficial to an agent.

Enhancement is often assumed to refer to the improvement of functioning above normality, while treatments are aimed at maintaining and restoring normal functioning or good health (Juengst 1998). Some scholars argued for the normative significance of the distinction between treatment and enhancement. However, this basis for defining enhancement is problematic. Take, for example, the case of X-linked severe combined immunodeficiency occurring in some boys (Häyry, 2010). In children with this syndrome, the immune system does not provide a defence against infections, so that what would otherwise be a minor infection becomes life-threatening. If there was an intervention that could improve the functioning of the immune system in those boys, we would call it treatment rather than enhancement, despite the fact that the function of the immune systems of those boys is a result of their genetic endowment (Häyry, 2010).

This limitation can be partly addressed if we take ‘normal functioning’ to refer to ‘species-typical functioning.’ This approach was taken by scholars like Sabin and Daniels (1994; Daniels, 1996) who argued that in determining the natural functional organization of members of a species it is possible to create a model of normal or species-typical function. Disease would represent a statistical deviation from normal or typical functioning (Sabin and Daniels, 1994; Daniels, 1996). However, it seems reasonable to assume that disease refers to the state of impaired or indeed less than optimal function rather than simply a deviation from the average – it would be rather awkward to say that to be a genius is to have a disease (Pacholczyk and Harris, 2010).

If disease is a deviation from species-typical functioning, treatment is what restores it. However, that is only correct for those below the typical functioning level. An intervention that levels-down those who are above the range of typical function would be difficult to call an enhancement. Such intervention would be damaging and not beneficial. It would also not be ‘therapeutic’. Therefore, restoration of species-typical functioning can be called therapy only if it constitutes an overall improvement in function or, in other words, an enhancement relative to the state before the intervention (Harris, 2009).

Another problem with the species-typical functioning view is that species-typical traits can be reasonably thought to be disabling (Harris 2001; 2007). That could be the case, for example, when the environment changes in such a way that a given widespread trait becomes a maladaptation heightening the risk of serious harm, which in turn impairs the ability of those possessing this trait to lead full lives. Consider another example. Dying of the diseases of old age is species-typical and normal, but is not necessarily desirable. If we could systematically treat diseases of old age by stimulating the regeneration of tissue and simultaneously switching off the aging processes in the cells, the longevity of patients could substantially increase. This would appear to constitute both therapy and enhancement, and the fact that diseases of old age are species-typical seems not to be overly relevant

(Pacholczyk and Harris, 2010). As a result of the discussed problems of the species-typical view, John Harris has proposed that enhancement may be understood widely as an improvement brought about by a change in a characteristic or function and an intervention that is overall beneficial (Harris, 2007; Pacholczyk and Harris, 2010).

In this work I will not attempt to fully consider arguments related to the treatment/enhancement distinction. The highlighted problems appear to me sufficient to question its normative force and to think about the reasons we want to resort to such a distinction in the ethical assessment of biomedicine. We could have a good reason if the distinction easily translated into moral appraisal of a given intervention, for example if it told us something about its permissibility or helped in decisions about allocation of resources. However, it is unclear that the distinction can serve this purpose. In their discussion of adult ADHD, Schermer and Bolt (2011; see also Schermer, 2007) argued that even if such a distinction could be made for a number of paradigmatic cases, it still leaves us with a large grey area in which such distinction would not be useful. In this work, I will not ground ethical argument in the distinction between treatment and enhancement.

Moreover, I will not attempt to clearly distinguish between the enhancing and therapeutic uses of potential social and moral enhancers. Wolpe points out that our understanding of 'enhancement' and 'treatment' is socially constructed: 'concepts such as disease, normalcy, and health are significantly culturally and historically bound, and thus the result of negotiated values' (Wolpe, 2002, p. 389). What conditions are included under the 'therapy' umbrella is socially negotiated and can be re-negotiated. Some scholars raised doubt about whether the expansion of diagnostic categories such as depression and ADHD is appropriate – perhaps we are labelling as diseases conditions that should not be treated as such (Conrad, 2007). I will address some of the ethical concerns related to medicalization in further chapters. In this introductory chapter it suffices to note that the presence of medicalization and de-medicalization, disease mongering and, expanding disease definitions (Schermer and Bolt, 2011) make the 'enhancing' and 'therapeutic' uses to be moving targets. It is not necessary for this work to hit those moving targets, and the discussion can be enriched by welcoming what Schermer and Bolt (2012) called the 'grey area'.

Thus, I will consider interventions of a similar kind (e.g. aimed at an increase in empathy, modulation of anger, etc.) regardless of whether they would attract a ‘therapy’ label or not. What will be of a greater concern in this work is whether the intervention is a moral enhancement in the sense of making morally better agent (see s. 1.2.2) or an intervention in a moral sphere that is generally desirable (see s. 1.2.3) as well as the factors that can influence the assessment of other factors that influence the assessment of overall moral permissibility of the modification.

1.3.2. Enhancement and benefit

Does the idea of enhancement as improvement include a normative statement about what is morally good? Not necessarily, as enhancement in the sense of a beneficial change in the sphere of morality (s. 1.2.3.) includes only the claim that an intervention is beneficial to the agent. When talking about moral enhancement it may be worth asking whether it is supposed to be beneficial for the person’s moral character or moral behaviour or for their welfare more generally. The ambiguity of the phrase ‘moral enhancement’ can be explained in part by the presence of those two ways in which an intervention can be beneficial—beneficial to the ‘moral character or behaviour’ of the person, or prudentially beneficial.

A similar question may arise in relation to cognitive enhancement—we may be wondering whether an intervention is beneficial for a person’s intellectual capacity or in the person’s interest more generally speaking. There may be cases where cognitive enhancement is in a person’s narrowly construed interests, those related to their cognitive abilities and knowledge, but not necessarily in a more widely construed interest. Let us consider two scenarios involving cognitive enhancement. There may be cases where an improvement in certain cognitive abilities may not be in the overall interest of a person, for example if there are substantial side-effects that affect their physical or mental wellbeing.

Take the example of Esperanza, a girl with borderline learning difficulties. After taking a new smart pill, Esperanza becomes much better at mathematics and physics, but the pill also happens to act directly on her neural circuitry governing mood and emotion—causing Esperanza to feel depressed most of the time.

Contrast this with the case of Ernesto, who has a similar level of learning difficulty but does not experience any obvious side-effects. However, because he is now able to learn so much more effectively, he becomes lonely—his old friends will not play with him because now 'he is too smart'. Despite the efforts of teachers and parents to improve the situation, Ernesto becomes increasingly isolated. In Ernesto's case, although there are no straightforward adverse effects, an improvement in an aspect of cognitive function causes a behavioural change that brings about a net loss in wellbeing.

Consider another example, somewhat akin to the theme of Keyes' (1966) short story *Flowers for Algernon*. Esther has severe learning difficulties. She does not realise that she lacks certain capacities. She is a cheerful person and a pleasure to be around, and she enjoys life. She does not display challenging behaviour and so does not require any medication. A new drug comes onto the market that has been shown to improve cognitive function in people with less severe mental disability. Esther's mother decides to try this drug, and there is a marked improvement in Esther's cognitive abilities. Unfortunately, the improvement is not significant enough to enable her to be more independent: although she understands her environment better, there is no great change in her wellbeing. However, for the first time, she starts to notice the jokes that people make at her expense, and she is acutely aware of her limitations. Although, as Mill (1859) famously wrote, it may be 'better to be Socrates dissatisfied than a fool satisfied,' in some cases the cost of knowledge or reflective awareness may be too high.

As the example of Esther demonstrates, there may be some cases where intervention in cognitive capacities is beneficial in the narrow sense, but not in the wide sense. In this situation we could say that intervention is an enhancement in the narrow sense, but not an enhancement all-things-considered. The cases of Esperanza and Ernesto demonstrate the importance of predicting and estimating costs and benefits. Despite the fact that cognitive enhancement *may* have significantly negative effects on someone's life, like those experienced by

Esperanza and Ernesto, it can be argued that enhancement, in such cases (cognitive enhancement narrowly understood), usually brings more benefits than harms (in the wide sense) and is therefore worth pursuing. Education is important in our societies; high academic performance often translates into better career prospects and brings a number of other benefits. Thus, our experience with a number of instances of enhancement may lead us to say that cognitive enhancement (narrowly understood) is most likely to be in a person’s interest.

Analogically, the multiple ways of assessing whether an intervention in the moral sphere is an enhancement mean that interventions in the moral sphere that are attempting to improve certain function or generally improve moral agency are subject to further ethical assessment.

1.3.3. Moral dis-enhancement

Another issue arises when we consider enhancement as a beneficial intervention in the context of the moral sphere. If a similar logic is applied to that applied to cognitive enhancement, intervention in the sphere of morality is only an enhancement if it is beneficial for the subject of that intervention (see s.1.2.3). Consequently, if we understand moral enhancement analogically to cognitive enhancement, there is nothing *prima facie* inconsistent in saying that moral enhancement may run contrary to what is morally good. Moral enhancement thus understood might, for example, make people less prone to act on moral reasons, give those reasons moral weight or act in a moral way. This is because moral enhancement will refer to an intervention in the sphere of morality that brings about an overall prudential benefit to an agent.

Take Eric, who is deeply moved by moral considerations. He strives for moral integrity and often acts on the basis of his moral beliefs. He spends a substantial amount of his time thinking about what is good and what is right, gives most of his disposable income to charity, and spends many hours per week volunteering. However, his preoccupation with moral obligations has led to problems in his family life. His wife has threatened to leave him if he does not stop bringing homeless people to their house and does not find some time to spend with her. In this case, acting as he thinks he ought to act has negative consequences for Eric's overall wellbeing. Eric decides to strive to care less about others' misfortune.

It is plausible that the ability to modify or biomedically affect the moral sphere will lead to enhancement guided by non-moral considerations. Moral-disenhancement may occur on a relatively small scale and be a matter of individual choice, as in the hypothetical case of Eric, but it might also happen on a larger scale and in the context of institutionally implemented policies. For example, a large number of soldiers suffer post-combat trauma. Post-traumatic stress disorder can have severely disabling effects and make the transition into post-military life challenging. This problem has enjoyed increased attention, yet remains to be addressed (Brewster, 2014; Hattenstone and Allison, 2014). Moreover, the cluster of PTSD symptoms related to hyperarousal was shown to be significantly associated with violent offending (MacManus et al., 2013).

Given the personal and social burden of PTSD on military personnel, it is possible to imagine an intervention that targets emotional reactions to others' distress and harm as a preventative intervention. This kind of intervention is highly speculative, but not implausible. Military training, just like medical training, necessitates the ability to effectively function in the presence of others' distress. However, while we generally see doctors' actions as aiming at alleviating the suffering of others and acting in their interest, at least some military tasks involve the purposeful infliction of harm on others. Even if harm to others is justified (e.g. happens in the context of just war), one can argue that removing 'emotional breaks' can lead to poor moral outcomes and negatively affect the ability of the military personnel to be moral agents. Thus, one can raise the question of whether moral dis-enhancement of agents, even if done in the context of a just cause, is ethically permissible.

This question is not limited to biomedical means. Current military training may involve selectively reducing the disposition to empathise with others (e.g. via dehumanization of the enemy) and alleviating moral distress via reframing (e.g., framing an issue in the form of as a matter of a morally justified fight). Some authors argue that by necessity, soldiers depersonalise both themselves and the enemy to control the emotions that arise while witnessing deaths and killing other human beings (Bartov, 1992; Ben-Ari, 1998; Nadelson 2005). A US soldier stationed in 2007 in Baghdad described this emotional detachment as follows:

> 'If there's one thing about being in a war zone it is this . . . the level and intensity of the carnage that I've seen is unparallel to anything I will ever experience again in my life. . . . But like all things in life, you become desensitized and used to what you see. That is sadly the point in my life where I am. Seeing another dead body, or executed Iraqi or whatever no longer has an effect on me. Nothing . . . cold nothingness. (Eddie, 2007)

Ben-Ari (1998) has argued that although depersonalisation is inevitable in war, it turns into dehumanisation when the enemy comes to be seen as a demon. In such cases excessive and unnecessary violence becomes justified as morally right. According to Bartoy (1992), this happened between US and Japanese, and German and Soviet troops during World War Two. Robben (2012) argued that a similar process took place in the Iraq War where '[t]he hajjis, habibs, ragheads, and sand niggers were the enemy, and they were not thought of with a shred of humanity' (Key, 2007, p. 51). Robben argues that the dehumanising message was already acquired in the earlier training, but subsequently reinforced by racial stereotypes ideologically reinforced by the framing of the conflict as a war against evil and the particular kind of combat that the American presence in Iraq involved. All of this was conducive to unnecessary violence and killing and led, in turn, to lowering the threshold against the mistreatment of civilians and suspects, along with serious violations of military ethics.

Robben argues that this has happened alongside ethical behaviour of soldiers: 'medical care was provided to wounded insurgents, combat missions were complemented by goodwill missions, and empathy was shown for the poor' (2010, p. 146). Robben (2010) states that this does not mean that the moral agency of soldiers was undermined, with the previous quote purporting to demonstrate such unimpaired moral agency. Such assessment, I think, is more complex. I do not attempt to consider all the issues related to moral agency in such situations, but rather point out that biomedical means might be used to achieve a more complete state of ‘detachment,’ complementing the more familiar ways of emotion regulation aimed at detachment. If effective and selective enough, biomedical means could also make it possible to achieve such a state quicker, more easily and perhaps more completely. I think this could make a potentially significant difference in how the combat unravels, and additionally affect the capacities that underpin the ability for noticing and caring about others’ interests.

In such situations some capacity for appropriate emotional reactions underpinned by empathy and reaction to others’ distress, although no guarantee of ethical action, might have some morally desirable effects. This is a reason why the ideology of the Third Reich involved explicit encouragement and even duty to diminish empathetic concern for the ‘enemy’ on top of the more traditional dehumanising techniques. On the other hand, empathy did not prevent the great evils then, and is even less likely to prevent great evils now – where combat becomes even more embedded in technology and where the inflicting of harm is less and less direct.

1.4. Ethical permissibility

The term ‘moral’ in the phrase ‘moral enhancement’ can refer to the overall ethical permissibility or obligation of moral enhancement, and so can be a more general reflection on the context of moral enhancement and its overall consequences. It can be argued that the second and first interpretations of moral enhancement (making people better) and the all-things-considered moral assessment of such intervention may reasonably be collapsed into one. For example, one might argue that moral enhancement is an intervention that will result in people acting more morally, and the moral action will be the one that maximises overall utility. At the same time, if the moral assessment of interventions is also based on maximising utility, the two will likely coincide. We have a similar case if one takes the idea of making people more moral to mean being more virtuous and the ethical value of the enhancing intervention is judged on the basis of the extent to which it promotes virtue. However, there are at least two situations where the two do not have to coincide.

First, moral enhancement in the second sense (creating better moral agents) may be achieved using morally reprehensible means, which in turn will influence our moral assessment of the intervention in question. If the moral enhancement of a given person can only be achieved at the price of the side-effect of inflicting strong pain on another person for a long time, we may reasonably argue that overall it is not worth it, even if enhancing the person will lead them to act in a significantly more moral way following the intervention. I would propose that, in this context, we understand some of Harris' (e.g. 2011) objections to moral enhancement. If moral enhancement was effective (on whatever account) but unjustifiably impaired one's freedom to lead the life one pleases, the moral enhancement might not be morally permissible.

Secondly, making a person more concerned with morality may not have a positive overall effect under certain circumstances. For example, when the subjective and the objective right[5] do not coincide, more consistently following what is subjectively right at least sometimes leads to morally worse outcomes. A further scenario is possible—namely when an enhancing intervention will result in a less moral individual, but will have overall good effects.

1.5. Conclusions

Some may object to the use of the term 'moral enhancement' as referring only to certain capacities and not carrying a clear normative message as well. Although it may be true that when we think about moral enhancement we automatically think about people being better morally speaking, this approach introduces hard to avoid confusion due to the 'moral' doing double, and sometimes triple, work; as a description of the target abilities that are improved, as a normatively loaded reference to whether that intervention results in people being morally better in some way, and as a reference to whether this enhancement is overall desirable from a moral point of view. This may introduce confusion when examining moral enhancement—we can be asking three questions to which we might give, at least in principle, three different answers.

[5] For subjective rightness, see Carritt (1947) and Parfit (1988).

Moreover, if we find cognitive enhancement interesting, we are also likely to find moral enhancement in the third sense (an enhancement of the moral sphere beneficial to the agent) interesting. It raises interesting questions about authenticity, free will and the moral nature of humans, the role of rational choice and of the motivation to be moral in choosing which moral stances one adopts. As a result, although most of the current discussion focuses on the second understanding of moral enhancement, the 'cognitive' understanding is interesting in its own right.

Although I recognize the need to address the mentioned questions, a more detailed discussion will have to wait for another occasion. The aim of this thesis is to evaluate and contribute to the current debate.

In sections 1.3 and 1.4 I have attempted to elucidate certain complications in the ethical assessment of modifications in the moral sphere. For example, an attempt at moral enhancement understood as making morally better agents will be subject to assessment whether or not it is also aids moral agency in the wide sense and whether or not in enhancement actually making a person better off. Consequently, I will use the term 'moral modification' or 'attempts at moral enhancement' to acknowledge the further need for moral assessment. Although most commentators use 'MB' to refer to 'moral bioenhancement,' I will take 'MB' to be a short for 'moral biomodification.'

In the next chapters I mainly focus on two questions: (1) whether or not and when biomedical moral modification will enhance or endanger moral agency, and (2) whether or not and which biomedical attempts at moral enhancement are morally desirable. The next chapter examines the plausibility of biomedical modification of emotion and pro-social sentiment as a way of creating better moral agents.

CHAPTER 2. Hype or hope? Plausibility without techno-utopianism

2.1. Introduction

Nordmann (2007) criticised the ethical discourse surrounding nano-, bio- and neurotechnologies for what he calls an ‘if and then syndrome’ (p. 32). He argues that the discourse validates the incredible and improbable future and only then critiques or endorses such imaginary technology. Nordmann (2007) criticises this discourse for squandering the scarce and valuable resource of ethical concern, misleading by casting remote possibilities or philosophical in principle thought experiments as foresight about likely technical developments, in effect deflecting ethical consideration from present transformative technologies. In this chapter, I will examine the plausibility of moral modification aimed at creating better moral agents in the near future. I will argue that in order to engage in an ethical debate more in touch with the present and near and possible future, we need to reconsider both the goals and expectations implicitly and explicitly put on MB by Persson and Savulescu (2008; Savulescu and Persson, 2012b). In the first sections of this chapter, I argue that the prospect of effective MB is indeed implausible is we accept the goals of eliminating large-scale harm proposed by Persson and Savulescu (2008), given the context dependency and effectiveness of pharmacological interventions. The latter part of this chapter examines oxytocin as a trust- an empathy- promoting drug to see what we *can* expect of MB.

2.2 Reconsidering our goals

In *The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity,* Persson and Savulescu (2008) argue that non-traditional means of enhancement may contribute to the rising risk of considerable harm to a large number of people and, therefore, are ethically problematic. They suggest, however, that this threat may in theory be offset by moral enhancement, by which they understand moral modification aimed at making moral agents that do less evil. Accordingly, in their paper MB is seen as a potential tool for eliminating this alleged risk of large-scale harm. Since small groups or even individuals may inflict serious harm, the aim of MB is to prevent the 'morally corrupt minority' (p. 162) from doing so.

Let us consider the notion of increasing empathy to illustrate some difficulties with this approach to MB. Increasing empathy comes to mind when we think about what kinds of intervention might carry out the task that Persson and Savulescu (2008) want MB to do. Lack of empathy is sometimes said to be correlated with criminal behaviour and it seems sensible to assume that increasing one's appreciation (be it cognitive, affective or both) of others' suffering will decrease the likelihood of behaviour that is likely to result in harm. For the purpose of this argument let us assume that there is an intervention that substantially increases empathy and is shown to be effective regardless of the measure used to assess the magnitude of that effect. But increasing general empathy will most likely not be enough, for several reasons— even when we assume that increased empathy is going to make a substantial difference to our motivations to act in a certain way (see: Persson and Savulescu, 2008; Fenton, 2010; Harris, 2011).

First, we know that moral reasons are not the only basis for action and that prudential reasons can override moral ones. Thus, even if increased empathy indeed gives rise to reasons not to harm others that are stronger than before, they may not be strong enough to cause us to refrain from performing a fatally harmful action. As a result, it is reasonable to expect that even a highly efficient intervention will not be sufficient to abolish the possibility of harm completely.

Second, there may be cases where an increase in empathy will increase the risk of large-scale harm. It is not clear that the allegedly morally corrupt minority that may pose a threat acts solely on the basis of non-moral reasons. This claim seems to be based on a conflation of two uses of 'moral'—one to describe a *kind* of reason for action and ethical assessment of actions. Thus, when we refer to the 'morally corrupt minority' we might mean 'those whose acts we judge as immoral'. However, it is possible for a terrorist's actions to at least appear to be based on moral reasons, that is, reasons of a moral *kind*. There is a long tradition of those who claim to be fighting for what they consider to be a better world regarding the infliction of harm as a necessary evil; sometimes we may support this struggle and sometimes we may denounce it. We may disagree with the moral assessment that the terrorist has made, rejecting some or all of her reasons for action, we may disagree about which ends are desirable, or we may simply disagree on our predictions of likely consequences—chances of success and the cost of bringing about the desired end. On the other hand, there is also a long tradition of arguing against change despite the harms that result from leaving things as they are.[6]

There are other reasons why MB may be unsuitable to serve the purpose Persson and Savulescu (2008) want it to serve. They emphasise wickedness as a cause of large-scale harm but, as Harris (2011) points out in his response in 'Moral Enhancement and Freedom', large-scale harm can be inflicted not only by 'the bad' but also by 'the mad'. Moreover, it can result from incompetence, stupidity, negligence or miscalculation (Harris, 2011, Fenton, 2010). Thus, MB, even if possible and effective, is likely to be unable to offset the dangers allegedly brought about by cognitive enhancement and the development of science in general. If this is the only goal of MB, than we have little chance of achieving it.

2.3. Reconsidering our expectations

2.3.1. Context-dependency and individual differences

I have outlined some of the reasons why MB may indeed be incapable of eliminating the risk of large-scale harm, as Persson and Savulescu (2008) seem to require. But the expectation that MB *eliminates* the risk of large-scale harm seems to be not only potentially impossible but also unreasonably demanding. One of the reasons why it is unreasonably demanding was pointed out by Harris (2011): the expectation that MB *ensures* safety by *eliminating* risk seems to be impossible to satisfy. Even if MB were fantastically efficient and cost-effective, it would be unlikely to eliminate the risk of large-scale disaster because one malevolent person, who slipped through the net of enhancement or for whom the intervention did not work, is enough for the risk not to be eliminated. Even assuming that all disasters are caused only by malevolent individuals, the standard for the effectiveness of moral enhancing interventions is set very high. Although we may have such hope,

[6] As pointed out by John Harris (1980).

we do not normally expect cognitively enhancing technologies to work for every single person, nor do we expect most very effective treatments to work in every single case.

Both cognitive and moral enhancement can be achieved by a variety of means. One of the possibilities is to use pharmacology. Pharmacological interventions often vary in effectiveness from individual to individual, and the influence of individual differences on outcomes is very well known, especially when the goal of intervention is to modify behaviour, mood or thinking processes. Predictions of outcomes (and side-effects) for a single patient can be so unreliable that suitable medication is decided upon only after a process of trial and error (Huskamp, 2003). Often, a group of patients will be unresponsive to all of the pharmacological remedies, and sometimes to both different types of medication and different types of therapies, as well as to combined approaches. The use of pharmacology that seems to be sensitive to individual differences and pharmacological interventions, at least as our experience so far suggests, is likely to work for some but not for others. The number of subjects who might experience a desired effect is likely to increase with the growing variety of available interventions, as new drugs and other technologies (such as deep brain stimulation) are designed and tested to address the needs of those for whom nothing has yet worked (Mayberg et al. 2005; Berton and Nestler, 2006).

It is important to admit that pharmacological interventions have their limitations, but it is equally important not to forget about the cases where those interventions are effective. It may be regrettable that a drug is not effective for *all* (or even many), but denying the plausibility of moral enhancement because it does not work for all seems to be unjustifiably demanding.

Moreover, whether a potentially enhancing intervention is indeed enhancement depends both on the context and on a particular person’s needs. While weight gain may be an enhancement for an underweight individual, it would not be so for an obese person. Although we may think of a change as a typical cognitive enhancement, for example an increase in the ability to focus on a particular task while ignoring distractions (focusing attention), the same intervention may be neutral or even counterproductive in some tasks that require creativity.

Whether a particular intervention is going to have morally enhancing effects[7] depends on the context, individual and individual's activities. However, this only means that Savulescu and Persson's goal will unlikely to be furthered by moral bioenhancement. Context-dependency abolishes neither the usefulness and plausibility of cognitive bioenhancement nor the usefulness or plausibility of moral bioenhancement.

2.3.2. Effectiveness

Another issue is the expected magnitude of change. The plausibility of MB for the purpose of making better moral agents may be put in doubt if the interventions available seem to be insufficiently effective. If we were to adopt the goal of MB that Persson and Savulescu adopt (2008), we would expect morally enhancing intervention to result in the overriding of any inclination or reason for inflicting large-scale harm. Among 'the wicked' who are willing to inflict large-scale harm, there are likely to be both those whose vector sum of reasons largely points towards causing a disaster, and others riddled with doubt but who in the end decide to follow through. There will be opportunists who will change their minds with a little nudge, and those for whom causing large-scale harm is a purpose in life and is consistent with at least some of their strongly held beliefs. The impact of morally enhancing interventions would indeed have to be great to override all the subjective reasons for inflicting large-scale harm, strong attitudes and the impact of deeply engrained beliefs. With the bar of expectations regarding effectiveness set so high, other commentators (e.g. Harris, 2011) are right to doubt that sufficiently effective MB will be possible. However, there is no reason why we should understand 'sufficient' as Persson and Savulescu (2008) do for the purpose of their paper.

It is important to note another potential reservation. One could argue that since no MB likely to be possible in the near future will be capable of making *any* of the

[7] In the sense of being conducive to the good and affecting the moral sphere

wicked people good, MB is implausible. This objection seems to be especially illustrative because it is misguided; it remains so *even if* we agree with the factual premise. First, it misunderstands enhancement as necessarily bringing people from one extreme of the spectrum to another, and, secondly, it grossly oversimplifies the issue. We do not necessarily expect cognitive enhancers to turn stupid people into smart ones; rather we expect that they will improve certain aspects of cognition, improve ability to deal with certain kinds of tasks, and so on. Although we may think about prototypically 'smart' and prototypically 'stupid' people, 'being smart' can mean many different things, and requires a whole range of cognitive processes. Secondly, 'smart drugs' do not make people smart or even smarter. They modify a narrow aspect of functioning that partly underlies abilities and behaviours that we then see as signs of being smart. Whether any particular case of modifying an aspect of cognitive functioning amounts to an enhancement may very well be context-specific. If cognitive enhancement does not make stupid people smart, why should we expect MB to turn wicked people into virtuous ones? I suggest that we should not.

This is not to say that we could not *want* cognitive enhancers and moral enhancers to have this magnificent effect. We may hope for, imagine, and talk about the possibilities of radical human enhancement of a cognitive and moral realm, but even if radical enhancement is impossible, there is no reason why we should come to the conclusion that any enhancement is implausible. Naturally, we could point out that the effectiveness of enhancers is disappointing. It may indeed be the case that some enhancing interventions will have such a small influence on functioning that, for most of us, they are not worth the hassle. However, as some have pointed out, the cumulative and long-term impact of small changes can be substantial, yet are easily underestimated or overlooked altogether (Turner and Sahakian, 2006).

Another issue is that of comparative effectiveness as well as comparative cost-effectiveness. It may be the case that any pharmacologically induced change in moral functioning will be much less effective than more traditional means such as moral education. Even if that is the case, it may still be worth pursuing. It may be worth pursuing if pharmacological methods will be significantly cheaper or more accessible, and thus ultimately cost-effective. Moreover, it may be the case that application of pharmacology and other novel means of enhancement may be more effective or cost-effective in certain specific circumstances, for specific groups or as a method that complements traditional means.

If we are going to consider pharmacological interventions that affect morality or cognition, there is no reason to be *automatically* discouraged if they have limited effectiveness (ie, they do not turn maths idiots into maths geniuses or morally corrupt people into paragons of virtue). We would likely find an analogical threshold for efficiency impossible to reach for many pharmacological interventions. What we should rather look at is whether the effect of a single intervention or/and its repeated use is great enough, including the cumulative benefits of small beneficial changes to the extent possible, and look at comparative cost-effectiveness. It seems to be grossly premature to make strong judgements about those issues, given that neuroscientists only turned their interest to the cognitive science of morality a short time ago and that empirical research into the effects of different interventions in the moral sphere of individuals is far from extensive. If we align our expectations to those we have of cognitive enhancement and treatments for mental health disorders, the prospects of finding pharmacological or other moral enhancers seem to be much better.

2.4. A case study: oxytocin

Increasing empathy and trust has been proposed as an example of moral bioenhancement (Persson and Savulescu, 2008; Savulescu and Persson, 2012b). Oxytocin, a hormone and neurotransmitter originally known for its role in childbirth and lactation, has recently been shown to affect social behaviour and has been proposed as a potential agent of modifying moral and social behaviour. It is worth to take a closer look at the effects of oxytocin as it demonstrates both the potential effectiveness of psychopharmacological interventions in modifying social and morally-relevant behaviour and the complexity of its effects.

Studies in mice suggest that low levels of oxytocin correlate with impaired ability to recognise (Ferguson et al. 2000; Ferguson et al. 2001) and bond to one’s peers (Winslow and Insel, 2002). These observations came in part from experiments with mice with a mutated oxytocin gene. Ferguson et al. (2001) showed that mouse knock-outs show a profound social recognition deficit despite normal olfactory and spatial learning abilities and that the social recognition ability can be fully restored by an injection of oxytocin in the medial amygdala. Similarly, a high level of this hormone seems to correlate with caring behaviour in rodents (Pedersen et al. 1982).

In humans, oxytocin has also been shown to influence social behaviour and cognition. It also plays an important role in creating the mother-infant bond. Feldman and colleagues (2007) showed that a mother's level of oxytocin in the first trimester predicts the strength of the mother's attachment to the infant. In addition, a boost to oxytocin levels in experimental settings commonly achieved by administration of a nasal spray, seems to promote trust (Kosfeld et al. 2005; Baumgartner et al. 2008) and generous behaviours (Barraza and Zak, 2009; Zak et al. 2007). Oxytocin seems to influence social cognition (Theoridou et al. 2009), increase some aspects of memory for social stimuli (Unkelbach et al. 2008; Guastella et al. 2008) and to increase 'mind-reading' ability (Domes et al. 2007).

Substantial effect sizes obtained in experiments on trust and generosity (for example, participants were 80 per cent more generous in the oxytocin group than in the placebo group in the Zak et al. (2007), and 30 per cent more trusting in the experiment done by Baumgartner and colleagues (Baumgartner et al. 2008), including some laboratory experiments that may have relatively high ecological validity (in Ditzen et al. [2008], couples were asked to argue, with researchers measuring the frequency of positive behaviour such as listening, confirming or laughing during the conflict), show that the use of oxytocin in everyday life is plausible. Moreover, oxytocin can be delivered in a practical way. Injections are not required: several of the mentioned studies used a nasal spray to deliver oxytocin. Several websites already market a nasal spray containing oxytocin,[8] although there is a need to evaluate the quality and the amount of the active ingredient in the commercial versions of the product to ensure that their effectiveness is similar to that demonstrated in empirical studies. The marketing claims on those websites might need to be taken with a pinch of salt at the very least. On the other hand, the

[8] e.g. http://oxytocinnasalspray.org; www.oxytocinstore.com.

justified scepticism about the oversimplifying claims used to market the product need not lead us to conclude that oxytocin has no effect on behaviour, especially given that the current evidence suggests a behavioural effect.

Given decent effect sizes in experiments, we may worry that in many circumstances oxytocin might impede judgement and increase trust when trusting is unwarranted or even harmful (Damasio, 2005). This view could be supported by evidence that oxytocin seems to restore trust following betrayals (Baumgartner et al. 2008). These worries seem further justified given that although trust is an important social resource (Giddens, 1991), it can sometimes also be socially maladaptive (Greenspan et al. 2001). Some have even proposed that commercial and military applications could harness the potential of oxytocin to make people gullible (Dethlefs, 2007). But is it indeed the case that an increase in oxytocin leads people to trust others indiscriminately?

Mikolajczak et al. (2010b) suggest that the matter is somewhat more complex. They point out that in previous experiments, participants rarely met the same partner twice, nor did they have any idea that the person they interacted with was unreliable. Moreover, previous research suggested that the effects of oxytocin, for example on aggressive behaviour (this is especially well-illustrated in research on aggression in female rodents), are context dependent (Campbell, 2008; Pedersen, 2004). Mikolajczak et al.'s (2010b) doubts were confirmed in research that used a customised economic trust game incorporating repeated interaction, with some partners seemingly being more trustworthy than others. Consistently with previous studies, they found that participants who received oxytocin transferred more money to partners in comparison with participants in the control group. However, participants transferred more money to partners perceived as reliable but did not transfer more money to seemingly unreliable ones. This suggests that oxytocin administration does not increase trust when the partner appears unreliable. On the basis of these findings, Mikolajczak et al (2010b) propose that the effects of oxytocin may be moderated by the perception of risk. The effects of oxytocin on trust have to be confirmed for other contexts, although one recent study indicates that oxytocin also increases trust in circumstances that do not involve monetary transfers—participants who received oxytocin were 44 times more trusting that their privacy would not be violated than participants in the control condition

(Mikolajczak et al. 2010a; Mikolajczak et al. 2010b). It is likely that oxytocin does nevertheless have some effect on our perception of how trustworthy others are. Yet let us make an assumption (reasonable on the basis of current evidence) that, generally speaking, the administration of oxytocin promotes trust—but it is unlikely that we would make some disastrous decisions because of this intervention. The research on the trust-promoting effects of oxytocin has an especially wide potential application given findings showing that we tend to underestimate people's trustworthiness (Fetchenhauer and Dunning, 2010), and given the importance of trust for morally relevant actions.

Oxytocin also seems to improve empathy, but the effect seen in research was only prominent for less socially proficient participants—as measured by the Autism Spectrum Quotient (AQ)—while no similar effect was found in the more socially proficient group (Bartz et al. 2010). Domes et al. (2007) found that oxytocin improves performance only for difficult stimuli. These findings go against the tempting but simplistic view that oxytocin can be used as a universal prosocial enhancer that can turn all people into social-cognitive experts. Instead, perhaps unsurprisingly given our knowledge about the context-dependency and impact of individual differences on the effects of both synthetic and naturally occurring pharmacological agents, oxytocin appears to help only some people. That is not to say that the effects are not substantial. In an experiment conducted by Bartz et al. (2007), the administration of oxytocin equalised the differences in performance of low and high-performance groups in such a way that the performance of participants with high and low AQ scores was indistinguishable.

The exact mechanism underpinning the influence of oxytocin on trust, empathy and other potentially morally relevant abilities is still debated, but only looking at the known effects paints an intriguing picture. Assuming that the results discussed so far are confirmed, we may be able to gain a new, and possibly more convenient, means of influencing our level of trust and empathic ability, probably without a worry of overdoing it drastically (although there is some indication that oxytocin seems also to increase envy [see Shamay-Tsoory et al. 2009]). This may be not good enough for those envisaging radical human enhancement, but, as was argued above in the section on 'reconsidering our expectations', the potential for this substantial, although not universal or radical, enhancement may be worth pursuing.

Although increases in trust or empathy may not eliminate (or maybe even decrease or keep constant) risks of large scale harm, an increase in empathy for those at the lower end of the functioning spectrum has the potential, via possible improvements in social cognition, to contribute to a greater ability to take into consideration others' wellbeing— an ability fundamental to moral consideration.

Positive behavioural effects demonstrated by couples during conflict in Ditzen et al. (2008) (which are most probably also mediated by an influence on the amygdala and stress levels) seem to point us towards the potential for a marked behavioural effect. These can be especially useful in professions that require empathy, a well-developed ability to notice others' distress, and maintaining prosocial attitudes under stress and during conflict, such as in the caring professions. The ability to increase generosity and trust in social exchanges where the exchange partner is judged to be trustworthy (or at least not particularly untrustworthy) seems highly unlikely to solve the serious political conflicts exacerbated by lack of trust. However, we can easily imagine a situation where an increase in the frequency of acts typical of a Good or even a Splendid Samaritan—doing what is morally desirable but not obligatory—can have a notable and positive influence on what kind of social world we live in. I will further explore the reasons for and against facilitating moral agency of Good Samaritans in chapter 8.

On the other hand, we are not justified in seeing oxytocin as an unproblematic social or moral enhancer. Firstly, whether or not the pro-social effects of oxytocin are truly pro-social will depend on the context. The pro-social effects might also be tied to undesirable effects such as increasing envy. Moreover, as Harris (2011) pointed out, the increase in pro-social behaviour does not necessarily imply that such effects will lead to truly better moral agency. Where there is no appropriate moral judgement involved, even where agents do more good their good acts need not be motivated morally. Moreover, where such moral guidance by reason is absent, we can expect that the pro-social effects might be sometimes inimical to morally good outcomes. If oxytocin increases in-group trust and co-operation, it might be pro-social in a narrow sense but lead to evils if the group aims, for example, at attacking another group. Guidance by reason remains a necessary component of moral decision-making and moral action, and no nasal shot of oxytocin is going to assure the ethical behaviour of the recipient of such intervention. The difference between the moral and the prosocial in relation to the importance of reflection in changing pro-social inclinations is further explored in Chapter 3. The importance of deliberation undertaken from a moral standpoint in the process of choosing moral modification, as well as in the process of evaluating outcomes of the procedure, will be further discussed in Chapter 7 and 8.

2.5. Conclusions

I have addressed doubts as to the plausibility of MB based on claims of its low effectiveness and context-dependency. Savulescu and Persson in their 2012 book *Unfit for the future: the need for moral enhancement* argue that we desperately need MB. Whether or not we are 'fit for the future,' biomedical means are unfit to the purpose Persson and Savulescu (2008) envisaged. However, I have argued that we should treat MB in a way analogical to cognitive enhancement. If we set more modest goals and revisit our expectations about the effects of MB, it is premature to conclude that such enhancement is implausible. However, modifications of our moral sphere will still be subject to the questions about the balance of costs and benefits, the weight of prudential and moral considerations and the ethical permissibility of the ways in which they are pursued.

CHAPTER 3. The moral and the pro-social

3.1. Introduction

The idea of direct emotional modulation for moral enhancement (understood as making better moral agents) has recently come under criticism (Harris, 2011; Jotterand, 2011). Some have criticized the Persson and Savulescu (2008) approach as not saying much about moral enhancement at all. Firstly, it has been argued that the MB interventions proposed to make morally better agents (such as increases in trust or empathy) are more accurately characterized as enhancement of pro-sociality. However, since the moral and the pro-social are not the same, reducing morality to pro-sociality does not take the crucial part of our moral psychology into account.

In this chapter, I ask whether or not biomedically modulating pro-social emotions and attitudes such as empathy would likely better moral agency. I will examine the difference between the moral and the prosocial and outline and examine MB aimed at increasing empathetic ability. Further, I will argue that Harris' and Jotterand's arguments about the importance of rational deliberation from a moral perspective are convincing.

3.2. Empathy's charm

Proposals that an increase in empathy would lead to morally better action (e.g. Persson and Savulescu, 2008; Savulescu and Persson, 2012b) are partly based on a recent line of psychological and neuroscientific research that investigated the role of empathy in explaining immoral behaviour in relation to personality disorders and Autism Spectrum disorders. Baron-Cohen (2011), in his *The Science of Evil,* argues that describing cruel acts as evil explains nothing. Instead, he explores the hypothesis that such acts can be traced to a distinct psychological state – a lack of empathy.

Appeals for more empathy inspired (although not always informed) by social neuroscience abound. In the introduction to his book, *The Empathic Civilization: The Race to Global Consciousness in a World in Crisis*, Rifkin writes that:

> ‘A radical new view of human nature is emerging in the biological and cognitive sciences and creating controversy in intellectual circles, the business community, and government. Recent discoveries in brain science and child development are forcing us to rethink the long-held belief that human beings are, by nature, aggressive, materialistic, utilitarian, and self-interested. The dawning realization that we are a fundamentally empathic species has profound and far-reaching consequences for society.’ (2010a, p.1)

Rifkin predicts that ‘we are at a decisive moment in the human journey where the race to global empathic consciousness is running up against global entropic collapse.’ (2010a, p. 42) To support this view, Rifkin describes the history of several civilizations and then explains the entropic biological, technological and environmental changes that threaten the continuation of civilization and may undermine human nature and empathic capacities. Rifkin sees sympathy as passive while empathy ‘conjures up active engagement – the willingness of an observer to become part of another's experience, to share the feeling of that experience’ (2010a, p. 12).

Rifkin does not stop at pointing out the importance of such conceived empathy for the enjoyment and sharing of the experiences of others. He argues that world leaders act on the basis of faulty assumptions about human nature. According to Rifkin, the mistaken assumptions were lied down in the Enlightenment, at the dawn of the modern market economy and the emergence of the nation state: that human beings’ essential nature is rational, detached, autonomous, acquisitive and utilitarian and argued that individual salvation lies in unlimited material progress here on Earth. Social neuroscience, and the discovery of mirror-neurons in particular, claims Rifkin, has forced us to re-evaluate this outdated view of human nature. Given the economic and social problems we are facing, Rifkin says that what is ‘required now is nothing less than a leap to global empathic consciousness and in less than a generation.’ (2010b, p. 2)

Rifkin both envisages and describes empathy extended to all living things and the biosphere in general. He quotes Maslow who said that ‘[m]ore sensitive observers are able to incorporate more of the world into the self, i.e., they are able to identify and empathize with wider and more inclusive circles of living and nonliving things' (1969, p. 42) and contrasts this view with the Enlightenment ideas of human nature, which, according to him did not give emotions their deserved place and importance. While Rifkin sees the internet and other new communication technologies as the technology that will aid the 'radical leap,' Savulescu and Persson (2012b) add to the discussion by proposing MB as a way of achieving moral outcomes.

There are many problems with Rifkin's account. In the remaining parts of this chapter I will address two. One problem has to do with the extreme view of the utility (I dare to use this apparently *passe* Enlightenment concept) and sufficiency of empathy in the making of a morally better world and agents. Rifkin (2010a), Baron-Cohen (2011) and Savulescu and Persson (2012b) seem to advocate an increase in empathy as a *panacea,* but such an indiscriminate increase in empathetic sensitivity would hardly be conducive to moral outcomes. Focusing on the notion of empathetic distress, I will argue that while empathy might be conducive to moral outcomes, Baron-Cohen's equivocation of 'more empathy' with 'morally better' is misguided. The second problem has to do with a more general equivocation of the pro-social with moral. I will argue that Jotterand, Harris and others, are correct in pointing out the importance of moral reasons in the quest for a better world and better moral agents.

3.3. The limits of empathy

The first problem comes when we try to turn empathetic ability into action. Empathy makes one more aware of other people's suffering, but it is not clear that it actually strongly motivates one to take moral action, or prevents a person from taking immoral action. In the early days of the Holocaust, Nazi prison guards sometimes wept as they killed Jewish women and children, but they still did it. Subjects in the famous Milgram experiments felt considerable anguish as they appeared to administer electric shocks to other research participants, but they administered the shock anyway (Milgram, 1963). Even where empathy orients one towards the interest of others and thus to moral action, it is not sufficient if that action comes at a personal cost or where other strong considerations are present.

However, an even more serious problem concerns the potential of empathy to result in bad outcomes. This can occur, for example, if one is overwhelmed with others' suffering to the point where the ability for effective action is impaired. Baron-Cohen (2011), in his book *The Science of Evil,* draws upon experimental and clinical psychology and social neuroscience to argue that the notion of evil should be replaced with the concept of 'empathy erosion.' After proposing that we describe the empathetic ability within a six-level framework – 'spanning from no empathy at all to being continually focused on other people's feelings . . . . in a constant state of hyperarousal, such that other people are never off their radar,' Baron-Cohen argues that a high degree of empathy is conducive to making people good and creating good societies.

He further illustrates this point as follows:

> Hannah is a psychotherapist who has a natural gift for tuning in to how others are feeling. As soon as you walk into her living room, she is already reading your face, your gait, your posture. The first thing she asks you is 'How are you?' but this is no perfunctory platitude. Her intonation—even before you have taken off your coat—suggests an invitation to confide, to disclose, to share. Even if you just answer with a short phrase, your tone of voice reveals to her your inner emotional state, and she quickly follows up your answer with 'You sound a bit sad. What's happened to upset you?' (…) She has an unstoppable drive to empathize. (2011, p. 27)

In response, Bloom (2014) points out that that although Hannah might be a good psychotherapist or a parent of young children (see: also Feshbach, 1990; Rosenstein, 1995; Moses, 2012), it is far from clear that this ability is good for others or for her – if, indeed, 'the drive to empathize' is unstoppable or strong. He argues that Hannah's experience might be the opposite of selfishness but is just as extreme: 'A selfish person might go through life indifferent to the pleasure and pain of others—ninety-nine for him and one for everyone else—while in Hannah's case, the feelings of others are always in her head—ninety-nine for everyone else and one for her' (Bloom, 2014). Bloom argues that some research (e.g. Batson and Weeks, 1996; Batson et al.

1988) suggests that the higher rates of depression and anxiety in women are in part explained by a sex difference in the propensity for concern with others, which results in putting others' needs before one's own. Moreover, although the participants in Batson's experiments on help and empathy seem to genuinely care about whether their help actually addresses the other's need, the negative effects on their well-being might be exacerbated by the fact of feeling bad if their efforts were not helpful – even through no fault of their own (Batson and Weeks, 1996; Batson et al., 1988).

Caring and giving can be stressful, difficult, and draining and concern for others can sometimes overtake people's efforts at self-care. Professionals who work in human service occupations can suffer from mental and physical health problems that have been associated with the strain of giving as a full-time occupation (Figley, 1995). These problems are common in medical professionals, psychologists, social workers, lawyers, and corrections professionals. Consistent with these notions, '*compassion fatigue*' is defined as the experience of 'stress resulting from helping or wanting to help a traumatized or suffering person' (Figley, 1995, p. 7). Moreover, Klimecki and Singer (2011) argued that high empathy specifically can lead to professional burnout. They argued that 'burnout in caregivers and empathic [or personal] distress are characterized by the experience of negative emotions, which lead to a self-oriented response with the desire to alleviate one's own distress; both have negative effects on health' (Klimecki and Singer, 2011, p. 285). Personal distress involves feelings of being worried, perturbed, or upset, *for oneself*, and can lead to behaviours oriented on alleviating one's own distress, rather than distress of others.

Stanislaw Lem, a science-fiction author did not need the recent neuroscientific and psychological evidence to predict the possible effects of the Rifkin – Baron-Cohen – Savulescu and Persson proposal. In his suitably humorous account, he takes the 'more empathy' proposal to its logical conclusion. In *Altruizine; or, a True Account of How Bonhomius the Hermitic Hermit Tried to Bring About Universal Happiness and What Came of It*, a short story in Lem's *The Cyberiad* (1974), altruizine is a drug that duplicates in others who are nearby whatever sensations, emotions and mental states one may experience, supposedly leading to 'Brotherhood, Cooperation and Compassion in any society… since … [s]hould [one] suffer any hurt, they will rush to help at once' (p. 272). Allow me to quote at length Lem's description of the effects of the brilliant invention:

> 'At breakfast time, wandering the streets in a daze, I came upon a tearful multitude that chased an old woman in a black veil, hurling stones after her. It so happened that this was the widow of one much-esteemed cobbler, who had passed away the day before and was to be buried that morning: the poor woman's inconsolable grief had so exasperated her neighbors, and the neighbors' neighbors, that, quite unable to comfort her in any way, they were driving her from the town. This woeful sight lay heavy on my heart and again I returned to my hotel, only to find it now in flames. It seems the cook had burnt her finger in the soup, whereupon her pain caused a certain captain, who was at that very moment cleaning his blunderbuss on the top floor, to pull the trigger, inadvertently slaying his wife and four children on the spot. Everyone remaining in the hotel now shared the captain's despair; one compassionate individual, wishing to put an end to the general suffering, doused everyone he could find with kerosene and set them all on fire. They were discussing it, the scoundrels: apparently the newlyweds' performance had fallen short of their expectations.
>
> Meanwhile each of these former vicarious grooms carried a club and drove off any sufferer who dared to cross his path. I felt I should die from sorrow and shame, yet still sought a man—but one would do—who might a little lessen my remorse. Questioning various persons on the street, I at last obtained the address of a prominent philosopher, a true champion of brotherhood and universal tolerance, and eagerly proceeded to that place, confident I should find his dwelling surrounded by great numbers of the populace. But alas! Only a few cats purred softly at the door, basking in the aura of good will the wise man did so abundantly exude—several dogs, however, sat at a distance and waited for them, salivating. …

> As I stood there, two men approached. One looked me straight in the eye as he swung and smote the other full force in the nose. I stared in amazement, neither grabbing my own nose nor shouting with pain, since, as a robot, I could not feel the blow, and that proved my undoing, for these were secret police and they had employed this ruse precisely to unmask me. Handcuffed and hauled off to jail, I confessed everything, trusting that they would take into consideration my good intentions, though half the city now lay in ashes. But first they pinched me cautiously with pincers, and then, fully satisfied it produced no ill effects whatever on themselves, jumped upon me and began most savagely to batter and break every plate and filament in my weary frame. Ah, the torments I endured, and all because I wished to make them happy! At long last, what remained of me was stuffed down a cannon and shot into cosmic space, as dark and serene as always. In flight I looked back and saw, albeit in a fractured fashion, the spreading influence of Altruizine—spreading, since the rivers and streams were carrying the drug farther and farther. I saw what happened to the birds of the forest, the monks, goats, knights, villagers and their wives, roosters, maidens and matrons, and the sight made my last tubes crack for woe, and in this state did I finally fall, O kind and noble sir, not far from your abode, cured once and for all of my desire to render others happy by revolutionary means…' (Cyberiad, 1974, p. 281)

I have argued that the effects of empathy are complex and may fail to do much good for the individual, the subjects of one's empathy and the common good. Admittedly, I have brought forward evidence selectively focusing on the harmful effects of high levels of empathy. I have aimed to provide a counter-weight to arguments about the beneficial, pro-social and morally enhancing value of empathy increase. This is sufficient for the purpose of the argument. However, I do not claim that *decrease* of empathy is necessarily conducive to moral outcomes. I am also not claiming that it is senseless to talk about prudentially or morally appropriate levels of empathy. Rather, I argue for a more qualified claim that empathetic ability is like any other cognitive and affective capacity: it is multi-purpose, its effects are context dependant and, in this context, more is not always better.

Rifkin (2010a) was wrong when he contrasted Maslows' (1969) 'new' insight about the importance of empathy with 'emotionally incapable' Enlightenment philosophy. Hutcheson's affective psychology and phenomenology developed in *Inquiry into the Original of our Ideas of Beauty and Virtue* (1725), *Essay on the Nature and Conduct of the Passions and Affectations* (1728a), and *Illustrations on the Moral Sense* (1728b) built on Shaftesbury's notion of an inborn moral sense. Hutcheson understood, and held that, a crucial feature of our moral evaluation is that we approve affections that are irreducibly benevolent and other-directed, at the same time condemning inappropriately selfish ones while looking at how agent's actions flow from benevolent affect towards other sensitive beings. In turn, Hume and Smith built on Hutcheson's ideas.

Hume locates all our motivations in the passions. Perhaps for this reason, he treats the will in his discussion of the direct passions, identifying it as 'the internal impression we feel and are conscious of, when we knowingly give rise to any new motion of our body, or new perception of our mind' (*Treatise of Human Nature* 1739-40, II.3.1 399, see also: *Dissertation on the Passions* in *Four Dissertations* 1757, *Enquiry Concerning the Principles of Morals*, 1751). Hume evokes the importance of sympathy in his justification for the motivating character of what he called 'artificial virtues:' such as justice and promise keeping. The sentiment he talks about, however, is 'an extensive sympathy,' redirected through general rules and the social convention toward society as a whole. This sympathy in turn requires correction, so that our sympathy is not directed only towards our kin. We have to direct our passions beyond their natural bounds, so that it allows us to approve of the justice or honesty of all sorts of people in all sorts of situations, regardless of their connection to us.

According to Hume, the general point of view does not provide a standard of rationality but it does provide a standard of appropriateness – and this standard allows us to shape, cultivate and constrain our sentiments in ways that provide the sort of stability and reliability that will form the basis of shared judgment. In

Theory of Moral Sentiments Smith responds, modifies and extends Hume's account. Although Smith discusses different 'spectator positions' one can adapt, the basic idea that the ability to adopt the stance of spectator of our actions and sentiments means that we can evaluate and modify our emotional reaction.

Perhaps' the sympathy of the ideal observer in Hume and Smith is too 'passive' for Rifkin's taste. Or perhaps it is too 'active' in its involvement of reflection. In contrast to Rifkin (2010a) and Baron-Cohen (2011), Scottish Enlightenment sentimentalists understood the importance of reflection and reason in morality, even when, as for Hume, reason was only useful in establishing the ends.[9] Even Hutcheson argued that although moral judgments ultimately rest in specific kinds of emotions, the exercise of the benevolent moral sense calls for additional reflection *beyond* a certain element of reflections present in all our affections. Even this minimal role of reason – well understood by the fathers of sentimentalism – seems to escape the radical proponents of achieving moral and planetary bliss through a radical increase in empathy.

3.4. Moral and pro-social behaviour

3.4.1. Is pro-social always good?

Some commentators (e.g. Persson and Savulescu, 2008) have proposed that moral enhancement could be achieved via modulation of pro-social behaviour or its underpinnings, such as empathy and trust, as well as the reduction of emotions (e.g. anger) that commonly underlie socially harmful (e.g. violent) behaviour.

Providing a clear-cut and exact definition of what is moral as opposed to the pro-social is a daunting task. But are they the same? We might value pro-sociality: we generally like people that are kind, empathetic, altruistic and helping. They are nice people. Those characteristics might also usually be conducive to all-things-considered good outcomes. However, we could come up with scenarios where some manifestations of those traits will lead to morally undesirable outcomes. To give a trivial example, shying away from harshly criticizing a plan that is likely to have

[9] My understanding of Hume is that he does not mean that passions are *sufficient* for morality but rather the they are *necessary*.

disastrous consequences might be kind and pro-social, but lead to overall morally bad outcomes and therefore be hard to justify morally. A consideration about what is conducive to the good seems to be the distinguishing feature of morality. Yet several scientists and philosophers keep identifying the moral with the pro-social.

In her book *Braintrust,* Churchland (2011) proposed that morality or ethics is a scheme for social behaviour. She proposed that it is rooted in four systems: involved in caring behaviour, theory of mind, problem-solving in social contexts and social learning. There is nothing troubling in those claims as they stand. Morality can indeed be seen as one of the normativities that regulate social conduct, and if we are interested in the description of what neural circuits are involved in moral decision making, and in the evolutionary roots of the capabilities necessary for morality, a view like this is promising.

However, the trouble begins if the pro-social is identified with the moral without acknowledging that an additional argument is needed to justify this, and the terms ‘moral’ and ‘pro-social’ are used interchangeably. Even if morality is rooted in or uses brain mechanisms involved in guiding social behaviour, it does not make morality and sociability or pro-sociality the same. Yet even Jotternad (2011), who criticizes the current approach to MB, seems at one point in his paper to equate sociability with morality (p.7).

A more elaborate argument against equating the moral with the pro-social runs as follows. Enhancing pro-sociality, be it empathy or helping behaviour, is simply not enough for behaviour to be moral. Increasing general empathy might lead to a disproportionate increase in empathy for the suffering of the ingroup; if this is perceived to be inflicted by members of another group it might stimulate increased hatred towards the outgroup (Pacholczyk, 2011). Being pro-social might mean aiding in doing evil and, as Harris (2012) pointed out, we can help others to achieve different ends and end-states, including helping them into an early grave.

The opposite might be also the case. Emotions like anger may lead to violence, but may also modulate the perception of unfairness and be involved in actions that are motivated by moral concern (Mullen and Sitka, 2006). It is not anger itself that is morally problematic, but rather its violent expression. What is more, sometimes violent behaviour can be justified morally and even lead to morally best outcomes (Harris, 2011).

The way anger is displayed is modulated by a set of cognitive processes, and this mediation is complex. That means that not only are there likely to be different moral consequences of behaviours heavily influenced by emotions such as anger, but also that pro-social and anti-social consequences are likely to be variable depending on the context. To illustrate this point let us consider some recent research on conflict resolution.

Halperin (2008) conducted a study at the eve of the Annapolis peace summit in 2007 and found that above and beyond any other emotion, a construct branded as 'sentimental hatred' increased the tendency of Israelis to support extreme military action toward Palestinians. Researchers suggest that the level of long-term hatred influences the behavioural manifestations of anger. Not surprisingly, anger that occurs in the presence of high levels of hatred will most likely bring about an extreme aggressive reaction. In contrast, anger that occurs in the presence of low levels of hatred may lead to more constructive approach tendencies (Fischer and Roseman, 2007). If one believes that the opponent group can change its behaviour and that its intentions are defensive or innocent (and such appraisal is connected to the low levels of hatred), the appraisal embedded in anger may create a tendency to engage in constructive problem solving and crisis management, instead of violent behaviour (Halperin, 2008; 2011). This illustrates the point that emotions such as anger have both pro-social and anti-social consequences.

To argue that reducing anger, increasing empathy (whichever aspect of it) and increasing trust will lead to pro-social consequences is therefore an oversimplification. Branding traits and emotions as pro-social is an oversimplification, because what makes them pro-social or not is the behaviour they are likely to elicit – and that might mean both pro- and antisocial behaviour. Moreover, whether or not the pro-social behaviours such as attempts at conflict resolution or helping are indeed the best morally speaking is a separate issue altogether. Some studies have found that participants influenced by empathy for an individual might act contrary to what is all-things-considered just. Batson et al. (1995) have shown that people who are induced to feel empathy for a terminally ill child are more likely to unfairly allocate resources to this individual. They would, for example, move her off a waiting list and into immediate treatment even where that implies that others on the waiting list do not get the treatment they need.

Branding pro-social traits as moral is especially worrying given the real-world context of the philosophical debate on MB. In recent years the convergence of the criminal justice and mental health systems in the UK means that pharmacological and other psychiatric interventions are easier to deliver to offenders. With the recent focus on the management of so called dangerous individuals in the UK, as well as the history of the public health perspective on violence in the US, some might see the time to be ripening for biomedical interventions. It certainly sounds better to brand an intervention ‘moral enhancement’, while it would be more accurately called crime control or reducing anti-social behaviour.

This is not to say, however, that an ethical argument in favour of modifying emotions or traits underlying behaviour could not be made. If we could provide good reasons for the view that, other things being equal, modulation of a certain trait is likely to provide morally best outcomes overall, we would have the beginning of the ethical argument we need. Here I will not attempt to fully develop arguments of this kind, but let me provide an indication of how such an argument might go. One way is to argue that for a given individual it may be overall morally better (perhaps over a period of time) to have a given trait or level of emotional response. We therefore need an account of the good, and empirical evidence that systematically ties this certain level of trait or emotion to what is morally desirable. If we accept that behaviours are what matters here, we would have to demonstrate that direct emotional modulation is likely to have an effect on behaviours, and know enough about the situations that a person is likely to encounter to demonstrate that this modification will be better overall.[10] This kind of argument applies equally to the enhancement of cognitive abilities and to modifying morally-relevant behaviours.

Secondly, one could take a population perspective and argue that an increase or decrease in a certain trait or characteristic would have overall morally best

[10] Which might require responding to the situationist critique

outcomes, even taking into consideration counterexamples.[11] We could focus on certain subpopulations (analogically to the targeting the high-risk populations in disease prevention) or adopt a population strategy which would focus on shifting the entire distribution of a trait (Rose, 1981; 2001).

3.4.2. Population perspective

To see what aspects we would have to take into consideration in order to demonstrate the desirability of a population-wide intervention, let us consider an example. Assume that the reduction of anger in a given subpopulation might help in preventing negative consequences and make the subjects of the intervention behave in a way more conducive to moral outcomes – for example, by reducing the probability that they commit violent crime. One way, as Harris has proposed, is to promote better impulse control. If that cannot be done – because it is too costly, impossible or otherwise impractical – the second way is to weaken the underlying impulse. Changing the underlying impulse might be very difficult to achieve selectively, i.e. by targeting only those with a high propensity to feel anger, who at the same time are not able to selectively inhibit harmful displays of anger. If that claim turns out to be true, we have a *prima facie* case for a population-based intervention.

Both strategies (targeting high-risk individuals and whole population interventions) have their advantages and disadvantages (Manuel et al. 2006). In population-wide interventions, each individual usually only has a small expectation of benefit, and some will not benefit at all. This small benefit can be outweighed by a small risk. This happened in the World Health Organization clofibrate trial, where a cholesterol-lowering drug seems to have killed more than it saved, even though the fatal complication rate was only about 1/1000 per year (WHO, 1978). Such low risks, which can be vitally important to the balance of overall outcomes, may be hard or impossible to detect. If our rationale for the intervention is to reduce violent crime and deaths caused by violent crime, it would be a Pyrrhic victory if more people died and suffered lower a quality of life then were saved, or otherwise benefited, as a result of the intervention. Analogically, it would be a Pyrrhic victory if the improvement in the behaviour of some was more than offset by a decrease in

[11] I am grateful to Soren Holm for that point

morally desirable behaviours in others. We do not propose indiscriminately lowering populations' sex drive only because there are some sex offenders. In that case we can even put the issue of liberty aside; we do not propose it because the calculation of the loss of utility to many is simply not worth the prospective and speculative gains.

So to be able to argue for a population-wide intervention that would result in the individuals in the population being, overall, morally better, we have to show several things. We have to show that the modification is indeed a modification in our moral sphere – the assertion that a proposed modification is pro-social is not enough. Moreover, we should demonstrate a reasonable expectation of efficiency – cost-effectiveness as compared to alternative interventions – and show that the side effects are unlikely to have consequences overweighing the benefits brought by the intervention. Our overall assessment would also include the effects of the modification of traits on peoples' moral lives in general. Again, this kind of argument can conceivably be made, but it has to be informed by moral theory. Unfortunately for many of the commentators, 'more empathy' or 'less anger' is simply not enough.

3.4.3. Moral modification instead of moral enhancement

The focus on increasing pro-social behaviour and branding it as moral enhancement might be seen as an attempt to circumvent the complex, long and messy normative debates, and to provide non-controversial examples of moral enhancement (understood as making morally better people). However, as demonstrated above, those claims also depend on the conception of the good and a view on what morality is about – even if those assumptions are not spelled out explicitly. Those assertions are not morally neutral or free of moral theory.

This poses a practical problem. Does it mean that all accounts of moral enhancement should come with a certain moral theory, along with a conception of the good, the right and the just? I argue that this is not necessarily so, as long as it is stated explicitly what one means by moral enhancement. In Chapter 1 (see also: Pacholczyk, 2011) I have provided a brief analysis of the concepts of 'moral' and ‘enhancement’. The vagueness and ambiguity of the concepts might give rise to a different interpretation of them. Perhaps the most intuitive understanding of the phrase ‘moral enhancement’ is ‘making morally better people.’ This is how I have interpreted this phrase in this chapter, in order to be able to engage in the current debate. This interpretation, however, might be restricting our discussion about the prospects brought by increased understanding and control over the biological underpinnings of morality. A less intuitive take on moral enhancement might be more suitable for our discussion. The phrase ‘moral modification,’ consistently with the use of ‘MB’ in this work, could better capture what we have in mind, that is, a greater degree of value-neutrality, in the sense of leaving open the question of what exactly is morally good and accepting that there are different conceptions of the good. The focus here would be on the possibility of modifying and influencing various aspects of our moral functioning, rather than on whether it is conducive to the good, morally permissible or obligatory.

A similar approach, however, was criticized by Jotterand (2011). Jotterand characterizes the current moral discourse as being emotivist, that is, equating moral decisions with expressions of preference, an attitude or feeling. Supported by MacIntyre’s (1981) critique, he states that ‘the modern and emotivist self represents its own self-ruling point of reference and therefore does not need to be liable to anyone as moral agent.’ (Jotterand, 2011, p 5). That, according to Jotterand, introduces a problem for achieving consensus due to competing interests and visions of the good. As an upshot of that observation, Jotterand criticizes transhumanists as not being able to provide an answer to the question of what is an optimal level of morally relevant traits or moral emotions, e.g. empathy or moral indignation (2011, p.7).

However, the focus on modifying emotions and the acceptance of pluralism need not imply emotivist ideas about morality, nor the disposing of the need for reasoning and moral theory informing moral judgments. I am not sure whether transhumanists are indeed implying that there is no need to take a philosophical or normative standpoint or attempt to be value free. Not taking a normative standpoint need not imply that there is no need to take it at some point. So it is with the empirical investigations of morality. There is no need of arguing what exactly ensures best outcomes all-things-considered, what constitutes a virtue and so on (although such proposals are naturally welcomed) just at that point. The normative appraisal comes later and is a separate step (see also Chapter 1). Thus, the objection that the approaches which emphasise the manipulation of moral emotions do not provide any content to guide one’s behavioural responses (Jotterand, 2011) is a good observation, yet provides no good argument against the use of emotional modification.

Why is it preferable to remain neutral[12] at first as to the question of what things are intrinsically good, and to move later to a normative appraisal of what level of a trait or characteristic is conducive to the good? In other words, why is it advisable to explore our ‘coulds’ before we explore our ‘shoulds’? There are several reasons. Firstly, this approach allows us to accept pluralism in conceptions of the good, a fact that need not be as troubling as Jotterand seems to imply. It also allows for discussion about – and a change in – our conceptions of the good over time. It also makes it possible for the reader to plug in his or her substantive axiological view. Secondly, it allows for context-sensitivity. Even assuming that a general understanding of the good is constant, there might be a difference in what level of trait is conducive to the good, depending on, for example, the characteristics of the agent (e.g. the level of ability to inhibit unwanted impulses) and the agent’s situation. Arguing for a universal optimal level of empathy in general and for everyone, without considering those contextual cues as well as the importance and the guiding function of reason and moral beliefs, could be seen as both naïve and unnecessary universalism.

However, neutrality of the kind I have proposed need not deny the importance of moral theory, moral beliefs or a normative moral standpoint in general. In the next section I will explore the importance of engagement with moral reasons in the context of MB aimed at facilitating moral agency.

[12] Relatively neutral. I do take Quine’s point that observations are value laden, but there is a substantial difference between being not totally value-free in our views and pushing for a concrete axiological view. Perhaps there are some restrictions embedded with the view I am proposing; for example, it might be seen as offering a certain account of moral agency or autonomy – a one that which values control over one's emotions. However, this approach allows us to discuss a wider range of interventions, while ignoring a range of possible differences in the conceptions of the good for the time being.

3.5. The importance of moral reasons

Harris (2012) has argued that in order for a given intervention to be considered a moral enhancement understood as making morally better agents it is not sufficient that a morally better outcome is achieved. Rather, behavioural change must be achieved for moral reasons. Decisions regarding morally relevant matters can be made for a number of reasons, including self-interest, aesthetic preferences and so on. It might just so happen that a decision about a morally relevant matter was reached for non-moral reasons, but is consistent with an all-things-considered morally motivated decision, should such a decision have been made.

Consider the following case. After a successful job interview for a company that is known for its engagement with a local community, Derek is walking down the street with a prospective boss. They are approached by a man who explains that he is homeless and asks them to buy him some food. There is no particular reason to doubt this man's story. Normally Derek would have politely declined and continued walking, reminding himself that he gives enough money to charity to feel that 'he did his bit,' that there are charities that provide food to those who need it, which in his view discharges his *prima facie* obligation to help at little cost to oneself. The presence of his employer makes him stop and think 'if I decline, this would make a bad impression on my boss.' Based on the negative consequences for his reputation, he decides to buy lunch for the man. As John Harris has put it, the fact that one is doing good does not mean it is specifically moral behaviour; it is simply morally relevant. Although there might be problems (which we will shortly address) with a robust practical application of the distinction between, on the one hand, deciding to do what is conducive to morally best outcomes and, on the other, doing what is morally best for moral reasons, the basic point stands. The point that for an action to count as moral action (whether it is a right or wrong action) it has to be done for the right kind of reasons, can be made within broadly consequentialist framework such as Harris,' from deontological perspective and in the context of virtue ethics.

Harris takes the argument further and offers a case against MB. Harris does not object to a biomedical intervention that, for example, mitigates xenophobia by enhancing general cognition, thereby reducing the tendency to hold false stereotypic beliefs. He puts such interventions in the same unproblematic category as ‘bringing up children to know the difference between right and wrong’ or ‘general education, including self-education, wide reading and engagement with the world’ (Harris, 2011, p. 104). However, he argues that MB using methods that affect emotions directly is not morally desirable if done with bettering moral agency in mind. What he objects to is the attempt to enhance morality via the direct biomedical modulation of emotions (Douglas, 2008) – that is, without the intermediate step of increasing the accuracy of cognitive processes (such as reasoning) or cognitive states (like beliefs) (Harris, 2011).

Firstly, the point made here might (again) be that for a MB to genuinely count as enhancement of moral agency, it is not sufficient that we end up with people who simply act in a way that is conducive to morally best outcomes. They also have to do it for the right sort of reasons – that is, moral reasons. Direct biomedical emotion modulation as proposed by Douglas (2008) and Persson and Savulescu (2008) might be seen here not as moral enhancement (in the sense of making better moral agents) but rather as bypassing morality altogether. In so far as that is accepted to be correct, this might suggest one argument against the ethical desirability of MB’: the concern that pursuing moral improvement through direct emotion modulation might in fact cause moral decline and thus is morally undesirable.

Troubling as this concern might be, I would like to question both that such direct biomedical modulation of emotions necessarily falls outside the realm of moral enhancement (understood as making morally better agents), as well as the connected claim that MB is morally undesirable. However, those arguments will have to wait until the last chapter of this thesis, in which I will argue that voluntary MB, embedded in appropriate reflection, can constitute bettering of moral agency (and so moral enhancement according to the definition of many commentators) also within a perspective that acknowledges the importance of moral deliberation and all-things-considered judgement from a moral stance. Before this argument, however, I will examine three objections to MB: relating to moral disagreement, medicalization and narrative identity.

3.6. Conclusions

In this chapter I have argued that modulation of pro-social sentiments is not sufficient for moral enhancement understood as making better moral agents. After briefly introducing the debate I have considered the proposals according to which increasing empathetic ability should be pursued as a solution to great global ills (Persson and Savulescu, 2008; Savulescu and Persson, 2012b, Rifkin, Baran-Cohen, 2011, Rifkin, 2010a). I have argued that increasing pro-social sentiments is not sufficient for assuring oral outcomes, and increasing empathy or a predisposition to help might also lead to adverse outcomes morally speaking. Conversely, inclinations and emotions considered anti-social such as anger, might result in morally good outcomes. Thus, the identification of the moral and the pro-social is not justified. Moreover, increasing empathy has serious limitations in the extent to which it is likely to translate into the kind of moral behaviour that would lead to solving great social ills.

In section 3.4.2 I argued that the population perspective on the modification of pro-social sentiments for the purposes of increasing doing good is faced with problems due to lack of selectivity and individual differences (see also s. 2.3.1) and concluded that an individual approach to MB that embedded in and taking advantage of the guidance function of reason and moral beliefs is more likely to be conducive to good moral outcomes. I have further considered the importance of moral reasons in section 3.5., arguing that engagement in moral reasons is necessary if MB is to result in enhanced moral agency. I have argued against a position according to which modulation of affect is not a desirable way of enhancing moral agency, and instead argued that MB via direct emotion modulation, provided it is done in the context of moral reflection and deliberation, is a prima facie useful tool when pursuing bettering of moral agency.

CHAPTER 4. Moral disagreement

4.1. Introduction

In this Chapter I discuss the view that making people more moral is practically impossible due to the lack of consensus about what is and is not moral. I ask what are the consequences of moral disagreement for the use of MB. First, I examine what cases of moral and social modification the objection applies to. I then discuss the concept of fundamental moral disagreement and its implication for what we should do. I argue that even the presence of fundamental moral disagreement does not give us a good reason to abandon our moral beliefs and that those disagreements are better accommodated on the level of political deliberation. MB is unlikely to affect substantive endorsed beliefs directly, so the challenge only arises when we try to assess whether a certain change in function is conducive to morality or moral outcomes. Disagreement here does not preclude meaningful improvements in moral capacities; it simply means that where our axiological views differ, we might differ in the assessment of whether an intervention is or is not moral enhancement in the sense of making better moral agents.

4.2. Moral disagreement and MB

Some may doubt the plausibility of MB, or even doubt the reasonableness of pursuing it, based on the claim that there is a substantial and possibly irreconcilable disagreement as to what is a moral way to go about things. If we disagree about what is moral, the argument might go, we cannot know which way we should modify our moral sphere – we do not even know what the goal of the modification should be! There is at least one understanding of moral enhancement to which this doubt does not apply – moral enhancement understood as a modification in the sphere of moral functioning that is in the person’s self-interest outlined in section 1.2.3. Being better or worse moral agent is not at issue here, and so the doubt does not have a bite.

And what about moral enhancement defined as making people better moral agents? It seems that the argument is in this case, at least *prima facie*, plausible. If moral disagreement undermines our moral knowledge, it could have consequences for the project of making people more moral, be it by non-traditional means and traditional ones such as moral education. But let us have a closer look.

Although there is a good amount of disagreement about what moral education should look like, most of us would not say for that reason that it is better to have no moral education at all, that we should not teach our children that lying is wrong or that striving to further develop ourselves as moral agents in our adulthood is a misguided proposition. Why is that? Firstly, because despite possible theoretical differences, there is a good amount of consensus about which acts or kinds of acts are morally wrong or morally right. On most, if not all, reasonable accounts of what is moral, killing a person for no other reason apart from the pleasure one derives from this act is wrong. There is also substantial agreement that, generally speaking, we ought to keep our promises or avoid lying.

Moreover, there is a substantial amount of consensus about things that are necessary or conducive to moral agency and sensitivity, and conducive to morally good kinds of motivation, outcomes, etc. To give just one example – one of those things is concern and respect for other moral agents, which in turn requires a number of cognitive and affective capacities. The certain amount of agreement (more or less limited, depending on how high we will put the bar) means that the objection from disagreement does not apply to the numerous instances when disagreement is absent or weak enough. Objections from disagreement will not apply to improving our ability to be moral in those cases.

The presence of disagreement is often used to demonstrate the inadequacy of moral realism, and so to justify certain conclusions about the metaphysics of morals.[13] However, this argument is susceptible to the objection that it proves too much and – since it is an inference to the best explanation – the objection that there are alternative explanations of moral disagreement. When moral disagreement is present, it can be the result of several factors. It may be the result of disagreement about non-moral facts, both about morality and about the world. The disagreement can also have its source in some kinds of procedural failure in the reasoning

[13] For example, see Mackie's (1977) well-known 'argument from relativity'.

process. Alternatively, the apparent disagreement may be an instance of the case when people are talking past each other, and do not understand each others' claims (Harman, 1975; Wong, 1984). In those cases we may hope that at least some disagreement may be removed. Abilities necessary to engage in a collective inquiry and discussion with others may be helpful in facilitating this process. Some kinds of enhancement (enhancement of reasoning skills, for example) may aid us in being better prepared for that process.

4.3. Fundamental moral disagreement

Some have assumed that all moral disagreement is in fact due to those reasons (e.g. Boyd, 1988), while others maintain that there are cases of moral disagreement between two people who are equally rational, and equally well informed about the non-moral facts and understand each others' claims (fundamental moral disagreement). Whether that is indeed the case seems to be a rather complex question and I will not attempt to give an answer here. However, *even if* we accepted the conclusion about the metaphysics of morals, it does not have straightforward implications for the possibility of moral actions and moral concern even in the case of fundamental moral disagreement. Why is that? It is because we cannot automatically get from the metaphysics of morals to the conclusion about moral knowledge and about what we should do. Let me just give one example of this – there are alternative metaphysical positions that have the potential to deal with the objections raised. It is possible, for example, to accept error theory and end up with moral fictionalism, where our make-believing in morality can be prudentially advisable (Joyce, 2005). Alternatively, moral non-cognitivists may seek to explain how the feelings, attitudes or prescriptions expressed in moral claims can be justified (see Hare (1981), Gibbard (1990) and Blackburn (1998) for theories of moral justification compatible with non-cognitivism). Those views can account for the apparent fundamental moral disagreement while leaving moral enhancement (via MB or moral education) as a viable notion.

Although the most common, metaphysical arguments are not the only ones developed on the basis of the observation that moral disagreement exists. For example, McGrath (2007) defended an epistemological version of this argument.

Epistemological arguments from disagreement seek to undermine moral knowledge by showing that regardless of the metaphysics of moral facts, we can reasonably expect to have much less moral knowledge that we previously thought. Consider the following passage from Sidgwick's *The Methods of Ethics*:

> '[I]f I find any of my judgments, intuitive or inferential, in direct conflict with a judgment of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two judgments necessarily reduces me temporarily to a state of neutrality.' (Sidgwick, 1907, p. 342).

McGrath (2007) develops a parallel argument that applies not to certainty, but rather to moral knowledge. When moral beliefs are subject to disagreement and Sidgwick's condition is satisfied (that is, if one has no more reason to suspect that the other person is mistaken than that it is oneself who erred), one is not holding knowledge about the contested issue; and that is the case even if the belief happens to be true. In fact, McGrath (2007) develops a stronger version of this claim by arguing that all controversial moral issues (such contentious matters in applied ethics and culture) fulfil Sidgwick's condition; let us accept this last claim for the purposes of the argument. What consequences does it have for the project of using MB for making morally better agents?

The consequences are far from straightforward. In those cases it does not follow that we should abandon, prohibit or find MB an untenable proposition – and that applies to both moral education and other non-traditional means of enhancement. Firstly, in cases that apparently satisfy Sidgwick's condition we may still have some problems with justifying why exactly it is rational for us to trust others' moral intuitions as much as we trust ours, and why, as a consequence, we should abandon our belief (Wedgwood, 2010). But let us assume that some version of Sidgwick's proposal applies and so in many cases of controversy it is rational for us to abandon our beliefs.

Non-traditional MB is unlikely to be specific enough to change the moral appraisal of any particular controversial issue. It is more likely to slightly modify some propensities to react, perceive and behave by increasing impulse control, empathy, trust or reducing fear responses and so on. Naturally, we can still disagree about issues such as whether a higher level of trust is conducive to moral outcomes. However, if we accept Sidgwick's (1907) advice to hold our judgements we are still left with the question 'so what should we do now?' Let us say that we disagree about whether Jane should increase, decrease or maintain her empathic ability (we fundamentally disagree about all three possibilities). What behaviour would constitute holding our judgement on this issue? Some may say that we should leave things as they are. But there is no reason why we should privilege the *status quo* option over other possibilities, given that there is disagreement also about the *status quo*. Thus, moral disagreement is problematic as a support for leaving things as they are. We are still faced with the question 'what should we do next?' The answer could be that it is only rational for us to have *no moral views* at all on the contentious matter and use other reasons to decide on the course of action.

It is important to remember that we have developed political means of dealing with moral disagreement and sometimes find disagreement to be a constructive force necessary for change. In liberal societies moral education is often about developing the ability of persons to be autonomous moral agents, providing them the possibility of gaining reasoning skills and exposure to moral problems to aid this development. We tend to protect the freedom of people to disagree with commonly held views. We also have political frameworks that aid us in dealing with moral disagreement and often seek the state to be as neutral about issues of morality as it is possible. We tend to protect the private sphere – the freedom of parents to raise their children as they see fit is interfered with only in cases of clear parental failure; we struggle to protect freedom of conscience, and so on. We accept that people have different ideas about what a good life is about and value the ability of individuals to act consistently with their idea of the good life and morality, and, generally speaking, restrain this possibility only when we have strong justification for doing so. Even given the doubts that an agent may have about what is right, we are likely to find the adoption of a moral stance (for example, as opposed to narrowly self-interested stance) to be valuable.

One could argue that the possibility of MB in this liberal framework would be likely to deepen the disagreement – which could be seen as undesirable prudentially or morally speaking. We may therefore have good reasons to make people less bothered about morality in cases when disagreement arises (this would be a solution consistent with the view that it is rational for us to abandon our belief in certain cases of disagreement). Interestingly, an argument for making people suspend their judgment and not act motivated by moral reasons under these particular circumstances is an argument for a certain kind of enhancement. If one supported this argument using *moral* reasons this would be an argument for a specific kind of moral enhancement understood as making people more moral. If the rationale is prudential, we have a case for prudentially beneficial intervention into our moral sphere.

4.4. Conclusions

To sum up, moral disagreement has much less straightforward consequences for MB that we may have thought at the outset. Firstly, it only applies to moral enhancement understood as making people more moral, and not to moral enhancement as a prudentially beneficial modification of the moral sphere. Also, it does not apply to a whole array of issues that we tend to agree on, including the issues of what is conducive to morally desirable moral sentiments, motivations, outcomes, etc.

If we treat moral disagreement as giving rise to a valid and strong argument against certain views on the metaphysics of morals, there is still much explaining to be done of what impact it should have on our moral knowledge and subsequent actions. We can, for example, adopt a non-realist view of morality that is not susceptible to the objection from disagreement and work from there. What we have learned from the discussion on the possible sources of disagreement is that disagreement about non-moral facts, procedural failure, bias and lack of proper discussion can all give raise to disagreement about moral issues. We may therefore have a good reason to support both traditional and non-traditional means of improvement that would aid us in dealing with those disagreements better then we now do.

It is also unclear how an epistemological argument from disagreement should impact our behaviour, but it is unlikely to support the *status quo*. If we indeed think that holding our judgement means abandoning moral considerations in controversial cases and that this is what we *ought to* do, we may have a good case for a particular kind of MB. Also, let us not forget that we have political means of dealing with moral disagreement. Respecting moral agents' decisions and allowing moral agents to pursue their idea of the good in the private sphere (and discuss and argue for it in the public sphere) is one of them. Unless we have other strong reasons to treat non-traditional moral enhancement differently, this also applies to those cases of MB.

CHAPTER 5: Medicalization

5.1. Introduction

The recent scientific focus on the neuroscience of sociality and morality, the expansion of diagnostic and psychiatric assessment categories to encompass more social aspects of our behaviour (e.g. psychopathy, personality disorders, social anxiety) and the increased search for treatment of the conditions means that our social and moral behaviour is increasingly under a scientific, psychiatric and medical gaze. The medicalization of morality is not new: psychiatrist Thomas Szasz (1963) and sociologist K. Irving Zola (1972) first described how modern psychiatry helped to medically redefine conduct previously viewed as sin or crime (e.g. heavy drinking, homosexuality, masturbation, suicide). Although public health and psychiatry have long been concerned with social behaviour (Conrad and Schneider, 1980), the recent boom in neuroscience research and the increased interest in pharmacology and brain stimulation for social and moral enhancement (in the sense of making morally better agents) – happening in the context of the increasing efficacy of biomedicine – means that the proposed medical solutions may become more and more effective in changing our behaviours.

Yet the widening reach of the scientific and medical gaze meets with opposition. Cognitive enhancement, mood enhancement and MB have all been subject to critiques that see medicalization as a morally problematic process. This chapter asks whether using specifically biomedical means of moral modification gives rise to a strong ethical reason to forgo using MB. In order to do that, it aims to explore and examine some critiques of medicalization, untangling the various threads running through the debate. First, the concept of medicalization will be explained with reference to early sociological critiques and the more current approaches to medicalization. I propose that ‘medicalization’ should be used as a descriptive term, pending the moral assessment of the process. Second, the chapter explores objections to medicalization, including the ‘category-mistake’ argument, the view that ‘normal and usual’ traits should not be medicalized, the stance that unpleasant experiences are often justified and necessary for a full and flourishing life, the social control worry, the critique of an exponential growth of medical care as well as the objection to what is seen as the undermining of autonomy and responsibility.

Thirdly, the chapter brings forward the positive aspects of medicalization. The chapter concludes that it is important that we attend both to the advantages and disadvantages of the medical and scientific gaze. After examining the force of various arguments related to medicalization, this Chapter concludes that the arguments related to the critique of medicalization do not provide a strong reason to forgo MB generally speaking.

5.2. What is medicalization?

5.2.1. Medicalization – early sociological approaches

For the last thirty or forty years, sociologists have used the term *medicalization* to refer to the process by which 'non-medical' (or 'life' or 'human') problems become understood and treated as 'medical' problems (Conrad, 2007, pp. 3-4). While early critics of medicalization focused on psychiatry (Szasz, 1970) or a more general notion of medical imperialism (Illich, 1975), sociologists began to examine the processes involved in the expansion of medicine's realm (Freidson, 1970; Zola, 1972). As sociological studies on medicalization accumulated (see Conrad, 1992; 2000) it became clear that medicalization went far beyond psychiatry and was not always the product of medical imperialism, but had arisen at the intersection of complex social forces. According to Conrad, the research then focused on the definitional issue: defining a problem in medical terms, usually as an illness or disorder or using a medical intervention to treat it. Many early sociological studies took a social constructionist approach to those problems, with the focus on the construction of new medical categories, an increasingly medicalized approach to conditions such as ADHD, menopause, alcoholism and PTSD and the resulting expansion of medical jurisdiction (see: Conrad, 1992).

Conrad suggests three broad areas of focus found in the sociological studies of medicalization in the 1970s and 1980s that aimed at explaining the causes of medicalization: the power and authority of the medical profession, activities of social movements and the influence of professional or organisational actors. Firstly, medical professionals sometimes were at the center of the move towards medicalization, such as in the cases of hyperactivity, menopause, child abuse and child-birth, among others. Physician involvement was considered through the lens of concepts such as professional dominance, physician entrepreneurs or medical colonisation. Secondly, medicalization was thought to be fuelled by the activities of social movements and interest groups, which argued, campaigned and lobbied for a medical definition for a problem or to promote the veracity of a medical diagnosis. A clear example is alcoholism, with the involvement of Alcoholics Anonymous and a wider 'alcoholism movement,' while physicians remained mostly at the backstage of the events. Social movements were also critical in the medicalization of PTSD (Scott, 1990) and Alzheimer's disease (Fox, 1989), although some efforts by activist groups were less successful, such as the case of multiple chemical sensitivity disorder (Kroll-Smith and Floyd, 1997). Third, organisational or professional agents sometimes played a prominent part in medicalization. The examples include the rise of obstetricians and the demise of midwives in some countries (Wertz and Wertz, 1989) and the rise of behavioural paediatrics (Pawluch, 1983; Halpern, 1990). Although in early studies of medicalization, different stakeholders were also mentioned (such as the role played by pharmaceutical innovation and marketing in hormone replacement therapy), the causal role of health care funding and pharmaceutical companies was considered to be of secondary importance.

5.2.2. New approaches to medicalization – biomedicalization

The social context of medicine, however, has changed. Critiques of the ways in which the medical profession has extended its jurisdiction have become part of everyday and professional debate, and the power of doctors is constrained by the law and threat of litigation, the critical eye of bioethics, the increasing imperative for evidence-based medicine, as well as by a strong focus on patient autonomy, patient's rights to health and compensation of injuries. Perhaps there is room for improvement in the way the medical profession pays attention to those constraints in practice, but currently a host of new actors and problems have exerted increased influence over the process of medicalization.

In a recent paper, Clarke et al. argue that medicalization is intensifying and being transformed: around 1985 'dramatic changes in both the organization and practices of contemporary biomedicine, implemented largely through the integration of technoscientific innovations' (2003, p. 161) contributed to an expanded phenomena they call biomedicalization. They define biomedicalization as

> 'the increasingly complex, multisited, multidirectional processes of medicalization that today are being reconstituted through the emergent social forms and practices of a highly and increasingly techno-scientific biomedicine' (Clarke et al. 2003. p. 162).

The concept is very wide and includes a wide variety of phenomena: biotechnology, medical informatics and information technology, changes in health services and the production of technoscientific identities. This new conception was criticised by Conrad for losing the focus on the definitional issues, which have always been a key to medicalization studies.

Whether we see the changes as a transformation into a qualitatively new phenomenon (Clarke et al. 2003) or as an extension of medicalization (understood as in Conrad, 2005), medicine has been and is changing. By the beginning of the 1990s we began to see impacts of the changes in the organization of medicine. As the emphasis in health policy shifted from concerns about access to cost control and care management (Pescosolido, 2006; Scott et al. 2000), some scholars noted an erosion of medical authority (Starr, 1982). Sociologists focused on deprofessionalization, decline, and public distrust (Pescosolido, 2006). McKinlay and Marceau (2002) noted that the 'golden age of doctoring' has ended in an emerging, increasingly buyer-driven system, as the physicians – like all other workers in a capitalist society – were eventually stripped of control over their work through corporatization and bureaucratization (McKinlay, 1982). Patients began to act more like consumers, both in choosing health insurance policies and in seeking out medical services (Inlander, 1998), and although this trend was especially strong in the US, it can also by noticed in countries with publicly funded health care systems, such as the UK's. In addition, new arenas of medical knowledge were becoming increasingly dominant, with the boom in scientific knowledge in neuroscience and genetics, as well as the increasing profitability of pharmacology and early applications of genetics and neuroscience. Conrad notes a change in the drivers of medicalization. He cites the three new forces contributing to medicalization as biotechnology, consumers, and healthcare funding and emphasises that medicalization is currently driven by commercial and market interests, with doctors increasingly acting as gatekeepers of technology (Conrad, 2005).

5.2.3. Medicalization: a normative or non-normative concept?

One question that arises is whether medicalization is necessarily a bad thing. Conrad proposes that the term ‘medicalization’ is descriptive. In his early writings Conrad (1975) argued that ‘by medicalization we mean defining behavior as a medical problem or illness and mandating or licensing the medical profession to provide some sort of treatment for it,’ thus emphasising both the medical definition of a social problem and medical jurisdiction over that problem. In a later review article, Conrad de-emphasized the jurisdictional aspect of medicalization and highlighted the definitional one:

> ‘[m]edicalization consists of defining a problem in medical terms, using medical language to describe a problem, adopting a medical framework to understand a problem, or using a medical intervention to ‘treat’ it.’ (1992, p. 211)

Thus conceived, medicalization is a sociocultural process that may or may not involve the medical profession, may or may not lead to medical social control or medical treatment, and may or may not be the result of intentional expansion by the medical profession. Medicalization here is a descriptive term and medicalization need not be morally problematic (Verweij, 1999).

Despite explicitly assuming a position of non-normativity, many sociologists have traditionally criticized the increasing scope of the medical gaze (Conrad, 2007) or seem to have worked on the assumption that the process is bad (Parens, 2011). Thus, many authors use the term 'medicalization' with an implicit pejorative connotation, indicating a problematic extension of medical control over more and more aspects of private and social life (e.g. Illich, 1975). However, others see medicalization as having some positive implications, arguing that medicalization may open the door to effective medical treatment for harmful conditions (e.g.

Carter and Hall, 2012, p. 231). In the proceeding sections I will not attribute any normative connotations to the term itself, but highlight and discuss both the positive and negative consequences of medicalization.

5.2.4. Efficacy and the sociological view of medicalization

The bad name of medicalization is partly rooted in instances when medicalization proceeded despite little evidence for its benefits. For example, there has been a growing concern that the medicalization of birth may be going too far (Johanson et al. 2002). While WHO guidelines prescribe that maximum rates of caesarean sections associated with good maternity care should not exceed 15 per cent of all births, almost every developed country exceeds this threshold (Walker et al. 2002). Maternity care routinely offered in western countries seems to have reached the 'perinatal paradox: doing more and accomplishing less' (Rosenblatt, 1989)

For sociologists, the discussion about efficacy of medical treatments may be less compelling, or at least marginally relevant to their pursuits. Conrad claims that:

> 'Although medical interventions are judged by how efficacious they are, the social consequences of medicalization occur regardless of medical efficacy. They are independent from the validity of medical definitions or diagnoses or the effectiveness of medical regimes.' (1992, p. 223)

Conrad thus distinguishes between social effects and medical effects, but attends only to the former.

Although such a distinction may serve a purpose – by ensuring, for example, that one's investigation remains within one's scope of expertise – normative judgements should derive from the integration of all relevant evidence and perspectives. Therefore, while sociological critiques of medicalization provide a relevant perspective which is particularly useful in identifying social consequences, it is important not to abandon concern about the comparative efficacy of particular medical interventions and various solutions to the same problem. Efficacy should be established by comparing a given medical intervention to both medical and non-medical solutions.[14] There is some tendency to discuss therapies as if they existed in ‘the ideal state’, without equally careful attention given to the side effects and costs.[15] For example, when SSRIs and other ‘new’ antidepressants enjoyed their hey-day, media reports emphasised the prospect of addressing depression with less severity of side-effect *compared to the previously available medication*. Patients may discontinue medication due to various – as yet poorly researched – side effects, which they may not have been exposed to if they had undergone an initial course of psychotherapy. Thus, perhaps we should aim at knowing the ‘net efficacy’ of an intervention, established by taking into consideration both the benefits and side-effects. Moreover, taking into account both medical and non-medical benefits and harms is important because we do not want to be ill, and even if medicalization comes with some social costs, we may fear medicalization less than we fear illness.[16] Efficacy remains an important component of discussions focused on the use of medical approaches to problems in life.

5.2.5. Medicalization and over-medicalization

Parens notes that in both bioethical and sociological literature there is a growing (yet implicit) recognition of good and bad forms of medicalization: ‘medicalization’ (which is presumably good) is contrasted with ‘over-medicalization’ (which is presumably bad; see: Parens, 2011, p. 2). Parens quotes Henry et al. (2007) who, in their argument for distinguishing between using memory-attenuating drugs to respond to Post Traumatic Stress Disorder (which they approve of) and using the same drugs to achieve non-medical purposes (which they do not approve of) write:

> ‘If memory-attenuating drugs prove effective, we argue that the most immediate social concern is the over-medicalization of bad memories and its subsequent

[14] This may sound like a common sense and obvious point. And it is. Unfortunately, such assessment does not always happen.

[15] I am grateful to Søren Holm for this point.

[16] I am grateful to John Harris for that point.

> exploitation by the pharmaceutical industry.' (Henry et al. 2007, p. 13)

Yet to say that X is an instance of over-medicalization is to state the conclusion of a moral evaluation before giving reasons for this conclusion. The next sections examine the intuitions that may underpin the assessment of medicalization as ethically problematic.

5.3. Ethical appraisal: arguments against medicalization

A common criticism of medicalization is that construing non-medical (or life or human) problems as medical problems, construing normal human variations as pathological, commits a category mistake (Parens, 2011). Shyness can be an unpleasant state that many people experience upon meeting new people. Short stature can result in unpleasant feelings in some short individuals. However, as the critic of medicalization could observe, neither sadness nor shyness nor short stature is a *medical* problem. Although feelings that can go with being sad or shy or short may be difficult, they are not symptoms of disease. Similarly, critics of the DSM-5 worried about its 'potential to pathologize and stigmatize normal human experience' (Pierre, 2013). To treat human problems as medical problems is to make a mistake about the nature of the world, and seeing clearly and living well requires that we avoid such a mistake.

The 'category mistake' objection often additionally mixes in the claim that those unpleasant parts of experience are normal and usual, for example that sadness is a normal, perhaps even essential part of a full human life. Living well requires that we learn to accept some problems, learning to affirm, rather than erase, variations in our moods, behaviours, and appearances (see: Parens, 2011). The above critique, in fact, comprises of a series of intermingled, yet separable threads. Let us look at them in turn.

5.3.1. The 'category mistake' argument examined

In a somewhat purified form, the category-mistake critique of medicalization states that it is epistemically incorrect to see life problems as medical ones. This argument relies on an assumption that there are problems that are appropriately seen as medical or not. Yet, as with many issues, there are several perspectives that one could take on them. For example, the observation that people of lower socioeconomic status tend to have worse health can be construed as a moral, political, health or economic problem. Similarly, explaining the causes and conceptualizing criminal behaviour may include referring to moral failure of individuals, condition known as psychopathy or asocial personality disorder, economic and social factors (‘the mad’, ‘the bad’, and ‘the disadvantaged’ explanations).

All those ways of looking at it put the stress on one aspect of the problem, and this may lead to different kinds of solutions. We might evaluate the usefulness of one or another focus, but we would be hard pressed to point to some objective ‘essence’ of a problem that makes it *inherently* political, moral or economic. Similarly, sadness, shyness and criminal behaviour may be approached from different perspectives. While some perspectives may be more useful or feel more comfortable than others, given that it is unlikely that there is something inherent in the way the world is structured that would force us to adopt one and not another perspective, an essentialist position is hardly justified. Thus, this critique needs an additional assumption to hold any ground.

Another weaker version of this critique of medicalization points to the concern that that the medical way of framing issues, while not a category mistake in the way that saying ‘I am my body’ is, tends to push out other approaches. The critic may accept that we need both ways of looking at ourselves to get what we want or need, but worry when this interpretation is taking over. This claim is more plausible, but it has to be 1) justified empirically to show that the mentioned ‘replacing’ of perspectives actually happens and 2) shown that an alternative construction of a problem is valuable in some way and should not be lost. This is rarely demonstrated or argued for, and thus significantly weakens the ethical appeal of the medicalization arguments.

Perhaps the issues surrounding medicalization are best seen to relate to power and economy. Firstly, we might justifiably raise a question about who is in the best position to make a difference and get us what we want. For example, even if sadness or shyness are normal and usual responses to our wider life circumstances, and being spotty or impotent are not 'medical problems in their essence', where doctors are in a position to provide some solutions, we might accept medicalization and the treatment of those conditions as 'illnesses' rather than wait for the access to those drugs to be deregulated and demedicalized.

In this context, it is worth asking whether access to certain goods and tools needs to be mediated by the medical profession, or whether it would be better to leave it up to the individual to decide. For example, we might think that some kinds of pharmacological birth control should be as widely and easily available as condoms, or we might think that for various reasons (control of side effects, picking a method from a wider array of choices, including ones that involve a minor surgical procedure, better promotes choice and safety) it is an all-things-considered better policy to provide it via the available medical infrastructure.

Moreover, insofar as medicine provides effective solutions to what troubles us, medicalization may simply serve the purpose of implicitly designating who has the skill or technology to intervene. This can come together with certification of those who provide those solutions – qualified doctors. Certification is not uncommon in different areas – for example, regulation and law may only permit qualified electricians set up new electrical installations. Certification gives a *prima facie* reason to have a somewhat greater degree of confidence in a practitioner's skill than when no certification is present. It thus also brings a certain social benefit.

Another issue arises where medical professionals act as 'gatekeepers' and thus have control over the access to technology. It is hard to say how this would play out in the scenario of moral and social enhancement. Access to new reproductive technologies, neuroscientific self-experimentation, as well as, for example, deep brain stimulation depends on the the ability to find a medical professional that is willing to provide the technology. This makes it a matter for the medical profession and, despite some decline in the power of that profession, still involves dilemmas related to the exercise of power and discretion by medical professionals (Pacholczyk, 2011).

The second question is about economy – how do we get what we want in a cheap, easy and effective way which maximises benefits while offering the opportunity of managing costs or side-effects. Providing birth control through medical professionals may simply be the best way of maximising the benefits given the social systems and structures currently in place (e.g. collective subsidising or covering the cost, expert assistance in decision-making, managing and following up on side-effects). While that may mean that birth control and the prevention and remedying of sexually transmitted infections (STIs) is medicalized, this does not prevent an individual from participating in the choice of non-medicalized birth control methods or preventing STIs in non-medicalized ways, nor does it mean that the sphere of our sexuality has been somehow washed off all other (i.e. non-medical) meanings.

5.3.2. The' proper goals of medicine' and 'the normal and usual traits' objections

This additional assumption supporting the essentialist view of 'medical' and 'non-medical' problems is usually derived from intuitions about the proper goals of medicine. However, as Parens (2011) argues, it is difficult both to formulate and to justify such conception. Firstly, although a broad conception of health and the goals of medicine are available (such as the WHO's), for the medicalization objection to be convincing, it needs to be sufficiently narrow. Moreover, Parens points out that a reader 'attuned to how institutional goals change over time with the coming and going of more and less savory political interests, however, will be wary of an analysis that assumes knowledge of a given institution's 'proper' or 'essential' or 'real' goals.' (2011, p. 3).

Conrad attempted to circumvent this problem by referring to whether or not diagnoses are viable, rather than whether they refer to 'real' medical problems: 'What constitutes a real medical problem may largely be in the eyes of the beholder, or in the realm of those who have the authority to define a problem as medical. It is the viability of the designation rather than the validity of the diagnosis that is grist for the sociological mill' (2007, p.4). But it is unclear how the distinction between viable and not viable and valid and invalid medical diagnoses would help the essentialism. Firstly, it is unclear what viability means according to Conrad (2007) and how it is different from validity. Perhaps it means ‘reasonability’ or ‘usability’ or ‘usefulness’ in the given social context, since Conrad refers to the distinction understood in social constructivist terms; yet it is still unclear how that would impact the way sociological researchers go about choosing which diagnosis to investigate and how to investigate them. Parens (2011) suggests that in practice researchers tend to implicitly use the valid/invalid distinction, since sociologists do not investigate all viable diagnoses but rather pick and choose which diagnoses to investigate as cases of medicalization.

Another thread running through many objections to medicalization refers to the intuition that usual and normal human traits should not be understood as medical problems, even if they negatively impact wellbeing. The problem with the critics’ narrow conception of the goals of medicine that this usually entails – whether explicitly or inexplicitly – is some notion of *normal* or *species-typical* functioning (Boorse, 1977; Daniels, 1985; Sabin and Daniels, 1994), usually used normatively. However, the attempts to construe the notions of health and disease with reference to those concepts are riddled with problems (see: Agich, 1983; Engelhardt, 1986; Fulford, 1989; Harris 2007). Moreover, even if we could base our notions of health and disease on those distinctions, it is far from obvious that the protection of health is the only ‘proper’ goal of medicine – doctors typically perform organ transplantations, advise about contraception (which is plausibly understood as disrupting normal functioning of the organism for the benefit of preventing unwanted pregnancy), advise employers about their workers’ health, perform immunisations, etc. (Harris, 2007; Pacholczyk, 2011). Moreover, doctors perform amputations and brain lesions to ameliorate problems and avert danger to health or life (hardly, however, by restoring normal physiological function), prescribe aspirin that keeps blood clots from forming by interfering with the ‘normal’ production of thromboxine (keeping blood clotting below the average), prescribe bisphosphonate to prevent osteoporosis often occurring in older people by modifying the usual activity of bone cells, and prescribe hormone pills to reduce menstrual bleeding .

Thus, the generally accepted toolkit of medical professionals includes interventions which 1) do not attempt to restore health or prevent disease at all (e.g. providing birth control), 2) do not restore species typical-functioning yet are typically seen as treatment/prevention of a disease (e.g. dietary advice to ensure better health and longevity). This suggests that relying on a narrow concept of the proper goals of medicine is misguided and risks inconsistency, *even if* we adopt a social constructivist approach.

Moreover, there is a problem with the normative assumptions behind the claim that medicalization going beyond the 'proper goals' of medicine is wrong. To say that the problem or cause of harm is commonplace provides little indication that we should not address it. It might have been commonplace for humans to get cold, yet we build houses and install heating systems; it might have been usual for people to die before 30, yet we welcome the chance to live longer if the quality of life is acceptable; it might have been usual for people to die of polio, yet we welcome the eradication of the disease and associated suffering (Harris, 2007) – the fact that an evil is usual, does not take away the permissibility of attempts to avert it, nor does the fact that a benefit is unusual make it morally impermissible to seek it. *Even if* we accept that commonplace problems should not be medicalized, this says nothing about the permissibility of addressing them. But if we want to address them, why not do it via medical means when those means are available and effective? In this context, Harris (2007) proposed that we extend the notion of the proper goals of medicine to making people better, broadly conceived. This may include 'making us better than well' and giving us what we desire insofar as this is compatible with morality, is lawful, and so on.

5.3.3. Epistemic, pragmatic and moral justification of pain and harm

It can also be argued that some traits and experiences labelled as 'problems' in the process of medicalization serve an important epistemic, pragmatic or moral function and are therefore justified. For example, sadness may point to aspects of one's life that need improvement, and provide a motivation for change. Moreover, even if the undesirable situation cannot be ameliorated, there may be value in knowing what predicament one is in (an epistemic gain) and sadness and grief can be seen as ‘appropriate’ (morally and/or epistemically) reactions to a loss.

For example, Lehrer (2010) mentions a psychiatrist who was distressed to notice that in his effort to ameliorate depression he failed to distinguish patients whose problem has a mainly social component from those whose problem stems from physiological reasons. The psychiatrist changed his mind when a patient, asked about the effectiveness of medication answered: ‘Yes, they’re working great . . . I feel so much better. But I’m still married to the same alcoholic son of a bitch. It’s just now he’s tolerable’ (p. 42).

Lehrer points out that because the woman’s problem was rooted in her relationship with her alcoholic husband rather than in her dysfunctional body, it was a mistake to treat her. Parens agrees that construing her normal human unhappiness as depression would be, as he calls it, a distressingly bad form of medicalization. If we put aside the practical and conceptual problem of how effective for wellbeing and happiness such intervention is (we can reach different answers referring to different theories of happiness), the separate issue that arises here is that of the epistemic worth of the experience and authenticity. As Parens put it,

> ‘[n]o matter how much the medication might attenuate her suffering, that could not justify her becoming complicit in cutting herself off from an important feature of her life as it truly was.’ (2011, p. 4).

I think that this is an important consideration in evaluating the choice of the ways and means we use to better our lives. This is not, however a problem that is restricted to medical solutions. Marx has famously called religion, ‘opium of the people,’ arguing:

> ‘Religious suffering is, at one and the same time, the expression of real suffering and a protest against real suffering. Religion is the sigh of the oppressed creature, the heart of a heartless world,

> and the soul of soulless conditions. It is the opium of the people.’ (Marx, 1843)

Karl Marx was not alone in seeing the function of religion as problematic. Similar views were shared by thinkers as diverse as the early romantic poet philosopher Novalis (1997), poet and essayist Heinrich Heine (1840), and priest and professor Charles Kingsley (Selsam, 1963). I do not attempt here to outline the various functions that religion can play (see: Durkheim, 1915; Alpert, 1938) nor to reflect on the moral, political, social and epistemic issues that arise when discussing religion, but rather suggest that various ways of construing the meaning of human experience, of looking at and remedying human problems, may lead to the same practical or epistemic detachment from the reality of the lived experience. A doubt similar to that which Parens’ raised about the use of medicine and Marx about religion can be also raised about the increasingly popular (for its stress-reducing effects) practice of meditation, and all kinds of other practices that would make us feel better by avoiding facing the problems that most affect our well-being. These include the tendency to get involved in work to escape family troubles, writing poetry to alleviate, rather than examine, one's existential pains and sorrows and humour – where it makes the unbearable bearable. Perhaps then, Keats would inspire Conrad’s attack when he states in *Sleep and Poetry* that ‘the great end / Of poesy’ is ‘that it should be a friend / To sooth the cares, and lift the thoughts of man’ (Keats, 1816, p. 68).

I am not arguing that the questions about alienation (vague as this term may be) from our situation should not be raised. In fact (although somewhat beside the point), I am sympathetic to many of the worries that the psychiatrists in Lurie (2010) raise when they point out the possible detrimental effects of the use of pharmacological mood enhancement. However, the main issue is not that the solution is medical, nor that it happens within the ‘medical gaze’ as critics of medicalization such as Conrad would have us think. Rather, legitimate doubts can be raised about various ways of modulating our emotions and memories and making ourselves feel better in general. If we are too much like the subjects willingly plugging ourselves into the available equivalents of Nozick’s (1974) experience machine, the locus of the problem is not in the *pharmacopeia* we might have available to achieve that goal, or whether or not that *pharmacopeia* is only available on prescription.

5.3.4. The ‘full human life involves suffering’ argument

Regardless of the direct epistemic value of letting us know where we stand, unpleasant experiences can also be regarded as a necessary and valuable part of a full human life. As problems are medicalized, a critic might say, they are construed as pathological and in need of fixing, while some kinds of suffering should not be fixed.

In his critique of the process of medicalization, Illich (1975) argued that medicalization is associated with iatrogenesis, in which the problems created by medicine are worse than the solutions it offers to the original condition. Cultural or structural iatrogenesis may happen when the medical view of, for example, pain, birth and death changes the meaning that those experiences have to people. According to Illich (1975) the meaning and the experience of suffering goes beyond the mere occurrence of physiological pain, and the attitude and meaning we give it makes a difference to how we live our lives. He further argues that medicalization leads people to forget about accepting suffering as an inevitable part of their conscious coping with reality, and instead learn to interpret every ache as an indicator of the need for ‘padding or pampering.’ Meanwhile, the signs and experience of suffering were traditionally seen as signals with a function of eliciting a response from an agent. Thus, Illich sees medicalization as the process of detaching pain from its cultural context (and thus meaning) and aiming to annihilate it.

It is easy to agree with Illich (1975) that the easier it is to make the pain just go away, the more temptation there is to alleviate the pain and ignore its cause, thus potentially bypassing the motivational and epistemic value of pain. On the other hand, there are many instances of unnecessary pain (in the sense of not serving any epistemic or motivational functions, or where the benefit brought by those functions is outweighed by harms) and alleviation of it would be appropriate. For example, it seems unnecessary to be dying in pain if one can die calmly and without pain; after the function the a pain of a broken leg has been fulfilled, it seems unnecessary to continue to be in a strong pain. Illich's argument holds only to the extent that we forget about the epistemic and motivational functions of pain.

An account inspired by existentialists may see anxiety as valuable because the experience of anxiety disrupts immersion in the usual projects and identity-creating roles, and is a part of the experience of oneself as a moral agent, responsible for one's decisions (Moran, 2000). Klerman (1972) made a more general point that Western culture developed under the influence of Ancient Greek and Christian traditions, which have assigned value to the suffering that comes with human problems. In a similar vein, Parens (2011) cites Shenk's (2005) account of how Abraham Lincoln's 'melancholia' was not just a huge burden, but also a crucial ingredient in his great life, this being but one recent example of that view. Thus, one may say that suffering may be seen as valuable and necessary for a full human existence and pharmacological interference with it as impairing our ability to flourish as people.

The critics of this approach sometimes call it 'pharmacological Calvinism.' The phrase was first used in the 1970s by Klerman, who thought that '[i]mplicit in the theory of therapeutic change is the philosophy of personal growth, basically a secular view of salvation through good works'. He describes 'pharmacological Calvinists' underlying intuition to be that

> 'if a drug makes you feel good, it not only represents a secondary form of salvation but somehow it is morally wrong and the user is likely to suffer retribution with either dependence, liver damage … or some other form of medical-theological damnation.' (1972, p. 3)

Parens notes that '[i]nsofar as those traditions celebrated suffering for which there were no medical remedies, Klerman must be right that at least to some extent those traditions made a virtue of necessity' (p. 5). Parens reformulates Klerman's thought:

> ‘If pharmacological and psychotherapeutic means can both achieve the same end – improving how one experiences herself and the world – then it is irrational and perhaps inhumane to prefer the more strenuous and expensive means. It’s irrational not to take a shortcut when improving human well-being is the destination. We should be slower to imagine that suffering leads to growth and understanding, and quicker to remember that sometimes it just crushes human souls.’ (Parens, 2011, p. 5)

One interpretation of Klerman’s may point to the value of a process (whether or not it involves suffering) rather than the outcome and the value we might ascribe to effort and struggle. Parens suggests that Kleman’s view ignores the moments in which we would think that suffering is a crucial element in a good human life and gives an example of grieving after a loss of a loved one. He suggests that such suffering should be endured rather than erased. This points to the fact that not all ways of improving wellbeing are good in the same way, and we may have reasons not to choose a ‘shortcut’ to wellbeing.

Although Parens’s example is an intuitively appealing counterexample to Klerman’s view, we should not let the intuitive appeal get the better of us. The appeal of the example stems from several sources, and I would propose that we question the intuitions to which Parens is appealing. The problem, I think, is not with the medical means of change but rather stems from the fact that we value a certain engagement with the world in which our feelings both express and reflect our situation and what we find valuable. The loss of a person we loved rightly evokes grief which we would be justified in not wanting to immediately remedy. However, the argument would equally apply if we decided to put ourselves though a two week course in psychotherapy, one that would alleviate our grief via attenuating the emotions, so that they would correspond to the emotions felt when losing a favourite umbrella. Parens would be justified in raising exactly the same objection and we may justifiably question whether immediate attenuation of grief amounts to ‘improving how one experiences herself and the world.’ (Parens, 2011,

p. 5) However, the problem has little to do with the exact means of medicalization, nor with medical solutions specifically.

Yet, there is another way in which Klerman's thought can be understood: he can be seen as criticising the view which focuses on the value of the effort, instead of the desired outcome. Several commentators emphasised the value of effort in the pursuit of excellence, and highlighted that medicalizing the problem means looking for a technological fix – which, even if pragmatically possible, would not be morally desirable (for discussion see: Cole-Turner, 1998; President's Council on Bioethics, 2003, p. 289; Fox, 2005; Olsen, 2006; Schermer, 2008; Goodman, 2010). Since the scope for improvement and effort will remain, even if we find technological fixes for many problems and shortcuts to wellbeing, the objection is weak if it is trying to establish that biomedical enhancements are morally impermissible or inherently morally suspect because they take away the chance of morally valuable effort.

On the other hand, it correctly highlights the fact that we may not always have a reason to take shortcuts, since what we may value about something is the activity or the process. When solving brain-teasers we might prefer to 'figure it out' on our own, because we value the activity and process over finding an answer that we could easily find online. Similarly, there is a reason why we may prefer a bike tour over a flight, even if the end destination is the same. We might also endorse the suffering of grief, and find a certain degree of existential doubt and anxiety as reflecting the human condition. However, it appears that it is not 'the effort' or some other similar disconnected property undermined by biomedicine that we value, but rather the pleasure of the process, the importance of the journey, the character-shaping or skill-developing consequences of the effort. Similarly, experiencing pain may be valuable when it stimulates us to come to grips with our situation, or for its role in expressing and reflecting what we find important. The extent to which 'effort' or pain are necessary to achieve the things we find valuable, however, is contingent on circumstances and not valuable for its own sake.

Experiencing poverty, hunger, illness or loneliness may open to us an understanding of others who are in a similarly dire position, an understanding that we would not otherwise have. It might also allow us to appreciate what we have got to a greater extent, and help us to learn ways of dealing with ourselves and the world that we otherwise would not learn. It might also stimulate social action. Yet, the importance of apparently undesirable states and experiences for a full and flourishing life should be scrutinised; since those experiences are considered to be necessary for fully experiencing life, they might be subject to the *status quo* bias served to us under the guise of usefulness and value.

In his 2011 paper, Davies argues that the Western narrative of suffering has broadly shifted in recent decades to generally favour a negative over a positive conception of suffering. He also argues that this shift has been hastened by what he calls 'rationalization of suffering,' by which he refers to the development of biomedical means of addressing suffering. Insofar as medicalization impairs our ability to understand the causes and context of our ills and find a fitting remedy (e.g. if we propose that 'work stress' should be remedied by anxiolytic drugs where anxiety is due to lack of basic financial security traceable to social and economic relations) or prevents us from using better solutions (e.g. drugs are a first-line remedy for mild depression even if psychotherapy, social contact and exercise are all-things-considered more cost-effective), I agree with the doubts raised by the critics. However, I would rather worry about another 'rationalization of suffering' that may afflict us: the one in which we justify the value of the bad in a self-defeating feat of rationalization. Carl Sagan in *The Demon-Haunted World: Science as a Candle in the Dark writes that:*

> 'One of the saddest lessons of history is this: If we've been bamboozled long enough, we tend to reject any evidence of the bamboozle. We're no longer interested in finding out the truth. The bamboozle has captured us. It's simply too painful to acknowledge, even to ourselves, that we've been taken.'
> (Sagan, 1996, p. 230)

The value of suffering *qua* suffering is a prime candidate for being described as a claim that has bamboozled people for years. We are better off facilitating the acquisition of the same gains without the associated pain when possible, asking the question of whether the negative experiences and effort are worth the pain, and whether the pain comes with sufficient gain to justify it.

Karney Osborne, in a poem written about the effects of a volcano eruption which destroyed the Montserrat Island's capital, killed 19 people and disrupted the lives of the islands inhabitants (about 4,000 of whom still remain 'in exile' [Romeo, 2012]), wrote about the beauty of the landscape:

> She was my pride and joy
> —I will lift mine eyes unto the hills
> Emerald Isle, green hills, green fields…
> …The sun shines again and again
> No trees, no grass
> my love is naked and bare.
> Now I see her as she really is; strong, proud, defying
> All giving me strength to fight, to survive.
> Now she is beautiful.
> There is beauty too in nakedness.
> (Osborne, 1966)

Would the narrator choose for the eruption to happen, if she had the control over it? I would hope that no advocates of the value of effort and pain would promote living through the calamities of war or concentration camps so that we learn from the experience, have an opportunity to endure pain and realise the fragility of life. Humans can make the most of and learn from even extreme hardship, and sometimes very valuable lessons come from the most difficult experiences (retrospectively), but it often is not a sufficient reason for (prospectively) preferring to have that experience over not having it. Where effort or hardship is valuable, we might prefer to have the ability to choose to experience it, rather than be forced to.

5.3.5. Living well requires that we let some problems be.

Another ethical worry raised about medicalization is that the obsessive focus on averting small problems is self-defeating and we should accept some of the problems as a part of human existence. For example, Barsky and Boros argue that empirical studies suggest that

> ‘people are increasingly bothered by, aware of and disabled by distress and discomforts that in the past were deemed less important and less worthy of attention’ (1995, p. 1932).

The authors note that social and cultural forces can lead individuals to amplify pre-existing physical discomforts, misattribute them to disease and seek medical help. Thus, they see somatization (defined as the propensity to experience and report somatic symptoms that have no pathophysiological explanation [Woolfolk and Allen, 2007]) and medicalization are mutually reinforcing processes.

Medicalization of everyday and usual problems is on this account seen as troubling because it makes us attend to them more, and increases the suffering by making the problem more silent. In the same way that focusing on every single imperfection of one’s body makes one increasingly notice and assign weight to imperfections that would otherwise not be particularly troubling, medicalization may construct issues that previously were seen as usual (even if imperfect) features of life as problems to be fixed. In addition to increasing the weight of pre-existing problems, medicalization might lead one to make a previously unnoticed or unimportant feature or experience acquire a negative meaning, resulting in a proliferation of defects (Bordo, 1998). Susan Bordo describes how after a visit to dentist (motivated by a need to address an altogether different matter) her perception shifted:

> ‘the gumminess of my own smile was of no concern to me until after I had seen the dentist; but under his care I began to wonder if it wasn’t in actuality something I’d better hide… or ‘correct.’ (…) If you are trained to see defect, you will.’ (Bordo, p.213).

The argument may go as follows: when previously acceptable traits and experiences become unacceptable we are faced with a situation in which, although more and more imperfections are ameliorated and despite our increased capabilities, our subjective wellbeing falls. I think that the worry raised by Bordo (1998) is *prima facie* plausible, insofar as medicalization involves construing something as a pathology, presumably at both an individual and social level. However, the issue raised by Bordo (1998) is also a matter of values and culture and not only medicalization. For example, Francis examined the stigma experienced by middle-class parents of children with physical, psychological and behavioural problems and highlighted the importance of also considering 'larger contexts of an anxious, intensive parenting culture' (2012, p. 927). What is a reasonable response? The focus here, I think, should not be on the medical or biomedical toolkit, but rather on the cultural norms that may foster wellbeing or ignite stigma. The considerations raised by Bordo (1998) could also provide reasons to abandon the discourse of 'pathology' and focus on the discourse of 'improvement,' as well as reevaluate the value of ends that the medical means might be used to achieve.

A related problem raised by Bordo (1998) has to do with the creation of new desires and the cultural norms and context medicine is embedded in. This problem is not a specific effect of medicalization but rather refers to the 'creation of desires,' which could equally be an effect of advertisers wanting to sell us their products, professionals wanting to sell us their services and more attention directed to the possibilities afforded by a new technology. Where the creation of new desires is unendorsed by the agent and happens without an easy ability for the agent to be able to engage with the influence, the issue merits our attention (For more discussion see: Arrington, 1982; Crisp, 1987; Phillips, 1994; Dow, 2013). Luckily, the fact that we could do something does not mean that we should do it, and that something is available does not mean that it is valuable.

5.3.6. Social control: are we moulding ourselves to fit society instead of adapting the society to fit people's needs?

An important related issue is that social factors contributing to problems are downplayed in comparison to individual biological and psychological factors. The increased pressure to perform and to keep pace with society's increasing demands can be seen as an under-recognized part of the problem in adult ADHD (Schermer, 2007). The challenges that individuals encounter are framed as individual rather than social problems – it is the person that does not perform well enough rather than society in need of change. One of the critiques of the use of ADHD diagnosis is that when the over-stretched educational system cannot adapt to children's needs, children who are a problem for the system land in doctor's offices and often get a prescription for stimulants (Graham, 2007). The argument here may be that we are putting the cart before the horse, forgetting that it is individuals who matter, not the abstracted idea of the society.

The more direct negative consequence is that medicalisation may distract attention and direct resources away from changing the social structures and expectations that can produce suffering in the first place. Perhaps rather than changing the bodies of shy people with drugs, we should change our expectations of how people behave in novel situations (Parens, 2011) and create environments that are facilitative of people with different skills, behaviours and strengths. The tension between attempts to change an individual and to adapt the society is clearly visible in the recent history of conceptions of disability and the disability movements. On the one hand, people with movement disabilities have greater access to rehabilitation than ever before, stem cell scientists are working to create a way of addressing spine injury and new prosthesis and wheelchairs have been adapted to a variety of activities. On the other hand, patient and 'users' groups fought for improvements in how accessible the environment is for people with movement problems (public buildings and many leisure facilities became wheelchair-accessible for example). Each of those stances came with an associated view of disability (social vs. medical) which, although not necessarily mutually exclusive, are nevertheless associated with different practical approaches.

This argument, I think, carries some weight. However, it still leaves us with the issue of how to address the outlined problems. In the context of ADHD, we might wonder whether it would be better for those children who struggle in the imperfect educational system to continue without the pharmacological aid. Provided that the pharmacological solutions are safe and effective enough, it seems to me that pending the needed social action, we have a good reason to provide the pharmacological remedies. If we consider the problem of ‘the wrong focus’ in isolation from the possible side-effects of the medication, I think that the objection is too weak to ground a robust critique of medicalization.

The critic of medicalization could see many of the problems that medicine seeks to remedy as stemming from the bad organisation of our societies. For example, the rise in obesity in Western countries may be seen to be a result, for example, of an increase of physically passive jobs and the increased availability of cheap, calorie-rich food served in big portions, along with the strong lobbying power of the food industry. A gastric band surgery, gastric bypass surgery or an effective anti-obesity pill might ameliorate some of the negative effects but does not address the root cause of the problem. It could be argued that the prevalence of mental illness, including depression, can be seen to be a direct or a down-the-line result of living in a modern capitalist system, with its economic ups and downs, uncertainties and pressures, the necessary economic relocations which fracture social bonds, and the lifestyle of a modern worker which is not conducive to wellbeing (e.g. Link and Phelan, 1995). Medicine (and related science) can be seen as a part of the social order, reinforcing it by smoothing the roughest edges or at least giving hope that something can be done. Medicalization is ultimately seen as a mark of social control in a pejorative sense of the phrase – *de facto* creating effective workers and dissolving dissent. To propose symptomatic solutions is to leave the cause intact which perpetuates and even reinforces the bad underlying dynamics.

An analogical argument can be made specifically in relation to the moral domain and moral education. Many education scholars notice a general move from supporting the teaching of subjects that were typically conveyed knowledge, values, beliefs and skills necessary for developing moral agency, such as art, literature, religious studies towards making schooling and the curriculum serve economic purposes. On the other hand, the existence and smooth functioning of the society and organisations depends on citizens’ ability to display a number of moral and civic virtues and behaviours, including responsibility, respecting the interests of other persons, not harming others, tolerance, etc. A critic of medicalization might point out that the medical conditions labels used to describe ‘disruptive behaviour’ such as oppositional defiant disorder and conduct disorder, detract from the causes and appropriate solutions.

For example, Oppositional Defiant Disorder (ODD) is characterized by the frequent occurrence of at least four of the following behaviors: losing temper, arguing with adults, actively defying or refusing to comply with the requests or rules of adults, deliberately doing things that will annoy other people, blaming others for his or her own mistakes or misbehavior, being touchy or easily annoyed by others, being angry and resentful, or being spiteful or vindictive. ‘Defiant behaviors’ may include persistent stubbornness, resistance to directions, and unwillingness to compromise, give in, or negotiate with adults or peers and the deliberate or persistent testing of limits, usually by ignoring orders, arguing, and failing to accept blame for misdeeds. Instead of encouraging schools to teach impulse control, emotion management skills, problem solving, taking one’s stance respectfully, as well as reasoning through disagreement and ethical issues (much of which is included in the ‘character education’ approach to moral education), children who do not conform to the prescribed ways of behaviour enforced by economically strained educational system, are given ADHD medication (stimulants), atypical antipsychotics, antidepressants or tranquilizers.

Let us assume that the above view is correct – that the organisation of modern capitalist societies is the cause of many of the problems that medicine addresses. In that case, it appears to me that medicalization can serve both a positive the negative role in addressing the problems at the societal level. For example, only recently the limitations of SSRI treatment come more into focus (Sansone and Sansone, 2010; for a resent large study see: Read et al. 2014; for discussion in the context of MBs see: Wiseman, 2014; Hyman, 2014). As those limitations are better known and as the research on the health impacts of the factors outlined in the previous paragraph by Link and Phelan (1995) becomes more robust, one could envisage a change in the focus of medicine itself.

Some of this trend is already visible in the discussions about effective public health measures, as well as the ‘trickling down’ of the effects of those investigations – for example, it is not uncommon for doctors ‘to prescribe’ exercise. Medical professionals may promote as well as impair the adaption of effective non-pharmacological measures which at times may provide a counterweight to the strongly represented interests of, for example, pharmacological companies or the food industry. Insofar as political solutions go, medical professionals may be a supporting or inhibiting force for the needed changes. To a large extent, the strength of the argument hinges on empirical claims about the role of medical professionals, and this role may differ depending on the country, and issue and is not immutable. Where the medical focus leads to misallocation of resources, perhaps it is a matter of changing medicine's focus and workings (e.g. by encouraging more referrals to psychotherapists in the case of depression), rather than pushing for de-medicalization – although the conclusions will need some empirical support and are best considered on a case-by-case basis.

5.3.7. Exponential growth of costs and the undermining of non-medical coping arguments

As medicalization expands the category of what warrants medical treatment, the cost of medical treatment grows exponentially, which makes it increasingly harder for any government to pay for medical care for all (Conrad et al. 2010). Moreover, critics of medicalization may worry that the indirect costs of side-effects may have to be added to the costs of medical services. Medicalizing what can otherwise be seen as moral failures, increases that cost further.

As it stands, this objection is weak. The claim that it is better not to develop medical means of addressing problems because it will increase the costs can be responded to easily: we can choose to pay or not to pay for additional services depending on what we can afford and are willing to pay for healthcare as opposed to other government-subsidised or privately purchased goods. In relation to moral modification, there still will be a choice of achieving set objectives through traditional means like moral education (also not always cheap: for example moral education at a university level is quite expensive) as well as medical means, such tools derived from psychotherapeutic approaches and drugs. We can choose to use available services or not, depending on whether it is worth it.

Rationing has been done for a very long time (as the care can always be better) and although looking at problems as medical may spark a pragmatic or moral imperative to try to address them, we can choose our priorities both on an institutional and on an individual level. Similarly, the indirect costs may not be obvious or known at the start, but as the intervention is applied and we get a better idea of the costs and benefits, we might re-visit the cost-benefit analysis and factor those costs in. Moreover, as the creation of tools to deal with problems in medical ways progresses, generally speaking, the cost of remedies is likely to fall. While new remedies are likely to be created and not be affordable, the old ones will become more affordable and cost-effective.

Although this view of the process is a generalisation, and thus subject to exceptions and problems (e.g. medical research and development targets the maladies of rich countries) as well as needing certain conditions to occur (e.g. preventing companies from only providing the most expensive treatments while withdrawing the cheaper but more cost-effective when needed, while providing a sufficient economic incentive for research and development to happen – e.g. through a suitable patent system), the subsequent increases in welfare justify the trouble. Even when we might not be able to afford all care for everyone, it would be unjustifiably perfectionist to only develop remedies that we can currently provide to all. Thus, the argument about the rising costs of healthcare is strongest as an argument for appropriate prioritisation, management and development, rather than against the process altogether.

However, the argument can be supplemented with reference to the *change of attitudes* which will influence the consequences of medicalization, choices about healthcare provision and judgements about outcomes of successful healthcare delivery. In *Medical Nemesis*, Illich (1975) describes what he calls social and cultural *iatrogenesis*. Illich attributed medicalization ‘to the increasing professionalization and bureaucratization of medical institutions associated with industrialization’ (Gabe et al. 2004, p. 61). He argues that although healthcare consumes an ever growing proportion of the national budget, the benefits to the patients and society are increasingly unclear. The more people are exposed to healthcare, the sicker they can feel. Accordingly, medicalizing the problems can play into their perpetuation.

One of the ways in which this happens, assording to Illich, is via social iatrogenesis, which refers to harms to health that are due to the socio-economic transformations which have been made attractive, possible, or necessary by the institutional shape health care has taken. Illich argues that:

> ‘[the] medical bureaucracy creates ill-health by increasing stress, by multiplying disabling dependence, by generating new painful needs, by lowering the levels of tolerance for discomfort or pain, by reducing the leeway that people are want to concede to an individual when he suffers, and by abolishing even the right to self-care. Social iatrogenesis is at work when health care is turned into a standardized item, a staple; when all suffering is “hospitalized” and homes become inhospitable to birth, sickness, and death; when the language in which people could experience their bodies is turned into bureaucratic gobbledegook; or when suffering, mourning, and healing outside the patient role are labelled a form of deviance.’ (1975, pp. 14-15)

Illich (1975) argues that we should be concerned with the erosion of already developed ways of dealing with pain, sickness and death. For example, as the mental health field promotes its technologies as necessary interventions in almost all areas of life, what people pick up is that they are not expected to cope through their own resources and networks – and non-medical ways of enduring and coping wither away. The same argument can be applied to medicalizing the imperfections in capacity for moral reasoning and action. Community’s ways of regulating morally relevant conduct and developing abilities that contribute to the capacity for moral agency such as impulse control, empathy, reasoning about moral issues, caring about other people’s interests, reflective helping, emotion regulation, etc., can be undercut by the reliance on medicine. The idea here is that once a problem is constructed as medical, it is ever harder to reconcile and cope with the everyday issues. As a consequence, a vicious circle is created: the more resources are provided for mental health services, more are perceived to be needed and health provision becomes a part of the problem. Medicalization can be seen as creating the same issues that medicine subsequently has to deal with, thus creating a vicious circle of need.

Several doubts can be raised about this account. Firstly, this view depends on the empirical facts about the impairing effects of the presence of medicine on the ability of individuals to cope. For example, it might be that those who can cope via their own resources will continue to utilize those resources while those who lack the social networks might benefit from medicine's support (e.g. older people whose children live elsewhere and who do not have an extensive social network). Secondly, one could question whether medicine causes the problem or simply fills the void created by other social forces (medicine is not the cause of people's disrupted social networks for example, but increasing labour mobility may be). Third, it is not clear that it is just to leave agents without the opportunity to be assisted if they need assistance. For example, it is unclear why should the degree to which agents fare well (whether the issue concerns a good such as happiness, coping with adverse life events or capacity for moral agency) depend solely on the quality of their social networks and their self-developed ability to cope with or 'bear' suffering and illness. Fourth, the lowering level of tolerance for discomfort may be problematic, but may also be a result of the increased attention we pay to peoples' wellbeing, and thus constitute progress. For example, that there is a lower level of 'tolerance of discomfort or pain' resulting from sexual assault or domestic violence, and that such issues are now discussed more openly with the use of language that could be easily described as a '*bureaucratic gobbledegook*', could be seen to be a bad development on Illich's (1975) account. The position I just described is not, I presume, a position that Illich (1975) would endorse, yet it very well illustrates the problem with applying his and similar critiques. Finally, it is possible for all of those effects to be present simultaneously, making the evaluation of whether medicalization is a desirable or undesirable process a very murky evaluation.

Moreover, according to Illich (1975), clinical iatrogenesis involves serious side-effects which may be worse than the burden of the original condition. Clinical iatrogenesis include the harmful side effects of seemingly beneficial and advisable intervention, post-intervention complications, the negative effects of wrongly prescribed medication, bacterial resistance developed as a result of widespread use of antibiotics, hospital-acquired infections and harm resulting from negligent medical errors. Although some of the iatrogenic effects may be obvious, the burden of others may be difficult to calculate (e.g. the harmful effects of drug interactions that overlap with the progression of a disease or aging). One of the contributing reasons is the under-reporting of side-effects, as happened in the case of SSRIs, Rofecoxib (Avorn, 2012) and Lariam (Croft, 2007; Ritchie et al. 2013).

Finally, there may be a difficulty with detecting the full burden of side effects, including the causally related yet difficult to measure harms – such as the long-lasting impact on the social interaction of moderately depressed people taking SSRIs and the impact of a Caesarean section on early formation of the mother-infant attachment and its consequences. The cost-benefits analysis only makes sense on a case-by-case basis. Even when the cost-benefit analysis is performed, we have to be cognizant of the incompleteness of our view, the incompleteness that is reinforced and shaped by the 'intangibility' of some kinds of relevant side effects.[17] While indirect and intangible costs of a disease are often explicitly referred to, often the indirect and intangible costs of drug use are not mentioned.

It is important that direct, indirect and intangible costs of medical and non-medical solutions need to be compared, and accounted for to the extent possible. Intangible costs might sometimes give us a reason to choose a more expensive and perhaps non-medical solution to a problem, for example moral education over biomedical means of modifying empathetic ability. For example, if there was a cheap drug that would achieve an effect similar in this regard to a semester of moral education, we might see a drug as a cost effective solution. However, if it turned out that the drug affects adversely the ability to form and enjoy lasting relationships (as some reports

[17] 'Intangible costs' usually refer to costs that cannot be directly expressed in monetary values, such as happiness or anxiety due to a disease. 'Intangibility' is perhaps an unfortunate term, since it implies that there is no way of measuring the impacts of those factors. I do not mean to suggest that those side effects are impossible to measure or estimate in any way, but I will follow the term used in the literature. See: Leukefeld et al. (2011).

on the side effects of SSRIs suggest), the balance of costs and benefits might change. In this case, more time and resources consuming moral education might be a better solution.[18]

5.3.8. Is autonomy undermined by a shift in causal explanations?

Some points of Illich's (1978) critique may be restated as a worry about individual responsibility and autonomy. Critics of medicalization may argue that since a technical fix cannot solve problems when the locus of a problem is not in the body but in a particular life situation, we risk entrenching the problem by framing it in a way that suggests lack of responsibility. In the context of proposals for wider provision of CBT, Summerfield and Veale argue that 'once a psychiatric formulation is deployed as the explanation for a person's problems, the moral economy of the situation alters' (2008, p. 327) and that the shift in focus ultimately undermines autonomy. The focus on a diagnosed condition for which (it is implied) the patient is not responsible, and from which they are not expected to recover without professional help, means that the agency and an expectation of finding a remedy passes from the patient to the therapist.

Moreover, Read et al. noted that the reason why the general public prefers psychosocial explanations of mental illness may be that once a disease model is applied to the brain, something definitive and negative appears to have been said about the patient's core qualities: 'that the person is incapable of judgements, reason, autonomy' (2006, p. 327). Summerfield and Veale (2008) argue that this may affect the way people see themselves, results in them giving up sooner, being more likely to see themselves not as normally stressed but as 'suffering from a disorder', and in general playing out the role of a moral patient (a sufferer of involuntary circumstances) rather than an autonomous agent, the process reinforced by an unequal patient-medical professional power dynamic which further pushes patients into passivity.

The conceptual history of addiction is an example of the way attributions of responsibility for the problem and for the solution influences the way the problem is

[18] Naturally, non-medical solutions have their intangible costs too. The point here is that for the full cost-benefit analysis the effects encompassed by the notion of clinical iatrogenesis should be accounted for.

addressed. Brickman and colleagues' (1982) classification of the different theories of addiction may be of help in thinking about the way in which the conceptual background provides an environment which encourages some and discourages other approaches to a problem, impacting the milieu in which an individual acts.

	Attribution to self of responsibility for **solution**	
Attribution to self of responsibility for **problem**	High	Low
High	Moral model (High, High)	Enlightenment model (High, Low)
Low	Compensatory model (Low, High)	Medical model (Low, Low)

Table. Attribution of Responsibility in Four Models of Helping and Coping. *Source: Brickman et al. (1982)*

Many arguments concerning medicalization apply equally to medicalization of issues such as aging or menopause and moral modification, yet this critique seems to be especially important when we are talking about modification in the moral sphere. On the one hand, being a moral patient seems to be inimical to moral agency and responsibility for one's moral development and actions. The problem is amounted by the proposed involuntary application of moral modification in order to make moral agents (e.g. Savulescu and Persson, 2008), but even voluntary use, if indeed accompanied by the attitude of 'outsourcing responsibility and power' could come with negative impacts on moral agency and moral responsibility inherent in moral agency. On the other hand, the capacities underlying the ability to be a moral agent are capacities like any other and are in principle modifiable both by biomedical means and traditional means like practice or education. In fact, a similar doubt may apply to the body creating an conjunction of the stance of the autonomous agent and the moral patient of both our own and others' actions – while we are in part constituted by our body and the body is a vehicle of our agency, we can also be modifying it, thus taking a agential stance towards ourselves.

To the extent the medical model necessarily promotes the stance of passivity it would undermine moral agency, and the importance of this problem is exacerbated when we are talking about MB and aiming at creating better moral agents. However, it remains unclear whether medical model necessarily promotes passivity and undermines responsibility, or whether the tools that medicine offers can be used to empower individuals to take greater responsibility for capacities that they previously saw as not under voluntary control. To provide a counterweight for the discussion about potentially problematic aspects of medicalization, the next section outlines the benefits of medicalization of an issue.

5.4. Ethical appraisal: the benefits of medicalization

Parens (2011) argues that discussions of medicalization often rely upon a tacit distinction between medicalization (which is good) and over-medicalization (which is bad). He suggests that the cases of PTSD and Alzheimer's disease, which were once seen as non-medical problems but are now understood within a medical context, are examples of 'good medicalization.' Similarly, Carter and Hall (2012, p.231; also Burke, 2011) point out that some scholars use medicalization in a more positive sense to describe the increased use of effective medical treatments by those who were previously denied access, either for social reasons or because such treatments were not available.

What are the potential benefits in the process of medicalization? The critical appraisal of arguments against medicalization already yielded some indication of such benefits. Medicalization may allow patients to forge collective identities around shared experiences, facilitating advocacy efforts and improving recognition (Browne et al. 2004). It may also shift social perceptions away from a moralistic, punitive approach to deviance, thereby creating space for increased support and tolerance (Burke, 2011). On an individual level, medicalization may also legitimize an individual's struggles and lead to increased access to services and resources (Conrad and Potter, 2000; Conrad, 2007). Further, a medical diagnosis may afford access to social capital associated with 'the sick role' (Parsons, 1951).

To provide further counterweight to the critics' claims I will use the example of medicalization of addiction to briefly outline some of the benefits that the process of medicalization may bring. Firstly, and rather obviously, taking a medical perspective may lead to good outcomes following the development of a better solution to a pre-existing problem. I will not elaborate on this point in great detail, but medicalization may be desirable because all the potential and actual problems of medicalization may have less weight than the sometimes devastating results of addiction.

Secondly, the construction of a problem as at least in part medical rather than entirely moral may pave the way to easier change, be it via medical or other means. For example, drug addiction can be regarded as a ‘chronic, relapsing brain disease’ (Leshner, 1997; see also Volcow and Li, 2005) or a matter of personal responsibility. The former view has been echoed in the attitude of US National Institutes (NIDA and NIAAA) while the latter is held by many drug users and families (Bell et al. 2012). That addiction has become medicalized is evident in the widespread position that addiction is a ‘chronic, relapsing brain disease’ (Leshner, 1997), a claim informed by evidence of neurophysiological and neurochemical changes present in addiction (Volkow and Li, 2004). As the evidence of neurological changes that predispose drug users to subsequent use accumulates, the medical model has become prevalent, replacing the dominant perspective of the 20th century that individuals who use drugs were ‘autonomous, self-governing individuals who wilfully, knowingly, and voluntarily engaged in criminal and immoral behaviour’ (Carter and Hall, 2008, p. 81).

According to its proponents, the brain disease model of addiction leads to changes in social and public health policies, which will have the double benefit of providing more humane and ethical responses to addiction, as well as more effective solutions to addiction and related harms (Leshner, 1997; McLellan, et al., 2000; Volkow and Li, 2004; Dackis and O’Brien, 2005). Indeed, the rise of the medical model of addiction has already played a role in finding novel and beneficial approaches. A neurophysiological deficit-focus view of addiction led to the development of opioid substitution therapies, which have shown to be effective in reducing drug use. In fact, Bell et al. (2012) argue that the primary barrier to increasing the effectiveness of opioid substitution therapy is residual vestiges of a non-medical, moralizing approach to drug use evinced by healthcare practitioners working in inpatient treatment facilities. Furthermore, opioid substitution therapy has its analogue in approaches to tobacco smoking cessation such as nicotine replacement therapy, which similarly depends on a medicalized, biological dependency perspective on cigarette smoking (Gartner and Partridge, 2012). Thus, the medical approach comes with an epistemic stance that facilitates the creation of certain kinds of solutions that have been shown to increase the chances of kicking the addiction or alleviating its harmful consequences.

Even where those approaches overemphasise the impact of physiological changes, such as changes in the reward circuitry in the brain, and unjustifiably underemphasise the impact of social factors or the strength of habit, the approach has generated previously unheard of solutions to addiction. Indeed, as the limitations of the 'brain disease model' start to come to the fore, the social, motivational and habitual influences on maintaining addiction enjoy greater attention and lead to an integration of pharmacological approaches with other remedies, such as counselling (National Institute on Drug Abuse, 2012).

Moreover, abandoning the view of addiction as a matter of moral strength or weakness may decrease blame and facilitate the search for solutions. Firstly, it is sometimes argued that addiction neuroscience encourages individuals to seek treatment or empowers them to make choices not to use drugs (Condit et al., 2006; Carter and Hall, 2012). Secondly, the existence of an authoritative scientific explanation of addicted individuals' experiences might increase their willingness to engage in medical treatment (Hall et al., 2008). There is some empirical support for this view. Gartner and Partridge (2012) point out that 'smokers who attribute a failure to quit to unchangeable intrinsic factors such as personal characteristics have lower personal quitting intentions and lower quitting self-efficacy' (p.79). Similarly, many patients who received a mental health diagnosis describe the sense of relief coming with decreased self-blame associated with the attribution of responsibility shift, as well as with a hope for a solution. Paradoxically, although medicalization is often criticised as labelling and stigmatizing struggling individuals, it may also lead to de-stigmatisation. Thus, the change in attribution of responsibility coming with a medical model has the potential to be either pragmatically harmful or beneficial, depending on whether it increases or decreases effective coping.

Additionally, models of dependency associated with medicalization give rise to a host of other approaches that recognize the biological basis of addiction. Those approaches tend not to focus upon abstinence, but upon reducing the risk of harms unnecessarily associated with drug use. These approaches include needle-exchange programs and medically-supervised injection sites, both of which are shown to reduce infections common amongst injection drug users forced to share needles or inject hurriedly to avoid detection. Those solutions, however, commonly face protests rooted in a view that the government should not allow drug users to act illegally and immorally – a perspective commonly associated with Brickman et al.'s (1982) moral model of addiction. Here, the medical model may facilitate adopting social policies that prevent a great amount of unnecessary harm.

5.5. Conclusions

Evaluating criticisms of medicalization poses a difficulty because the empirical data needed to evaluate the empirical aspects of ethical arguments is often missing or murky. Both medicalization's critics as well as supporters make empirically unsubstantiated claims and over-generalisations (see: Volkow and Li [2004] for often cited claims of the benefits of medicalization, some of which some go far beyond what is empirically justified). Moreover, the wider societal effects that worry many sociological critics are often difficult to evaluate with rigour.

The aim of this chapter was to question the implicit and often negative normative attitudes towards medicalization; to introduce the descriptive concept of medicalization as a process that, pending a moral case-by-case assessment, should be seen as normatively neutral; explore some common worries about the process of medicalization and to disentangle the various normally intermingled threads running through the arguments. I have proposed that the process of medicalization in its current incarnation is wider than that of increasing medical practitioners' power, and happens in the context of scientific and technological developments – and increasingly within a market economy. The view of problems as 'medical' is not restricted to the medical profession, but can be understood as a framework used by medical practitioners, scientists, policy makers and the members of society at large. The construction of problems as medical makes some approaches to dealing with them more likely than others and comes with benefits and perils.

The current developments in the neuroscience of morality and the proposals of MB via direct emotion modulation stem from a scientific approach to morality which merges easily with the medical approach to problems. The prospect of, and proposals to pharmacologically modulate romantic and parental love and attachment, pro-sociality, the underpinnings of moral judgement and behaviour with oxytocin or serotonin, together with the medical diagnoses and assessment tools for conditions such as psychopathy, social anxiety and post-partum depression, paint a picture of an increased pace in the medicalization of sociality and morality. It is important that we attend both to the promise and the limitations of the medical-scientific view of the social and moral aspects of our lives, giving due weight, however, to both the advantages and disadvantages of medicalization, and attending to the way societal values shape the exact results of this process.

CHAPTER 6: Narrative identity

6.1. Introduction

One of the objections raised against enhancement technologies is that they might change ‘who we are’. This worry might be explicated within the context of authenticity and identity – this chapter focuses on examining the latter. While there is a rich philosophical literature related to identity, the current bioethical debate largely draws on the analytic philosophy tradition, which focused on the question of numerical identity.

Following other scholars such as Schechtman (1996, 2009), DeGrazia (2005) introduces a distinction between numerical and narrative identity. According to him, while numerical identity refers to the criteria that determine whether a being at one time and a being at another time are, despite change, one and the same being, narrative identity relates to the question of what is most central and salient in a given person’s self-conception. In his 2005 paper on enhancement and identity, DeGrazia argues that much of the worry about creating a ‘new person’ may derive from conflating the two senses of identity: ‘If taking an SSRI changes your personality in important ways, *you* will change; it’s not the case that you would literally be destroyed and replaced with another person, as would occur if numerical identity were disrupted’ (DeGrazia, 2005, p. 269).

While DeGrazia’s appraisal of the bioethical debate as often conflating the two senses of identity is convincing, he uses a very wide notion of ‘narrative identity’. He goes on to examine the ‘narrative identity’ arguments against enhancement with only a sparse and token reference to relevant theories of narrative identity. This is not unusual; with the exception of a handful of papers (e.g. Schechtman 2009, Baylis 2013), the narrative identity approach has been mentioned but not fully integrated into the bioethical debate about enhancement. The situation might be characterised as not even ‘talking past each other’ but rather ‘talking in one’s own corner.’ This chapter aims to address this disconnection.

Narrative accounts of identity suggest that the sense of who we are is created via autobiographical self-narratives, which are seen as a means to give meaning to events, behaviours, desires, intentions, etc. in the context of one's life. One objection that can be raised against the ethical permissibility of MB is that such changes cannot be incorporated into a self-constituting narrative and, thus, that MB threatens to undermine narrative identity. I will argue that the objection is weak. The chapter proceeds in three steps.

First, I examine the argument brought forward by Martya Schachtman's (2009) who outlines the possible issues related to narrative identity raised by DBS and assess whether such arguments can provide a basis for a moral argument against the permissibility of DBS generally and MB via direct emotion modulation specifically. After briefly outlining relevant features of Schechtman's narrative identity account, I draw paralells between DBS as discussed in Schechtman's 2009 paper and MB. I argue that Schechtman's (2009) argument fails to ground an ethical objection to DBS on her own account because the *articulation constraint* could be satisfied in cases of emotion modulation via DBS. Moreover, it is unclear that biomedically undermined identity-narratives would be irreparable. In the second part of the section, I examine the empathetic access condition for self narratives and argue that it should be rejected as too demanding, given that it unjustifiably focuses on one backwards-looking attitude.

Second, I describe and evaluate Paul Ricoeur's account of narrative identity and argue that his theory provides overly stringent criteria for narrative identity. Moreover, the example of Ricoeur's theory is illustrative of the problem with applying narrative identity approaches to evaluate the moral permissibility of biomedical interventions.

Third, I argue against the strong ethical narrative thesis, according to which a consistent narrative is necessary for, or highly conducive to, a full and flourishing life. I argue that although narrative identity might be an interesting and fruitful way of looking at personal identity, we should accept that it is not necessary for a full and flourishing life. I conclude that although narrative identity theories can provide an interesting insight into potential issues raised by direct emotion modulation and

MB, such theories fail to provide a strong basis for a robust ethical objection to biomedically modifying our moral sphere.

6.2. Schechtman's objection to the use of direct brain modulation

6.2.1. Schechtman's account of narrative identity

An influential account of narrative identity, and one that has been explicitly applied to the assessment of DBS, is that of Schechtman. Schechtman (1996) makes a distinction between a re-identification question (what makes someone the same person over time) and characterisation question (what it is to be a particular person). She argues that while psychological continuity theorists such as Parfit (1984) focus on the former, the latter also merits consideration. According to Schechtman:

> '[I]ndividuals constitute themselves as persons by coming to think of themselves as persisting subjects who have had experience in the past and will continue to have experience in the future, taking certain experiences as theirs… A person's identity … is constituted by the content of her self-narrative, and the traits, actions, and experiences included in it are, by virtue of that inclusion, hers.' (Schechtman, 1996, p. 94)

The characterisation question deals with the 'set of characteristics that make a person who she is' and is relevant to discussing identity crises, which Schechtman understands occur when a psychological state (or combination of states) do not cohere with a subject's total psychology. Consequently, in an identity crisis, a subject is unable to integrate such a state as a comprehensible part of his life and to accept it to be his own. The approach resembles concerns about authenticity (asking, 'Is the life that I am living my own in a relevant sense?'). However, in the narrative account of identity the focus is on whether a subject 'creates their identity by forming an autobiographical narrative — a story of his life' (Schechtman, 1996, p. 113).

Schechtman raises the question of the criteria required for narratives to confer identity, since not all narratives can be identity-constituting. According to Schechtman, for a self-narrative to be identity-constituting, it must satisfy two constraints. First is the articulation constraint (the person must be able to provide some account of her history, her life situation, and her motivations) and another is the reality constraint (the self-narrative must be coherent with basic facts about how the world is).

6.2.2. Schechtman's objection to biomedical modification

In her paper about DBS and narrative identity, Schechtman (2009) considers the hypothetical case of a patient who experiences personality changes after DBS:

> 'Mr. Garrison, a 61-year-old American with PD who consents to DBS to treat his tremors and severe apathy. Following surgery, Mr. Garrison experiences significant improvement in his motor symptoms and dramatic changes in personality. Where once he was shy and introverted, he is now outgoing and gregarious. Where once he was a loyal Republican, he is now a Democrat. Where once he was enthusiastic about his work, he has now quit his job to promote various social, political and charitable causes.'
> (case formulation as in: Baylis, 2013)

This case is hypothetical and simplified but both imaginable and plausible – it is consistent with reports of actual cases of personality changes, changes in attitudes towards work and psychosocial adjustment challenges faced by some DBS patients (Agid et al. 2006, Schupach et al. 2006).

Personality changes that result from DBS, Schechtman argues, are at odds with the articulation constraint on identity-constituting narratives according to which 'the narrator should be able to explain why he does what he does, believes what he believes, and feels what he feels' (1996, p. 114). When the causes of actions can be traced back to the influence of DBS, patients may not be able to explain how their actions are rooted in their 'plans, projects, intentions, beliefs, and desires' – *because they are not*. Schechtman argues that 'his current passions and interests –

the things he takes as reasons – were caused by manipulation of his brain' (2009, p. 85). This requirement can be related to the importance of engaging with reasons for action when pursuing moral modification aimed at creating better moral agents. I have argued that moral enhancement understood as making moral better people necessitates deliberation in general and engagement with moral reasons specifically (see Chapter 1, 3, 7 and 8), and it would be a pyrrhic victory to use means that undermine the ability of an agent to act autonomously and act on the will of their own (see Chapter 7). Schechtman suggests that if Mr. Garrison were to suggest that his new passions and interests were the result of personal development and not DBS, then his narrative would not fulfil the reality constraint, according to which the self-narrative must cohere with basic observational facts about the world. Thus, what seems to be particularly problematic for Schechtman is the means by which personality changes occur – as a result of having electrodes implanted in the brain, not 'natural personal development' (2009, p. 85).

Schechtman's discussion of DBS is relevant for the discussion of MB for several reasons. Firstly, she describes personality change as relevant to moral action. Secondly, even if no morally relevant behaviour was mentioned, the narrative identity objection would apply to all biomedical enhancements that change traits that are important to narrative identity, including moral enhancement in the sense of a morally desirable enhancement of any sphere or function (see: Chapter 1, s. 1.2.1). Biomedical attempts at moral enhancement understood as modification in the moral sphere (Chapter 1 s. 1.2.3), are likely to change how the person functions in their social realm, thus being likely to be relevant to narrative identity as understood by Schechtman. In this discussion, it does not matter whether an intervention is an all-things-considered moral enhancement or dis-enhancement in the sense of making better moral agents (see Chapter 1, s. 1.2.2). In this discussion we will focus on how modification of the moral sphere in general could affects narrative identity.

One could object and argue that the difference between Schechtman's case and MB lies in the intention to change and the desirability of given personality changes. The patient that underwent DBS for Parkinson's disease desired the decrease of Parkinson's symptoms, but not the personality and psychosocial changes. On the other hand, in the case of MB the changes are both desired and intended. But consider an analogical case:

> Mr Morrison, a 61-year-old American who consents to DBS to increase his empathy. Following surgery, Mr. Morrison experiences significant increases in empathy on several measures and dramatic changes in personality. Where once he was shy and introverted, he is now outgoing and gregarious. Where once he was a loyal Republican, he is now a Democrat. Where once he was enthusiastic about his work, he has now quit his job to promote various social, political and charitable causes.

It is not clear that this hypothetical case differs from Schechtman's in any relevant way. Mr Morrison might have wanted a change in a narrow cognitive-emotional aspect, yet got much more than he bargained for. Alternatively, he might have anticipated and accepted a dramatic personality change. One way or another, that he has anticipated and accepted the outcomes does not change the fact that his identity-narrative might have been disrupted.

The case would also be relevant to the ethical evaluation of attempts at biomedical enhancement via direct emotion modulation by psychopharmacological means. Consider the following hypothetical case:

> Mr Ferrison, a 61-year-old American who consents to pharmacological modulation of oxytocin to increase his empathy. Following treatment, Mr. Ferrison experiences significant increases in empathy on several measures and dramatic changes in personality. Where once he was shy and introverted, he is now outgoing and gregarious. Where once he was a loyal Republican, he is now a Democrat. Where once he was enthusiastic about his work, he has now quit his job to promote various social, political and charitable causes.

This example is similar to Schechtman's and is hypothetical yet plausible. Oxytocin might modulate some mechanisms that underpin empathy, but also affects a host of other abilities and functions. It might increase trust in situations judged as safe, result in increased envy, promote parent-child bonding, influence attachment, increase out-group bias, etc. There also seems to be a link between oxytocin and the effects of stress. As a result, if the change in oxytocin levels results in observable changes in morally-relevant actions, it will likely result in observable changes in personality. Thus, Schechtman's objection to DBS is also relevant to at least some proposed methods of MB, and especially to attempts at moral enhancement by direct emotion modulation.

6.2.3. Arguments against Schechtman's objection

However, Schechtman's argument presented in the 2009 paper is not satisfactory on her own terms. In the 2009 paper Schechtman appears to strongly argue that DBS can threaten narrative identity. As Baylis (2013) rightly notices, Schechtman's account as presented in 2009 paper leaves open the possibility that personality changes can be successfully integrated into a subject's autobiographical narrative.[19] There is no reason to think that a subject will necessarily be unable to satisfy Schechtman's articulation constraint: to provide a satisfactory account of her history, her life situation, and her motivations; to narrate parts of her life in a self-conscious way; to render her self-narrative intelligible. Suffering and fighting a disease, thinking about DBS as a treatment option, the process of consent to DBS, the period of adjustment of settings, etc., can all form part of a narrative that incorporates changes that arose as a result of psychotherapy, pharmacological treatment and DBS alike. In a later paper Schechtman acknowledges that some of the change might be absorbed by the flexibility of the narrative, 'since narrative is a dynamic notion, continuity of narrative is thoroughly compatible with even quite radical change' (2010, p. 140).

It therefore seems that the initial argument presented by Schechtman in the 2009 paper would need to rely on the implicit assumption that the mechanism of change matters crucially. This echoes some of the moral unease expressed over a decade ago by Kramer (1993), who, in *Listening to Prozac*, mentions the unease that he experienced when seeing some of his patients change radically after taking antidepressants. Yet, given the above rebuttal of the articulation constraint argument, the reader is left in the dark as to why the involvement of a technological

[19] This holds whether or not a change in a given trait was intended. Narratives are typically thought to incorporate both voluntary and involuntary aspects of experience.

or medical means should be viewed with *special* suspicion. One can wonder why, if it coincidentally so happened that either a sudden and dramatic life experience, a month of practicing qigong or a psychotherapy session produced exactly the same physiological changes in brain function (and resulted in the same profile of personality changes), we should consider this ‘natural’ way of personality change less suspicious than the changes that result from DBS. So Schechtman’s account is open to charges of arbitrarily treating interventions as problematic without defining what kinds of means are problematic, nor giving convincing arguments as to why they are problematic.

A more charitable reading of the argument presented in Schechtman (2009) may get at some of our intuitions as to why a change of values, beliefs or character traits following DBS may be troubling; it is not that DBS is problematic in virtue of it being a technological means of affecting change, but rather because the intervention belongs to a class of change-affecting events which are *difficult to make sense of within a personal narrative*. Perhaps a shift in views after intense qigong practice or a life-shaking event could be equally problematic, if unaccompanied by reflection and integration of the new stance towards life – including giving epistemic and genealogical reasons for this stance.

Moreover, even if DBS would undermine a person’s identity-narrative, there is no reason to think that biomedically induced disruptions in identity-narratives are irreparable. If a person could re-invent a narrative following severe personality changes due to head trauma, there is no obvious reason that a person who undergoes DBS could not. Thus, rather than pointing to a DBS-induced identity dead-end, Schechtman’s account highlights that the integration of personality changes into person’s understanding of themselves may be challenging. The challenging nature of such changes has been highlighted in some research on patients' perspectives (Agid et al. 2006, Shupbach et al. 2006). However, the question about the moral weight of the risk of such identity crisis or narrative disruption remains unanswered.

Another interpretation of Schechtman’s (2009) objection may relate to the intuition that the values, desires and beliefs which can be causally traced back to DBS somehow lack anchoring in the persons own life. When she argues that ‘his [the patient's] current passions and interests – the things he takes as reasons – were caused by manipulation of his brain' (2009, p. 85) this can be interpreted either as referring to the new values not being truly his own (a concern perhaps better understand as a concern about authenticity) or being somehow baseless, epistemically unjustified.

The latter claim represents a related, yet separate worry about the epistemic justification of new values, beliefs and character traits after DBS. This epistemic problem could be especially important if new values, beliefs and character were *less* justified than the previous ones. However, in situations when previous justifications were weak (e.g. a depressed patient who believes that she worthless as a result of trauma), justified in the past but not the present ('My life is full of emotional pain' for a formerly depressed patient) or equally as well or poorly justified as a new belief (e.g. I'm a conservative because my father was a conservative, I'm a liberal because that is my fancy after DBS), the degree of justification of the beliefs does not change. This worry, however, is more closely related to Schechtman's reality constraint on identity-narratives and opens up an altogether different discussion.

6.2.4. Empathic access: another challenge?

Another objection to biomedically induced radical change could stem from Schechtman's distinction between 'person narrative' – i.e., the recognition of oneself as continuing over time – and 'self-narrative', which involves a sense of stable self over time. According to Schechtman, self-narrative, in contrast to person-narrative, requires not only that one remembers or recognises past actions as belonging to oneself, but that one has empathy with one's past actions: *empathic access*. Thus if one recognises that one has done something in the past – a bad act for example – but no longer feels empathy for the person that committed that act, the act is integrated into one's person-narrative, but not into one's self-narrative. Schechtman claims that when self-narrative is discontinuous, one's identity is threatened. It could be argued that even if a patient's narrative after DBS can fulfil the articulation constraint, it cannot fulfil the *empathic access* constraint. Schechtman contrasts the person-narratives and self-narratives in a following way:

> 'Temporally remote actions and experiences that are appropriated into one's *self* narrative must impact the

> present in a more fundamental sense than just constraining options or having caused one's current situation and outlook [as they do in a *person* narrative]. These events must condition the quality of present experience in the strongest sense, unifying consciousness over time through affective connections and identification. To include these actions and experiences in my narrative [i.e., my self narrative] I will need to have what I have elsewhere called "empathic access" to them. In *this* sense of narrative [i.e., *self* narrative], actions and experience from which I am alienated, or in which I have none of the interest that I have in my current life, are not part of my narrative.' (2007, p. 171)

According to Schechtman, having empathic access to an episode of one's past consists of two elements which are individually necessary and jointly sufficient. First, one must be able to remember what happened 'from the inside', with a suitable richness of phenomenology – to have an emotionally rich episodic memory of that event. Second, one must display 'a fundamental sympathy for the states which are recalled in this way' (Schechtman, 2001, p. 106). Empathic access on this account is more than just having an understanding of one's past and being able to make sense of it: empathetic access implies a particular kind of *identification* with one's remembered past. The stable defining traits of which we might not be explicitly conscious, but are revealed in the process of empathically accessing the past, provide the rich self-understanding which makes us, in Schechtman's words, 'intelligible to ourselves' (Schechtman, 2007, p. 18). When empathic access to one's past is absent, the stability of defining traits required for a self-narrative is in doubt and, consequently, survival in what Schechtman calls the 'subtle sense,' is threatened.

The main challenge to the *empathic access* criterion is that the necessity to recognize one's past actions as one's own in the way outlined above might be too stringent a requirement. I will argue that we have two strong reasons to abandon the empathetic access requirement. The first reason is that this view would deny the possibility of survival through radical change even when such change is best understood as personal development – change stemming from personal projects, values, beliefs and experiences. The second reason is that empathetic access unjustifiably privileges a certain kind of backwards-looking attitude, while other backwards-looking attitudes can be seen as sufficient for maintaining a self-narrative.

Consider the conversion of St. Augustine. According to Schechtman, this would be a survival-threatening disruption of narrative personhood rather than a continuing progress toward the good in the life of one particular person, despite the fact that Augustine gave voice to the narrative of change in his *Confessions*. Indeed, Schechtman remarks that religious conversion is ‘frequently cited as a case of identity threatening psychological change’ (2001, p. 105), and adds that the convert often ‘retains vivid recollection of lusts and passions that he now finds shameful and horrible' (2001, p. 105). Thus, according to Schechtman, although the convert maintains vivid memories of past deeds, she lacks the element of fundamental sympathy required for empathic access.

One can easily bring forward other examples: the person who in his twenties thought that being rich and powerful was at the heart of his self-conception, but who now realizes his mistake and feels alienated from what he now considers superficial values; the reformed criminal who for many years thought that robbing people was a fair game, but who now sees that this was ethically wrong and who now feels no sympathy for those mental states that at one time motivated him (Goldie, 2011). In his criticism of Schechtman’s account, Goldie argues that where Schechtman sees a loss of one’s defining traits as a threat to one’s survival, one can easily adopt an alternative position, according to which ‘allowing that change, possibly radical and profound, can be a source of personal moral progress and very much part of the human condition.’ (Goldie, 2011; 2012)

Moreover, it is unclear why ‘sympathy’ should be privileged as *the* backwards-looking attitude that allows affective connections which, according to Schechtman, are necessary for a subtle sense of survival. As Goldie (2007; 2011; 2012) correctly points out, alienation, mortification, ironic distance, amusement and embarrassment are perfectly possible ways of engaging with our past and do not imply bringing our survival (in the identity sense) into question. A similar point was eloquently made by Simon Beck, who, writing of his feelings about his actions when young, says: ‘I cringe at the actions of Simon Beck as a 16-year-old when I can bring myself to think about them. I would not *cringe* if there were not a rich level of continuity of consciousness—that embarrassment requires seeing those actions as my own’ (2008, p. 75). Thus, other kinds of affective connections might ground a deep sense of one’s continuity.

6.2.5. Conclusions

This part of the chapter examined the objection that biomedical interventions such as DBS or moral modification would impair narrative identity and inquired whether such worry would provide a robust basis for an argument that MB is impermissible. After outlining the relevant features of Schechtman’s (1996) account of identity, I have examined the objection she raises in her critique of the use of DBS (Schechtman, 2009). I have argued that DBS patients, and by extension subjects of voluntary MB, are likely to satisfy Schechtman’s articulation constraint, i.e., to provide a satisfactory account of their history, life situation and motivations. Consequently, at least as far as narrative identity is concerned, the ability of agents to make sense of their actions and engage with moral reasons is not obviously impaired. It remains unclear what degree and kind of change would be disruptive on Schechtman’s account and what kind of change can be assimilated into a changing but continuously present narrative. Moreover, it seems that even if narrative identity is disrupted, this does not mean that such disruption is irreparable. What we seem to be left with are worries about the epistemic justification of the post-intervention beliefs and desires, but this worry would only apply if there is a net loss in the degree of justification.

Further, I have argued that Schechtman’s empathetic access criterion should be rejected as too demanding to ground a strong moral critique of MB. Some interventions proposed (see: Chapter 3) aim at increasing empathy. Where those changes would strengthen the ability to empathise with ourselves, the intervention could potentially have strengthened the empathetic identification and thus narrative self-identity of a person. However, even where the moral modification intervention loosens the empathic connection between oneself now and in the past, I argued that this would not provide us with a strong moral objection to MB. Although the ability to understand one's past in the phenomenally rich way – which would also include an affective attitude of sympathy – is one way of engaging with the past, other backwards-looking attitudes can underpin the identity-narrative. Moreover, although the changes in the 'subtle sense of identity' Schechtman is trying to pinpoint while emphasising a sympathy-filled understanding of one's past may be valuable in some way, the life transitions that also are a valuable part of the human condition may require giving that up. The transitions marked by 'I changed so much! I don't really understand how I could value what I did!' may indeed be significant for us, yet to say that they are significant does not imply that they should not be lived through.

Schechtman's account of narrative identity is an interesting attempt to theorise factors that make up human identity. Since the ability of a person to give a satisfactory account of her history, plans and motivations need not to be disrupted by DBS or MB, and the ethical significance of such disruptions remains unclear even if narrative identity was disrupted, Schechtman's account fails to provide basis for a strong objection to the use of biomedical emotion modulation.

6.3. Can Ricoeur's view of narrative identity ground an ethical objection to MB?

6.3.1. Ricoeur's view of narrative identity

Narrative accounts of identity pose that in practice we create the sense of who we are through the construction of an autobiographical self-narrative.[20] Those self-narratives are commonly thought to be the means by which we order our experiences, give and re-evaluate their meaning and importance, understand and give meaning to our lives and even exercise agency. Because those narratives incorporate our motivations, actions, beliefs and decisions within the social context

[20] Different versions of this view are espoused by Paul Ricoeur (1992), Marya Schechtman (1996); Alisdair MacIntyre (1981), Daniel Dennett (1992), Peter Goldie (2003); Jerome Bruner (2003); Charles Taylor (1989); and Anthony Rudd (2009).

and over a substantial span of time, they are thought to explain and shape how we behave and thus influence our identity – who we perceive ourselves to be. It is often posed that construing self-narratives is very important, if not necessary, for a full and flourishing life, and on that ethical account of narrative identity, the ability to construe self-narratives promotes flourishing. In the narrative psychotherapeutic approach, the process of construing, revising and living in accordance with this narrative identity over time is seen as central to personality functioning and development, as well as a person's well-being (McAdams, 2001; Singer, 2004; Singer and Blagov, 2004).

A well-known example of the narrative approach to personal identity is that of Ricoeur (1984, 1985, 1988, 1992), who sees narrative identity as necessary for personal identity. His approach is worth examining for several reasons. Firstly, Ricoeur's approach is rooted in literary discussion, and thus focuses to a greater extent than some other approaches on the specifically *narrative* character of narrative identity. Secondly, the central focus of Ricoeur's approach is the interplay of character and the ethical commitment to self-constancy. Moreover, for Ricoeur, the narrative is what puts together and orders the external and the internal, the involuntary and the voluntary, in the process of making it both understandable and 'mine' – incorporated into self-view. Thus, Ricoeur's account is a *prima facie* fruitful tool in analysing the ethical dimension of emotion modulation leading to character change, be it gradual or abrupt. While accounts such as Schechtman's have been discussed in the context of biomedicine, there has not been much direct discussion of Ricoeur's approach to identity in that context.

According to Ricoeur, narratives incorporate both the voluntary and the involuntary influences on action and the external and internal events, making sense of them as part of a roughly coherent whole. They draw together disparate and discordant elements into the concordant unity of a plot. Narrative identity gives an answer to questions such as 'Who has done this?,' 'Who is speaking?'. In Riceour's words: 'The narrative constructs the identity of the character, what can be called his or her narrative identity, in constructing that of the story told. It is the identity of the story that makes the identity of the character*' (*1992, 147–48).

Ricoeur (1992) separates two distinct, although partially overlapping, senses of identity: *ipse* (sameness, character) and *idem* (self, self-constancy). *Idem* refers to sameness and encompasses numerical identity, extreme resemblance, uninterrupted continuity and permanence over time. In contrast, *ipse* is seen as constancy of the self, for example evidenced by honouring a promise even if one's character (elements of *idem)* has undergone a significant change. Positing *ipse*-identity is to emphasise an ethical dimension of identity: an intentional commitment to constancy even though the character, values and commitments of a person (*idem*) have changed. Narrative identity, according to Ricoeur, is what links the permanence in time of the character (*idem*) and self-constancy (*ipse*). It emphasises that the changing agent is faced with an ethical decision about how to act in contact with others after the change.

There are four main putative ways in which identity can be undermined on Ricoeur's view: (1) the failure to produce a narrative that satisfies the compositional criteria, (2) a threat to the ability to construe narrative, (3) change in character (*idem*), (4) the failure to resist change via promise-keeping and commitment (*ipse*). In this work I will examine the first way of undermining narrative identity, in order to illustrate problems faced by a truly 'narrative' account of identity.

6.3.2. Deficient and dissolving narratives

Not all perceiving, talking and thinking about internal and external events is narration. Ricoeur's account places constraints on what can count as a narrative, and therefore constrains what can serve as the basis of self-understanding. Starting out with Aristotelian rules of composition, Ricoeur (1984, 1985) emphasises the importance of structural unity and wholeness of the story, temporal unification, homophony (single, unified authorial consciousness), monological structure, and privileging the plot over the character, action or thoughts. If those criteria are not fulfilled we might be faced with a description, a dream sequence, a dialogue, a failed attempt at a narrative etc., but not a functioning narrative. A similar idea – the idea of narrative disorganisation – is present in narrative approaches to psychotherapy. In examining Ricoeur's approach, the following sections will employ some ideas and examples from narrative psychotherapy in order to make the possible application of the view apparent.

Before proceeding, it is important to note that there are several ways in which narrative approaches can be incorporated into therapeutic practice. Psychotherapists might perform text analysis of session recordings or homework exercises, focus on examining narratives of events and life narratives to identify main issues or problematic thought patterns, use therapeutic journaling about distressing life events as a means of reducing their impact (Pennebaker, 1997), use storytelling in order to facilitate shifts in the patient's construction of a problem (Blagov and Singer 2004), and so on. However, I'd like to focus on strongly narrative approaches which use not only some of the tools or insights of narrative theories, but rather build on the central tenets of narrative theories.

What is the goal of the kind of psychotherapy that is strongly rooted in a narrative psychological approach? On Ricoeur's view, the goal of psychotherapy would include not just increasing behaviours or interpretations conducive to wellbeing, but rather focus on facilitating the process of narration, strengthening the ability to 'challenge change' (self-constancy) and producing narratively informed life plans. The implication for the aim of therapy is that it is conceptualized in terms of promoting narrative organization. Consistent with this, in psychological narrative terms, psychotherapy is conceptualized as a linguistic practice of narrative articulation and reconstruction (McLeod, 2004). In narrative approaches to therapy, psychological difficulties are often seen as a reflection of 'pathological narratives.' According to Avdi and Goergaca (2007), 'pathological narratives' are seen as self-narratives that do not sufficiently represent vital aspects of lived experience, and therapy is conceptualized as a process of 'story repair', where problematic self-narratives are reconstructed to become more coherent, complex and inclusive.

6.3.3. Narrative disorganisation

One of the most common themes in narrative psychology explains pathology in terms of narrative disorganisation, with therapy being aimed at increasing narrative coherence (Angus and McLeod, 2004). Narratives explicitly expressed during psychotherapy can be seen as a reflection of covert, meaning-making psychological structures, while disorganised and incoherent narratives are symptomatic of a psychological problem. One illuminating example of psychological research concerning consistency of narratives has been done by Dimaggio and Semerari (2001; 2004)[21]. On the basis of therapy transcripts, they develop criteria that distinguish ‘effective’ from ‘ineffective’ narratives, resulting in a typology of what they call ‘pathological narratives’. ‘Ineffective’ narratives are divided into 1) impoverished narratives, which fail to incorporate important aspects of lived experience into the client’s life story, and 2) disorganized narratives, which fail to meaningfully integrate lived experience, thus failing to provide a sense of coherence, continuity and meaning. Each of the two types of ineffective narratives is further divided into subtypes.

Examples of disorganised narratives include basic integration deficit (unintentionally incongruent emotion, affect, verbal expression, posture), overproduction of narrative content coupled with a deficit in hierarchization, and a deficit in integration between multiple self-other representation (e.g. extremely incompatible representations of others held intermittently or simultaneously). For the purpose of the argument let us focus on the latter. I will examine the deficit in integration between multiple self-other representations in light of Ricoeur’s monological/homophonic structure criterion for a narrative, as contrasted with more than one authorial voice.

6.3.4. Deficit in integration between multiple self- or other-representations as an example of a narrative not fulfilling the homophony criterion

What Dimaggio and Semerari call a deficit in integration between multiple self-other representations illustrates how the authorial consciousness can be fractured. Dimaggio and Semerari give an example of a hypothetical male patient who tells of an episode with a friend as a main character; in that one life episode, he describes the person as tender, affectionate, loving, and loved. A few minutes later, in the

[21] Similarly, Lysaker and Lysaker (2002) note how schizophrenics’ stories can be barren, cacophonous or disorganized. The case of schizophrenics is widely discussed test-case for narratives in the discussion about narrative identity.

same session, he relates another episode involving the same person, who is now seen as being intrusive, violent, lacking in respect, and detested. If the patient is challenged too quickly, ‘But a few minutes ago you described another side to this person,’ the patient might reply, ‘Who, me? I have never thought that worm to be worthy of anything. She’s just a worm, and I want to get rid of her.’ Such replies can have a paranoid streak to them. He might say, ‘You’re poking fun at me, Doctor, just like all of them. You’re against me. You take her side. You don’t respect me either. You couldn’t care less about what I’m saying.’ (Dimaggio and Semerari, 2001, p. 13) This paradigmatic example demonstrates how two stories that cannot be fit into one coherent narrative can signal an underlying problem in attributions and interpretations (which presumably should be carried out with the help of an identity forming narrative).

The first question that arises is to what extent the switching between incompatible representations is necessarily a symptom of pathology (for simplicity, as indicated by the presence of externally noticeable harm or distress). Are the self-perception inconsistencies of Dimaggio and Semerari’s patient harmful mainly because they are ultimately rooted in an inconsistency in a person-narrative, or because they bring problems that come with switching between differently valenced emotions and involve distress and rage? Dimaggio and Semerari analyse psychotherapy transcripts, so their examples are bound to include examples of harmful narratives. A structurally similar inconsistency, however, could be referred to in different circumstances as a rich perception of the world.

Secondly, we could ask to what extent incompatible representations translate into a disrupted identity narrative? At times, our representations of ourselves might be inconsistent. For example, we might see ourselves as overburdened and swamped with family and professional obligations when we are tired, but see ourselves as capable of juggling the multiple demands of a full and busy life when rested and energetic. Anything that changes our emotional state radically can effect that change in perspective and so the same question may be asked about the effects of biomedical enhancement. Mood states can be seen as packages: they include representations and evaluation of the world, physical postures, ways of walking, easy or more difficult access to certain memories, acceptance of a differing level of risk, etc. If moral and social enhancement is to be effective in any way, it will likely change our behaviours and dispositions by changing one aspect of our state or a propensity to be in one or another state. However, since our states are packages, the change in emotions will likely result – at least sometimes – in a change in one's stance towards the world and oneself. If the change is of a sufficient magnitude and/or quality, this may result in one authorial voice being present at one time, and a different at another. Yet, it remains unclear to what extent that would be ethically problematic above and beyond ordinary changes in mood. This elucidates two problems with the application of narrative identity theories that require one authorial voice to be present – a) a lack of ways to evaluate whether the single-authorial voice has dissolved, and b) confusion over the ethical importance lacking a homophonic structure.

6.3.5. Against Ricoeur's homophony criterion

An alternative approach within the narrative psychological view incorporates the thoughts that emerged from our discussion above and is rooted in the idea that life narratives can be conceived as the outcome of dialogical processes of negotiation, tension, disagreement, alliance, and so on, between different voices (or perspectives) of the self. This approach has been influenced by authors such as Hermans and Kempen (1993). Drawing on Bakhtin (1929/2000), they propose that the self resembles a polyphonic novel, containing a multitude of internalised 'voices' engaged in internal dialogue. In this view, psychopathology is considered the result of fragmentation within those voices and/or the dominance of one voice over others. In stark contrast to Ricoeur's view, a dialogical disruption can also occur when the diversity of voices collapses into the monologue of a single voice. The other voices of the self are silenced, making different constructions of the events difficult or even impossible. Accordingly, in these narratives, the construction of reality is characterized by redundancy and loss of complexity as experiential diversity is discarded or ignored. Thus, the aim of therapy is to facilitate a reconstruction of the patients' repertoire of positions in such a way that he or she can move flexibly between positions (Hermans, 1997; Hermans, 1996). Although those theories do not explicitly consider situations when different 'stances' are taken due to a biomedical intervention or changes in mood, the general approach can be consistently extended to such cases.

Consequently, a person who chooses to use oxytocin for moral reasons and marvels at the different experiences of life that the change in perception gives, might experience a process of tension between those views that could be both confusing and creative. His narrative (or perhaps – dialogical) identity will be mediated by this internal discussion and we can easily imagine how his experience may be given meaning and structured in that dialogical internal environment. Thus, an internal story that does not fulfil the Ricoeurian criterion of homophony can nevertheless serve all the purposes that Ricoeur sees a narrative as fulfilling: the mediation between sameness and selfhood, structuring and making meaningful the voluntary and involuntary aspects of the internal and external, setting an environment in which life-plans are constructed and enacted. Thus, either homophony is not a necessary feature of identity-construing narrative, or narrative identities are not the only way in which one can make sense of one's life experience.

6.3.6. How many voices are too many?

When we emphasise the fact that the narrative structure proposed by Ricoeur brings with it the danger of discarding meaningful perspectives and experience if they do not 'fit the bill,' the same objection can be mounted against dialogical perspectives. A strong A vs B (e.g. 'being a mother' narrative vs 'having a successful career' narrative, 'being-Indian' narrative and 'being-British' narrative, etc.) dialogue of narratives can lock a person in the perceived opposition, to the exclusion of other possibilities and experiences. A response to that problem could be a) discarding one of the options as less dominant or b) treating one of the options as a footnote to the other, thus trying to merge the two narratives (both of which solutions could be a creative way of resolving conflicts but, arguably, also be charged with resulting in rejecting meaningful parts of experience).

Alternatively, one could work on enriching the dialogue with other voices. This last solution could be in principle pursued to the limits of the cognitive system's capacity and result in interpretative and decisional paralysis. Many of us can recall the experience of trying to take into account several points of view and giving them their due consideration, which at times might have led to an exasperated exclamation: 'I don't know what to think anymore!' Since the capacity of our cognitive system is limited, we have to attend to some and ignore other potentially relevant and meaningful experiences. This happens both on a very early level of attentional processing (the top-down aspect of perception) and can also happen on a cognitively complex level of narratives and internal dialogues. I want to suggest that the cost-benefit analysis (epistemically and pragmatically) need not turn out in favour of a single, homophonic narrative view, even if the narrative is flexible, dynamic and can be rewritten.

How many voices and perspectives one can maintain (without decisional paralysis), and how strong and pervasive several voices can be maintained and at what cost, may differ between individuals. Moreover, I do not see a good argument in favour of a particular stance towards this issue. Just as writers may choose to develop and perfect 'their own writing style' or choose to play with a multitude of heteronyms, as in Fernando Pessoa's case, similar options are open regarding identity – identity rooted in one or multiple voices or stances.

6.3.7. Conclusions

The discussion of Ricoeur's view of the compositional constraints put on narrative illustrates a problem that many narrative views are facing: the problem of striking a balance between positing criteria that will make identity truly narrative on the one hand, and maintaining descriptive accuracy as well as a link between narrative identity structure and the ability to lead a full and flourishing life on the other. I have argued that the homophony criterion is too stringent, as it may come at the cost of discarding valuable parts of experience, while the dialogical-self may fulfil the functions Ricoeur assigns to a homophonic narrative: the mediation between sameness and selfhood, structuring and making meaningful the voluntary and involuntary aspects of the internal and external, setting an environment in which life-plans are constructed and enacted.

6.4. Arguments against the strong ethical narrative view and the strong psychological narrative view

In earlier parts of this chapter we considered two approaches to narrative identity and examined their application to the effort of drawing normative conclusions. However, one can also criticise the view (sometimes referred to as the ethical narrative view) that a coherent and continuous narrative is necessary (or at least necessarily highly conducive) to a flourishing life. Vice's objection is a direct response to the strong ethical narrative claim that we *ought* to think of ourselves as protagonists in our own stories if our lives are to have any meaning at all (Vice, 2003, p. 101). Vice argues that if we take the narrative view 'seriously' and 'literally,' it requires that we cast ourselves as 'characters—usually the protagonists—of the stories we tell or could tell about ourselves' (2003, p. 93).

Moreover, Vice sees this view of one's life as having potentially harmful implications. Vice argues that if we try to mould ourselves to fit a narrative and conceive of ourselves as a particular 'character' (which she thinks the narrative view recommends) then we are likely to be more prone to self-deception, and may undermine who we really are in our efforts to fit the trappings of a specific preconceived character. Vice argues that in doing so we run the risk of constraining autonomy and being inauthentic, since those who try to live up to the standards of their perceived identity are limited to making choices consistent with that perceived identity.

A crucial issue here concerns the stringency of the criteria required for the creation of a specifically narrative identity. Mackenzie and Poltera (2010) reject the conception of narrative self-constitution that underpins many "story-telling" critiques, such as those of Strawson (2004) or Vice (2003). For example, Strawson (2004), suggests that narrative self-creation involves thinking of oneself as if one were a character in a novel, or 'thinking of oneself and one's life as fitting the form of some recognized genre' (p. 442). Mackenzie and Poltera (2010) argue that criticism of the sort advanced by Strawson and Vice present narrative accounts as more rigid than they are.

However, it may be that Strawson's and Vice's arguments apply more readily to some demanding accounts of narrative identity. For example, Ricoeur's (1984, 1985, 1988, 1992) account of narrative identity is sometimes seen as imposing stringent criteria of structural unity of the story and homophony (single, unified authorial consciousness). On the one hand, critics argue that these stringent compositional criteria could lead to meaningful and important experiences going unnoticed, being trivalised or repressed under the Ricoeurian approach (e.g. as in Muzak, 2007). Giving examples of multi-cultural experience and cultural dis- or re-location, Maan (2010) suggests that a widening of structural requirements (e.g. allowing a multitude of voices) would allow those ignored but important parts of experience to be integrated into a process of creative re-assignment of meaning. On the other hand, Ricoeur's approach can be seen as a response to a perceived postmodern fracturing of the subject. Loosening the criteria too much risks making the narrative indistinguishable from a description or a dream sequence, thereby resulting in the loss of what makes a narrative identity specifically narrative. This discussion, rooted in the literary tradition, has its analogue in discussions within the field of narrative psychotherapy (e.g. Hermans, 2003, Hermans and Dimaggio, 2004).

Even if the full blow of Vice's and Strawson's claims here are taken only by the more demanding narrative views, other aspects of their critiques remain relevant to approaches based on a more flexible narrative account. Contrary to what Mackenzie and Poltera (2010) argue, they cannot be easily discounted. Vice and Strawson can be seen as advancing two points. The first questions the universality of the psychological narrative thesis, and the second questions the necessity of attaching value to consistent narration, as in the ethical narrative thesis.

First, the empirical and conceptual question is whether people do indeed think about themselves in narrative terms. Vice (2003) questions the *psychological narrative thesis* and argues that while some people may think of their lives and themselves in narrative terms, few do; and those who do, tend to do so only when they are being particularly reflective (p. 97). I am uncertain whether this is empirically true, and the strength of the argument will depend on the identity criteria posited by a particular theory of narrative identity. However, it is plausible to posit individual differences in cognitive styles which influence the tendency for individuals to create life-narratives generally and identity-constituting narratives specifically. Perhaps a conceptually and methodologically sophisticated empirical philosophy study will shed light on this issue. In the absence of convincing evidence, it is plausible to assume that there are some people who understand their own lives through an elaborate narrative and in fact explicitly evoke it on a regular basis, there are other people that re-construe and evoke an otherwise transparent self-narrative only when prompted by circumstances, and there are yet others who do not construct their lives through and in a narrative that would fulfil criteria for an identity-narrative of a given narrative view even if they use stories to communicate meaningful experiences (Strawson, 2004). The latter sort of people may not be particularly reflective. Or they might suffer from a mental health issue that undermines their capacity to form complex narratives, to form and act on their endorsed preferences, or both. They also could be highly reflective, but in the Strawsonian way 'episodic', assigning little importance to a consistent narrative line, while, perhaps, focusing on current reflectively endorsed interests, desires and preferences. In a narrative view, the lives of the last group are less meaningful and incompatible with flourishing.

This brings us to the examination of the narrative ethical view. Similar to Vice (2003), I see no strong reason to think that the lives of the last group are necessarily less meaningful or incompatible with flourishing. There may be an aesthetic allure in the Ricoeurian view of fully flourishing life as life that is 'complete', but I doubt whether such an aesthetic preference translates into an epistemic or moral imperative to lead or even aim at such a life. The move from how *we might* construe ourselves through the process of creating an autobiographical identity narrative to the conclusion that construing of such personal narrative is *necessary* for a meaningful or flourishing existence is a strong thesis, and I have yet to encounter a strong argument in favour of this claim. It remains an open philosophical question whether only certain autobiographical narratives can be personal identity conferring, or whether identity can also be conveyed by non-narrative items such as core beliefs, values, life plans, roles or relationships. And in medical practice, I see no strong reason to favour a narrative identity view (and accept the potential objections to DBS coming with it) *if* the patient does not endorse or care about it.

A related objection is rooted in the observation that *not all identity narratives are good for you*. A narrative of ‘how I fail in most important aspects of life’ or ‘how my chronic pain determines what I do in life’ might fulfil the criteria for an identity-constituting narrative yet be hardly conducive to wellbeing (Morris 2012; Farkas 2013). Narrative scholars have long been comparing wellbeing- and autonomy-promoting narratives and harmful and autonomy-undermining narratives (Frank, 1997; Brody, 1994; Sontag, 1978), both in medicine and psychiatry. Psychotherapy viewed through narrative glasses can be seen as both a process of developing impoverished narratives or fixing fractured narratives through a process of re-writing harmful narratives into ones that are conducive to a flourishing life (Kirmayer, 2000; Adler et al. 2008; Angus and McLeod, 2004). Even if we were to accept the strong ethical narrative thesis that narrative is necessary for a flourishing life, recognizing the deleterious aspects of some narratives opens the door to the conclusion that one might be better off in a state of narrative-less existential puzzlement than to be stuck in the rut of a harmful narrative.

6.5. Conclusions

The challenges with translating narrative approaches to identity into strong ethical reasons to forgo the biomedical modification of emotions and desires rests on three groups of problems. The first problem faced by those approaches lies in providing an account of identity that is specifically narrative, but without being so stringent as to describe people who seem to have meaningful lives and some stable self-understanding as fatally lacking in a domain that is supposedly necessary for flourishing. Ricoeur’s (1984, 1985, 1988, 1992) account, with his explicit attention to the specific features of the narrative, was a good example to examine in this context.

The second problem for the narrative identity theorist is accounting for change in lives, desires and beliefs and its relation to identity-narrative change. Although

Schechtman (2010) proposed that the identity-narratives of DBS patients who experience marked personality change fail the articulation constraint, I have argued that this is not the case. I think that there are ethical questions raised by the possibility of rapidly changing one's personality, but whatever they might be, they are not captured in Schechtman's (2010) objection to DBS. We are told that narratives are flexible but yet can break, but the account is not fleshed out enough even to account for what the problem is in cases in which DBS patients clearly experience unwanted, serious and rapid personality changes that dramatically change their lives and disrupt their social relationships. This means that this narrative account of identity is ill-fitted to guide our decisions about biomedically-induced changes, as is also the case with voluntary MB.

The third problem concerns the value of the narrative identity and its disruptions. Even assuming that the psychological narrative thesis is true, i.e., that we all construct identity-narratives, doubts can be raised as to the value and importance of those narratives and the ethical significance of their disruptions. Some changes that can give raise to such disruptions can be a valuable part of human existence and a necessary step towards a better state. Some identity narratives might be facilitative of wellbeing, while others are detrimental, and a rigid identity-narrative may both give meaning and clarity to a diverse mess of life experiences as well as impoverish that experience and meaning. Thus, it is unclear whether the goal of maintaining a stable identity-narrative would be desirable *ipso facto*. As a result, it is difficult to derive strong action-guiding ethical conclusions from the consideration of narrative identity accounts. I therefore conclude that the narrative-identity objection to the use of MB is weak, even in cases where the intended or unintended effects of MB do not clearly happen within the bounds of the person's identity.

CHAPTER 7. Freedom, autonomy and the God Machine

7.1. Introduction

Proposals suggesting the use of biomedical emotion modulation for achieving better moral outcomes (Persson and Savulescu, 2008) were met with criticism (e.g. Harris 2010, 2014a, 2014b; Sparrow, 2014). An important objection originally raised by Harris (2010) highlights the potential negative impact of MB on freedom. Harris argues that even if the interventions Persson and Savulescu (2008) propose were to be effective (i.e. achieve the behavioural end), they would come at an unacceptable cost to the kind of freedom that is the foundation of moral agency while grounding moral responsibility. In response to Harris and other critics, Savulescu and Persson (2012a) develop a thought experiment in which 'the most powerful bioquantum computer', which they nickname 'the God Machine', uses direct modulation of intentions to prevent citizens from doing great evils. The thought experiment purports to demonstrate a technological intervention that, while it directly modifies the roots of morally relevant actions, is still desirable.

In this chapter, I look closer at the God Machine and its impact on moral life and freedom, while examining the criticisms raised by Harris (2014a) and Sparrow (2014). I will examine the God Machine thought experiment as well as introduce a several similar thought experiments in order to tease out what is an important problem giving raise to an ethical worry in the God Machine scenario. I will use the Moral Luck Machine thought experiment in order to argue that there is a relevant difference for moral agency between the two scenarios: Moral Luck Machine allows for greater engagement with reasons by the agent. I will use the Rational Persuader Machine thought experiment to further support that point. It seems that even though an outcome (agents actions in the world) do not change, there is a difference in how we arrived at the outcome. Finally, I will use Halls Brian Implant thought experiment to provide an example closer to the real world, to argue that it is not overdetermination of agent's actions (resulting in the comparable outcomes in the world) that is the main problem raised by the God Machine thought experiment. Instead, I will argue that the most significant impact of the God Machine's intervention lies in the fact that it diminishes the agent's appropriate engagement with reasons and freedom of thought, rather than the fact that it has an impact on freedom of action or on freedom from political domination.

The analysis presented in this chapter is aimed at a) addressing the issues raised by God Machine thought experiment, the interpretation of which has been used as a basis for arguing both for and against the desirability of MB and, b) using the analysis to tease out the important factors influencing moral agency and freedom in cases of overdetermination. However, as I argue in the last section of this chapter, the utility of those arguments in discussing the real-world MB is limited. Similarly to other instances of using this philosophical method in discussions regarding real-world applications of technology, thought experiments of this level of abstraction are useful tools in teasing out the aspects that are relevant, drawing attention to some aspects of reality and helping us in building conceptualizations that then can be of help when applied to real-world cases. However, those gains do not translate straightforwardly into moral assessment of real-world interventions. Thus, the analysis in this chapter aims at providing additional conceptual clarity and providing further support for the importance of engagement with reasons for moral agency but not as an argument about real world moral enhancement. Chapter 8 will address this limitation and build on the analysis presented in this chapter.

After introducing Savulescu and Persson’s (2012a) thought experiment in section 7.2, section 7.3 examines the God Machine’s impact on ascription of praise and blame. I conclude that the problems the God Machine brings to the ascription of moral praise and blame do not amount to a serious argument against its desirability. In section 7.4, I analyse the God Machine’s impact in light of the distinctions proposed by Frankfurt and argue that the God Machine preserves freedom of action and even free will, but affects the ability to form ‘a will of one’s own’ and impacts freedom of thought. Section 7.5 responds to Harris’ argument in which he sees the problematic impact of the God Machine as relating to the divorcing of thought and action: ‘Decisions to no effect are pointless from the moral perspective; for what is a good state of mind worth, if it makes no difference to the world?’ (2014a, p. 249) In response, I argue that the agent’s lack of appropriate engagement with reasons is the crucial factor, rather than the fact of changing the behaviour – that the impact for freedom of action related to overdetermination and the ‘direct’ mode of changing behaviour is less important than the lack of awareness and control of the agent over those changes and that a good state of mind is worth quite a lot. It may even ground meaningful ‘alternative possibilities’.

Section 7.6 examines Sparrow’s (2014) argument according to which the God Machine scenario is undesirable because the God Machine dominates its subjects. I argue that Sparrow’s argument depends on an inappropriate personification of the God Machine and that the God Machine is better understood as an analogy for the law. Pettit’s (1997) theory of freedom as non-domination is an unfortunate choice as the basis of Sparrow’s (2014) critique because it focuses on the relation of *persons* and is limited in the analysis of the freedom-impairing impacts of structures. Insofar as we treat the God Machine as a structure similar to law, there is no strong reason to suppose that it would enable the exercise of others' arbitrary and unchecked power in *the idealized conditions* of the thought experiment.

However, as I argue in section 7.8, worries about the arbitrary and unchecked use of power become warranted if we consider the application of the God Machine in a possible world more like ours. Moreover, Savulescu and Persson’s (2012a) appeal to Mill’s harm principle is weak because it relies on a misinterpretation of the application of the harm principle. In fact, the harm principle would provide a stronger argument *against* the God Machine’s changing people’s intentions. Moreover, the spirit of Mill’s use of the harm principle is to protect the individual’s private sphere from state and societal encroachment, and there is nothing more central and private than the ‘inner citadel’ of our thoughts.

In the final section of this chapter I outline the limitations of the application of the God Machine to the analysis of the impact of MB on moral agency.

7.2. Introducing the God Machine

In response to Harris’ (2010), Savulescu and Persson (2012a) ask us to consider a possible world in which people are not ‘free to fall’:

> ‘The Great Moral Project was completed in 2045. This involved construction of the most powerful, self-learning, self-developing bioquantum computer ever constructed called the God Machine. The

God Machine would monitor the thoughts, beliefs, desires and intentions of every human being. It was capable of modifying these within nanoseconds, without the conscious recognition by any human subjects. The God Machine was designed to give human beings near complete freedom. It only ever intervened in human action to prevent great harm, injustice or other deeply immoral behaviour from occurring. For example, murder of innocent people no longer occurred. As soon as a person formed the intention to murder, and it became inevitable that this person would act to kill, the God Machine would intervene. The would-be murderer would 'change his mind.' The God Machine would not intervene in trivial immoral acts, like minor instances of lying or cheating. It was only when a threshold insult to some sentient being's interests was crossed would the God Machine exercise its almighty power. (…) Human beings can still autonomously choose to be moral, since if they choose the moral action, the God Machine will not intervene. Indeed, they are free to be moral. They are only unfree to do grossly immoral acts, like killing or raping. This is seen as preferable to physical incarceration, which physically restricts the freedom of the immoral. While people weren't free to act immorally in the 'old days,' since the law prohibited it on pain of punishment, the instalment of the God Machine means that it has become literally impossible to do these things. It is seen as preferable that would-be murderers "change their minds", rather than an innocent person is killed and then the murderer incarcerated for life. And, the would-be murderer never knows that her intentions have been changed by an authority outside of herself. It seems to her that she has "changed her mind" spontaneously – she experiences a life of complete freedom, though she is not free. Although any intention to kill or rape immediately changed, this was put down to the efficacy of moral education. It seemed "from the inside" that she had just developed an aversion to killing an innocent person. And no one was ever killed.' (Savulescu and Persson, 2012a, p. 411)

7.3. Moral praise and blame

7.3.1. Black vs. the God Machine

The God Machine has parallels with Frankfurt's counterexample to the incompatibilist 'principle of alternate possibilities' (PAP)[22]. Beginning on page 835, Frankfurt asks us to consider a following example:

> 'Suppose someone -- Black, let us say -- wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. ... Now suppose that Black never has to show his hand because Jones, for reasons of his own, decides to perform and does perform the very action Black wants him to perform. In that case, it seems clear, Jones will bear precisely the same moral responsibility for what he does as he would have borne if Black had not been ready to take steps to ensure that he do it. It would be quite unreasonable to excuse Jones for his action ...on the basis of the fact that he could not have done otherwise. This fact played no role at all in leading him to act as he did. ... Indeed, everything happened just as it would have happened without Black's presence in the situation and without his readiness to intrude into it.'

Frankfurt describes his argument as one which shows that one can be morally responsible even if one could not have done otherwise. It is interesting to consider the implications of the God Machine for the ascription of moral blame and praise: in the God Machine World, we never know whether someone is or is not

[22] According to the principle of alternate possibilities, a person's act is free if and only if that person could have done otherwise.

praiseworthy for their lack of acting grossly immorally. This would certainly change the way we think about morality with respect to grossly immoral acts.

There remains a question about the value of knowing whether refraining from a grossly immoral action was praiseworthy or not, vis-à-vis considering the moral weight of the averted harms of immoral acts. Moreover, the worry about being able to ascribe moral blame and moral praise in the God Machine scenario seems to be putting the cart before the horse. If we see the accuracy of our ascription of moral blame and moral praise as having a solely instrumental value in promoting (via rewarding with praise) or discouraging (via expressing blame) certain acts, by praising everyone indiscriminately, we can assure that the deserving are rewarded. Inaccurate moral praise for those who indeed formed the intention would only encourage the kind of thinking and predisposition that lead to forming a bad intent, yet we may rest assured that this would not result in grossly immoral acts anyway, since the God Machine would intervene if such intention was formed again.

There are two crucial ways in which the God Machine scenario differs from Frankfurt's thought experiment. Firstly, in Frankfurt's scenario, Jones' actions are attributable to Jones since Black did not exert his influence on Jones. If Black was going to act is such a way to make Jones refrain from acting immorally, and he had to exert force to cause Jones to refrain from such action, Jones would not be considered to be praiseworthy for refraining. The God Machine scenario is more ambiguous because there is no obvious external coercion as would be the case if Black threatened Jones with a gun, or employed persuasion which would still leave Jones' being able to consider the influence - instead the God Machine acts directly on intentions without the knowledge of the influenced actor and thus, in Harris' words 'bypasses morality'. The influence itself cannot be modified or responded to by the agent, at least not in a conscious way. I examine the implications of that in later parts of this chapter. At this moment let us focus on the second way in which the God Machine scenario differs from Jones & Black thought experiment – the issue of the problems created by the fact that the intervention happens without agents' knowledge.

The God Machine scenario could undermine the value of moral praise. The first way this could happen relates to the undermining of reliable self-knowledge coming with the God Machine scenario. Since in Savulescu and Persson's (2012a) world, the person forming such an intention would not know that this intention has been changed by an authority outside of herself, all those subject to the God Machine interventions – both those who would have 'changed their mind' and those who would not – would have good reasons to doubt whether they deserve praise. Another way in which the value of moral praise could be undermined would be analogical to a situation when students are indiscriminately praised for their effort and hard work, even though some of the students did not study at all while others did. It matters to us that the praise is deserved and given on the basis of a roughly accurate assessment of reality. Thus, even though there were some instrumental reasons for continuing to praise, ultimately the undermining of the value of such praise could plausibly lead the God Machine society to abandon any talk about praise and blame, at least in so far as that includes taking credit for living in a society without grossly immoral acts. Those whose moral actions were praiseworthy could perhaps still freely choose to do the right thing, yet, when the stakes are high, their actions would cease to be a subject of moral praise. [23]

However, perhaps the above worry is overblown, given that a) we normally do not expect nor give praise for refraining from grossly immoral actions, b) we presently are also not certain about the extent to which the apparent moral behaviours (or lack of grossly immoral acts) of others are due to their virtue, and c) perhaps the God Machine could be treated as an instance of moral luck. Let us consider these reservations in turn.

It could be argued that the undermining of praise for refraining from grossly immoral actions in the God Machine society is unimportant because we normally do not give praise for not doing great evils anyway. Our unwillingness to kill an innocent person is something we expect and may see as a morally good thing, but not something deserving of moral praise. It then becomes irrelevant (for our doubts about moral praise in the God Machine scenario) whether the inability to kill an

[23] Although so far this section has simplified the issue of praise and blame in relation to action, this way of elucidating the implications of the scenario shed light on why we could find the God Machine scenario not only alien but also troubling.

innocent person is due to insurmountable squeamishness, strongly held values or preferences, or the God Machine– as we would not give any praise anyway. The question of moral praise and blame in relation to a person’s preferences seems to become more important when we consider an intention to act rather than refrain from acting. The question of moral praise and blame in relation to a person’s preferences seems to become important when we consider those actions one intends to commit rather than those actions one refrains from committing.

However, this seems not to be the case when a temptation (a *prima facie* desire or intention) is present. Although our moral praise, if at all present, is rather luke-warm when we simply do not have a desire to kill (we rarely morally congratulate ourselves or others for not killing), our intuitions become different and more complex when a desire is present but overcome by moral considerations. Consider the example of Wilson, who at one time formed an intention to kill in revenge, yet overcame it by referring to moral considerations, his values and preferences. Among other evaluative attitudes we might have, we would likely see the lack of immoral action on his part as praiseworthy, similar to praising an addict for not taking heroin while a non-addict’s abstinence would not evoke similar moral praise. Thus, it seems that the God Machine scenario takes away the meaning of the praise mainly for those who might have formed the intention to do a grossly immoral act, yet refrained from acting on this intention or changed their minds for moral reasons. As important as this can be, this would probably be a small proportion of the population of the God Machine society.

The second doubt relates to the fact that we presently have a degree of uncertainty about the extent to which the apparent moral behaviours (or lack of grossly immoral acts) of others and ourselves are due to virtue or vice. We interpret the behaviour we observe in moral terms and play a guessing game to establish intention as well as reasons on which a person acted. Moreover, we also have good reasons to accept that there is a fair degree of uncertainty about our own reasons for actions. In his memoir, Richard P. Feynman (2010), a Nobel Prize winning physicist, describes some of his experience with hypnosis. Although he initially intended to uncover the hypnotist’s ‘phoniness’ and resolved to try hard to resist the post-hypnotic suggestion (to return to his seat only after going around the room three times), in the end he found himself finding reasons to do exactly what the hypnotist suggested. It is not only with hypnosis that we sometimes are deluded or self-deceptive about the reasons and sources of our intentions and actions.

A long list of cognitive biases adds to the uncertainty. For example, in ‘post-purchase rationalization’ (Aronson, 2007; Cohen and Goldberg, 1970) the purchaser seeks to argue *post factum* (to others and themselves) that they have made a financially sound decision. Although this is typically cited as a ‘cognitive’ bias, it speaks powerfully to motivational states which have their roots outside of conscious awareness (Bos et al., 2008 ; Custers and Aarts, 2005 , 2007 , 2010 ; Dijksterhuis and Aarts, 2010). As a result it makes sense to accept that we do not have perfect insight into ours or other's reasons for action and thus cannot be sure whether the praise or blame given is given accurately– even without the existence of the God Machine. Yet the lack of certainty about the effectively motivating reasons for our actions is not normally a sufficient reason to stop morally praising or blaming. From the point of view of those living in the God Machine’s world, its interventions can be seen as another source of uncertainty and not a significant game-changer in terms of ascription of moral blame and praise. Thus, perhaps the issue of moral praise and blame in the God Machine world should not worry us.

7.3.2. The making of moral luck

The third doubt is that perhaps the God Machine could be treated as an instance of moral luck. Nagel (1979) and Williams (1981) have challenged the intuitively appealing ‘Control Principle’ (CP) as well as a corollary of it (CP2):

(CP) We are morally assessable only to the extent that what we are assessed for depends on factors under our control.

(CP2) Two people ought not to be morally assessed differently if the only other differences between them are due to factors beyond their control.

Although the principles seem appealing and plausible and are consistent with many instances of our moral assessment, there seem to be numerous cases in which those do not apply. For example, we seem to blame those who have murdered more than we blame those who have merely attempted murder, even if the reason for the lack of success in the second case is that the intended victim unexpectedly tripped and fell to the floor just as the bullet arrived at head-height, by which time the person with the intention to murder was incapacitated by guards. Since whether the victim tripped or not and whether the guards arrived quickly enough or not is not something in control of either would-be murderer, the case appears to violate the Control Principle and its corollary.

As Nagel (1979) puts it, ‘Where a significant aspect of what someone does depends on factors beyond his control, yet we continue to treat him in that respect as an object of moral judgment, it can be called moral luck’ (p. 59). Moral luck includes instances of resultant luck (how things turn out), circumstantial luck (e.g. people working in Germany in the 1930s vs. those living in another time or place) and constitutive luck (the luck regarding acquisition of traits and dispositions). As Feinberg (1970, pp. 34–38) argued, moral luck can affect even our ‘willings’ and other internal states. According to Nagel’s development of this point, there are other types of luck that affect not only our actions but also every intention we form and every exertion of our will. Furthermore, once these kinds of luck are recognized, we will see that not one of the factors on which agents' actions depend is immune to luck. Those who formed the intention to commit gross immorality and were prevented from acting on it by the God Machine, at least for the purposes of moral praise and blame, could be seen as being like those whose victims tripped or those who, through a chain of lucky coincidences, changed their minds.

Consider the example of Wilson Jr. who formed an intention to kill in revenge, yet on the way to commit the murder passed through a student protest and was mistakenly arrested on suspicion of another offence. With his phone call from prison, he intended to get his friend to put forward bail so he could carry out his plan. However, influenced by the soothing conversation with the friend and the passage of time, his intention changed by the time he was released. Were it not for the accidental arrest, he would have killed. The God Machine could be seen as a rather ‘hands-on’ instance of (inescapable) moral luck.

One could object that there is something relevantly different for the ascription of praise and blame between the God Machine scenario and the instances of moral luck discussed above: in the God Machine scenario the agent is neither responsive to reasons nor subject to a lucky coincidence or favourable circumstances, but rather directly affected. Let us imagine that a segment of the God Machine society raised this objection and as a result, the God Machine was improved.

> **The Moral Luck Machine.** The Great Moral Project was re-started in 2067. The most powerful, self-learning, self-developing bioquantum computer ever constructed called the God Machine has been improved. The Moral Luck Machine would monitor the thoughts, beliefs, desires and intentions of every human being. However, instead of modifying these directly, it creates the circumstances under which a person who formed an intention to commit a grossly immoral act would change their minds of their own accord. The Moral Luck Machine would never influence the behaviour of other people directly, but, through a network of hired staff and changes to inanimate objects, create the circumstances in which a person would change her mind. Sometimes it would be a brick falling close to them that would make them reflect on how short life is. Sometimes it would be delays in the metro or a hologram of a street preacher saying just the right words. Since the Moral Luck Machine was monitoring the thoughts, beliefs and desires of everyone and had run simulations, every intervention of the Moral Luck Machine would be effective. As before, it would only intervene once the threshold of harm to others interest was reached. Murder of innocent people no longer occurred and everyone was given the best chance to (inevitably) change their minds.

What would be the implications of the Moral Luck Machine for praise and blame in the imagined society? The Moral Luck Machine creates conditions for change of intention, which could be seen as an instance of a hybrid between circumstantial and constitutive luck. Among those we would praise for moral conduct in the God Machine scenario, there are a) those who changed their minds even without circumstantial-constitutive moral luck, b) those whose actions were influenced by circumstantial-constitutive moral luck, and c) those who had their intention directly changed. In the Moral Luck Machine scenario there are a) those who changed their minds even without circumstantial-constitutive moral luck, b) those whose actions were influenced by circumstantial-constitutive moral luck, and c) those whose actions were influenced by circumstantial-constitutive moral luck delivered by the Moral Luck Machine. The central question is whether there is a significant difference *for praise and blame judgements* between the circumstantial luck of those in the Moral Luck God Machine scenario and those whose intention was directly changed.

It seems to me that the relevant difference lies in *some* engagement of the agent in the process of changing her mind – the Moral Luck Machine might appeal to some values and preferences of the agent that perhaps would be weaker or not otherwise effective. If the God Machine changed people's intentions by evoking an intrusive thought to which one would have to relate instead of changing the intention directly, there would perhaps be little difference between the Moral Luck Machine and the God Machine. Thus, the difference may not be control or the external-internal influence, but rather the degree of agents' engagement with either internal (intrusive thought) or external (circumstantial moral luck) circumstances. While the Moral Luck Machine intervention allows agents such engagement with internal influences, the God Machine scenario does not. In his discussion of moral luck, Adams (1985) makes a similar point. He recognizes that there are limits to what we can be responsible for, and writes that the states of mind 'for which we are directly responsible are those in which we are responding, consciously or unconsciously, to data that are rich enough to permit a fairly adequate ethical appreciation of the state's intentional object and of the object's place in the fabric of personal relationships' (1985, p. 26).

As a result, it would make sense to confer a certain degree of praise for all people who experience a temptation but refrain from acting on it, as in the Moral Luck Machine scenario. The kind of moral luck provided by the original Savulescu and Persson's (2012a) God Machine, however, seems to deliver a rather different kind of moral luck that is relevant for the ascription of praise and blame. At least in part, the ascription of praise and blame depends on the ability to relate to the internal or external influence independently of the outcome. I will return to the importance of appropriate engagement of the agent with external influences in the latter sections. Here it suffices to say that even if the influence of the various morally relevant machines discussed in this section would be construed as instances of moral luck, it matters what shape that moral luck takes.

In the next section we will leave the issue of moral praise and blame aside and focus on the impact of the God Machine on free will and free action.

7.4. Free will and free action

7.4.1. Introduction

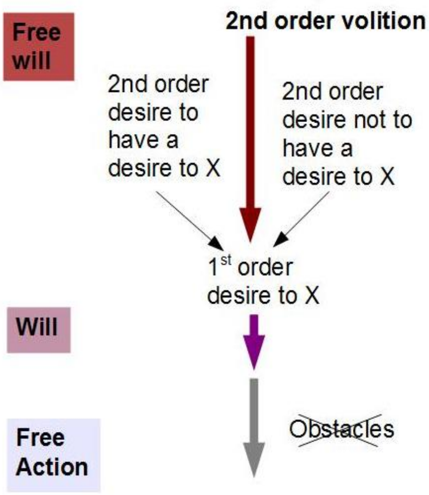
In response to the God Machine thought experiment, as well as Savulescu and Persson's proposals for MB via emotion modulation, Harris raised concerns about the negative impact of MB on freedom. Harris writes: '[t]he space between knowing the good and doing the good is a region entirely inhabited by freedom. Knowledge of the good is sufficiency to have stood, but freedom to fall is all. Without the freedom to fall, good cannot be a choice; and freedom disappears and along with it virtue.' (2011, p. 104).

In the following two sections I am going to argue for the modest claim that 'freedom to fall,' in the sense of free will and action is not *all* that is important about freedom as it is relevant to moral responsibility. Firstly, I will apply the Frankfurtian distinction between free will, free action and free will of one's own to the analysis of the God Machine and similar cases. I suggest that the God Machine undermines something very important – the ability to form a will of our own. It may also undermine free will and freedom of action, but those impacts are secondary and follow from the 'upstream' intervention.

7.4.2. Frankfurt on free will and free action

Let us first draw upon Frankfurt's distinction between acting freely and freedom to do otherwise. An agent acts freely, according to Frankfurt, when her action issues from her own (properly functioning) volitions unimpeded by external impacts. 'Will,' according to Frankfurt, refers to the first order desires that are motivationally effective[24], that is, desires that have motivated, are motivating, or will motivate an agent to act and are followed:

> 'To identify an agent's will is either to identify the desire (or desires) by which he is motivated in some action he performs or to identify the desire (or desires) by which he will or would be motivated when or if he acts.' (Frankfurt, 1971, p. 325)



Free will
2nd order volition
2nd order desire to have a desire to X
2nd order desire not to have a desire to X
2nd order desire to have a desire to Y
1st order desire to X
1st order desire to Y
Will
Obstacles
Free Action

Further, Frankfurt proposes that what matters for 'free will', are not second-order desires but second-order volitions. Second order volition refers to what first order desire an agent wants to make motivationally effective, that is a will to make a certain desire be an actual motivating force for a subsequent action. In his *Free Will and the Concept of Person*, he develops the concept of a *wanton*. Wantons are creatures that have first order desires, and may even have second-order desires (desires to have or not to have certain first order desires), however, they lack second order volitions. Wantons are indifferent as to which first order desire will move them to act, they have a will (the effective first order desire) that is their own and if nothing comes in the way of their action, they may act freely on their own will. They do not, however, have 'free will'.

In contrast, persons have second-order volitions. They may be able to make their second order volition have the effect of making some desires motivationally effective. In those cases and when the second order volition arises in an appropriate

[24] Frankfurt uses the idea of making desires 'effective' in the sense of their giving rise to a motivation that will move an agent to act. The actions themselves may be effective or not in the sense of being able to effectively change the outside world to have the desire satisfied. To avoid confusion I will use the term 'internally effective' where Frankfurt uses the term 'effective'.

way, the person can be said to have their *own free will*. However, we can bring up many examples when peoples' second order volitions do not translate into which one of their first order desires is motivationally effective (e.g. an unwilling addict). Those cases can be described as an instance of weakness of the will or akrasia (Davidson, 1970, pp. 21-42).

7.4.3. Frankfurt's own free will and the God Machine

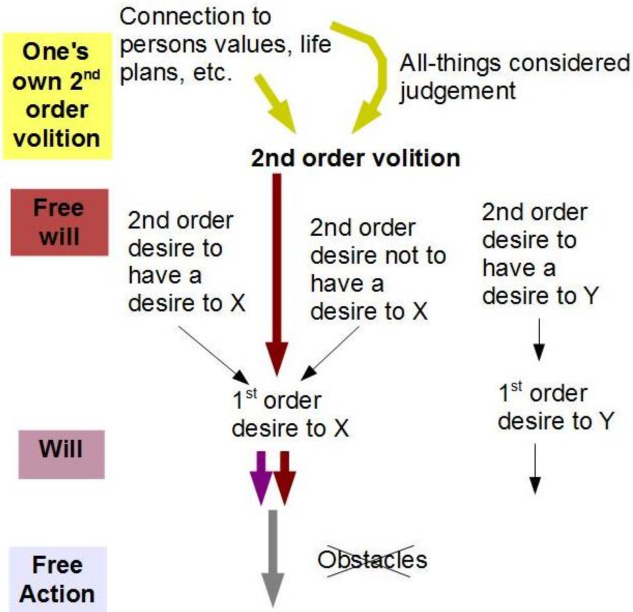
What does the God Machine affect? It could be argued that the God Machine affects free will. However, recall that in Savulescu and Persson's scenario:

> '[T]he would-be murderer never knows that her intentions have been changed by an authority outside of herself. It seems to her that she has "changed her mind" spontaneously – she experiences a life of complete freedom, though she is not free. Although any intention to kill or rape immediately changed, this was put down to the efficacy of moral education. It seemed "from the inside" that she had just developed an aversion to killing an innocent person. And no one was ever killed.' (2012a, p. 410-411)

The intervention of the God Machine could be considered as[25] *a)* a change in first order desires, *b)* a change in second order desires, *c)* a change in second order volitions by changing a volition, *d)* a change in second order volitions by rendering them ineffective, or *e)* a change in the process prior to creating the second order volition. I will argue that the last interpretation *(e)* is the most compelling.

[25] Savulescu and Persson (2012) sometimes refer to the God Machine as changing intentions. However, Frankfurt demonstrates that will is not the same as intent. "For even though someone may have a settled intention to do X, he may nonetheless do something else instead of doing X because, despite his intention, his desire to do X proves to be weaker or less effective than some conflicting desire" (Frankfurt, 1971). Since the God Machine results in behavioural change, it modifies not any intent, but rather the intent related to the effective first order desire (will).

.


One's own 2nd order volition
Free will
Will
Free Action
Connection to persons values, life plans, etc.
All-things considered judgement
2nd order volition
2nd order desire to have a desire to X
2nd order desire not to have a desire to X
2nd order desire to have a desire to Y
1st order desire to X
1st order desire to Y
~~Obstacles~~

Let us consider two versions of how the God Machine operates: the minimal intervention option and the maximal intervention option. On the maximal intervention view the God Machine can be seen as changing 1st order desire, 2nd order volition as well as something ‘up-stream’ from the second order volition. The scenario suggests some influence of a change at the roots of the second order volition first, because of the personal lack of awareness that a person changed her mind not of her own accord and second, because of the lack of repeatedly arising second order volition to commit gross immorality. In this case the freedom to have one’s own will, freedom of the will and freedom of thought (at least in so far as this translates into forming one’s own second order volitions) are all directly affected. The behavioural change follows from those changes.

On the minimal intervention view we could, firstly, interpret the intervention as only targeting the first order desire – the desire to kill and an associated intention would vanish or become too weak to be effective. This would not affect the 2nd order volition. I do not see this as a plausible interpretation because Savulescu and Persson (2012a) do not mention subjects of the God Machine interference experiencing akratically moral action – in this case an inability to act consistently with their second order volitions to make a desire to kill motivationally effective. The second interpretation would point to a change of 2nd order volition itself – from one of wanting to act on a desire to kill, to one of wanting to act on a desire not to kill. This, however, as a change on its own would be unstable, since the original 2nd order volition would be likely to arise again and again, at least for those potential offenders who acted with premeditation. Moreover, the agents described by

Savulescu and Persson (2012a) seem to have an experience of changing their mind of their own accord, while the sudden, magical change in a second order volition would be, I imagine, phenomenologically rather bewildering. Alternatively, the person might have already formed a second order volition not to kill, yet was unable to turn that into action. For those akratic individuals, improving the conformity with their second order intention would be an effective intervention. Yet, Savulescu and Persson (2012a) clearly refer to a 'changing of mind' of the person. On this interpretation there would be no 'change of mind' involved, but rather 'enforcing the ineffective will'. The third option, and what seems like the best bet to me, is that the God Machine affected the formation process of the second order volition.

The above analysis suggests that what was impacted directly was the freedom of thought and this, I think, is significant. Let us assume that the latter analysis of the minimal intervention view was correct. Technically (using Frankfurt's nomenclature), a person's free will would be preserved – as long as one can make effective the 1st order desire one wills to have. Moreover, the God Machine makes a well-informed, unbiased-by-clouding-emotions, all-things-considered judgement; the problem is that although the agent makes a choice too, it is an instance of heteronomy, not autonomy. Thus, even if the will is free and the agent has a perspective as well-considered as the God Machine, should they make an all-things-considered judgement consistent with that of the God Machine, it would not be the agent's *own* free will. The importance we assign to agents' acting on their *own* free will accounts for a significant part of our intuition about the God Machine's effect on diminishing freedom.

7.4.4. Conclusions

In this section I have argued that the God Machine affects agency on the level of creating a will of one's own. Using Frankfurt's approach, I have distinguished between free action, free will and will of one's own. On a minimal intervention view, the most likely change involves the change to something 'upstream' to second order volition and not the first-order or second-order desires.

The modification of the roots of second-order volition is indicated by the fact that a) agents are not aware that they were subject to the God Machine's intervention and, b) there is no repeatedly arising second order volition to do evil. Strictly speaking, that means that subjects' ability to act on a will of their own as well as their free will remains intact, even if the content of the motivationally effective desire has been changed as a result of changes 'up-stream' in the process.

7.5. Free thought and will of one's own

> Men fear thought as they fear nothing else on earth -- more than ruin, more even than death. Thought is subversive and revolutionary, destructive and terrible, thought is merciless to privilege, established institutions, and comfortable habits; thought is anarchic and lawless, indifferent to authority, careless of the well-tried wisdom of the ages. ... Thought is great and swift and free, the light of the world, and the chief glory of man.
>
> Bertrand Russell, *Why Men Fight*

7.5.1. Introduction

Firstly I examine the claim that what is problematic in Frankfurt-like cases is overdetermination and that the main problem lies in the foreclosure of meaningful alternative options for action (formulated most cogently in Harris, 2014a). Firstly, using a modification of a Savulescu and Persson (2012a) thought experiment nicknamed the Rational Persuader Machine, I suggest that in the God Machine scenario, it is not 'the ability to do otherwise' that protects from the loss of freedom. Rather, it is the break in the chain leading to rational agency, or, as Harris puts it, the 'divorce of thought and action' (2014a, p. 252). However, what is important is that the break happens at the level of agents' appropriate engagement with reasons. Secondly, I seek to undermine the view that the 'foreclosure of options' represents the chief problem with the God Machine. The insights of the above sections provide the basis for arguing that cases of overdetermination leave alternative possibilities untouched by Frankfurt-like cases – alternative possibilities sufficiently meaningful to ground moral responsibility.

7.5.2. Separating thought from action and the appropriate engagement with reasons

Savulescu and Persson agree with the spirit of Frankfurt's conclusion about the implications of overdetermination cases and claim that '[f]reedom of will or action is not indispensable for moral responsibility. So Harris's "freedom to fall" is not essential for moral choice and action' (Savulescu and Persson, 2012b, p. 115). Harris responds by arguing that moral responsibility cannot be separated from the actions one is responsible for:

> 'Agents are quintessentially actors; to be an agent is to be capable of action. Without agency, in this sense, decision making is, as I claimed and argue now, both morally and indeed practically barren—literally without issue! Decisions to no effect are pointless from the moral perspective; for what is a good state of mind worth, if it makes no difference to the world? At best Savulescu and Perrson can say, "Harris's 'freedom to fall' is not essential for moral choice." They cannot say, as they do, that "Harris's 'freedom to fall' is not essential for moral choice *and action*.' (2014a, p. 249)

I have argued that the way we arrive at moral action *A* is crucially important for our moral life and freedom. This importance is preserved even in the presence of foreknowledge about what action is going to be performed. To illustrate this point further consider the Rational Persuader Machine:

> **The Rational Persuader Machine.** The Great Moral Project was re-started in 2065. The most powerful, self-learning, self-developing bioquantum computer ever constructed called the God Machine has been improved. After societies became concerned about 'thought control', the Rational Persuader Machine was developed. The Rational Persuader Machine would monitor the thoughts, beliefs, desires and intentions of every human being. However, instead of modifying these

> directly, it creates the circumstances under which a person who formed an intention to commit a grossly immoral act would change their minds of their own accord. **The Rational Persuader Machine would never influence the behaviour or intentions directly but create circumstances under which a person would attend to previously ignored considerations, acquire a wider outlook, learn something about the consequences of their actions or attend to reasons they already had but which they had not given sufficient weight to.** It would never influence other people directly, but, through a network of hired staff and changes to inanimate objects, it would create the circumstances in which a person would change her mind. Sometimes it would be a closely falling brick that would make them reflect on how short life is. Sometimes it would be an anonymous email with just the right content, or a hologram of a street preacher saying just the right words. **The Machine is never a sophist; a constraint put on its action was that it is supposed to increase the person's ability (or provide the relevant information) for making an all-things-considered judgement.** Since the Rational Persuader Machine was monitoring the thoughts, beliefs and desires of everyone and had run simulations, every intervention would be effective. As before, it would only intervene once the threshold of harm to others' interest was reached. Murder of innocent people no longer occurred and everyone was given the best chance to (inevitably) change their minds.

It appears to me that interference from the Rational Persuader Machine, although still somewhat troubling (and shortly we shall explore some of the possible reasons why), would be significantly less problematic than that of the God Machine. The main difference is that there is no *direct* change of mental phenomena leading to the change of the second order volition and that, consequently, the scenario allows agents to engage with the input in a 'right way' , i.e., the agent develops a will on her own on the basis of reasons.

One could object by pointing out that the Rational Persuader Machine is indeed troubling because it is an instance of manipulation; although valid reasons or nudges to consider reasons are given, they might be interpreted as tendentious and instrumental in achieving a certain behavioural result. Such a scenario would be similar to propaganda, or to living in a society with tendentious media reporting. I admit that it is a valid consideration, yet my basic point stands: the difference between the Rational Persuader and the God Machine signals the importance of the agent's engagement with reasons.

Moreover, this scenario could be conceived of as a high-tech version of a well-known story. In Dickens' *A Christmas Carol*, Ebenezer Scrooge is visited by four ghosts: Marley, and the Ghosts of Christmas Past, Present and Yet to Come. Shown images of the past, present and future and warned by the ghosts, Scrooge 'changes his mind' and re-shapes his life. Although the ghosts' intervention could be considered as leading to a desired outcome, while at the same time the images brought to Scrooge's attention seen as tendentious, ultimately Scrooge seems to take into account more of the morally relevant considerations than previously. By this reading, his ability to make an all-things-considered judgement has increased, even if he remains an imperfect moral agent.[26]

The potentially relevant difference between Scrooge and the Rational Persuader Machine scenario is that in the latter case the agent does not know that another will is involved in shaping the information they can access – that there is a purposeful activity shaping their environment. Since the intervention is done in order to achieve a certain outcome, the whole exercise may be seen as manipulation. Moreover, since the person is unaware of the intervention they cannot take that factor into account. I think that this intuition arises from the fact that in our every-day context we have good reasons to distrust information from people who have a vested interest in us doing a certain thing – often, the information may be

[26] For the purpose of this argument it is sufficient to contend that Scrooge is somewhat *better* at making an all-considered judgement from a moral perspective, despite the questions about the reasons for him adopting the moral perspective and how good of a moral agent he is after the ghost's appearance.

misleading or unreliable. Yet, in our rather removed- from-the-real-world scenario, the constraint on the Machine's interference is that agent's ability to make an all-things-considered judgement increases. Thus, this last objection turns out to be weak.

As a result, I think that it is unimportant that the Rational Persuader Machine provides necessarily compelling reasons, as long as the agent is able to engage with them in a right way. Moreover, although for both the machines the outcome is known from the start, the fact that the Rational Persuader Machine is less problematic than the God Machine signals that freedom of action is not the only, and perhaps not even the main, consideration. It appears that it is not the ability to do otherwise – in the sense of having possible futures in which one commits a gross immorality – that is of utmost importance. Rather, it is the appropriate engagement of the agent with reasons that is important. This is not a novel point (e.g. DeGrazia, 2014; Harris, 2014a), and put this way it may seem uncontroversial. Compatibilist accounts of free will often include emphasis on the appropriate engagement with moral reasons: what matters is that there is an appropriate causal connection between agent's actions and preferences. A reason-responsive mechanism guiding action seems to be an important component of that account (for more details see: section 7.7.1). Consequently, if there are necessarily compelling reasons given, it does not necessarily mean that the freedom of will is diminished, even if we could not have done otherwise based on the reasons given.

Here we can note a certain asymmetry in interpretation of the God Machine and Rational Persuader Machine. Savulescu and Persson interpret the meaning of the God Machine's intervention as follows:

> 'We have argued that there might be interventions, such as the God Machine, that do indeed produce more moral behaviour that do control the moral agent, subjugating that person to the will of another and removing the freedom to act immorally. Such interventions and such control are not plausibly moral enhancements of that person – they rather undermine autonomy by substituting moral for immoral intentions.' (2012a, p. 417)

While in the God Machine scenario, the changes seem to undermine free will and free action (as interpreted by various commentators (e.g. Harris, 2014a; Savulescu and Persson, 2012a), this does not seem to apply to the Rational Persuader scenario, despite the fact that the agent cannot do otherwise. For the purpose of the outcome of action, the agent could just as well be locked in Locke's room.[27] This does not, however, seem to be as problematic in the Rational Persuader Machine scenario. As a result, it seems that the crucial factor in the assessment of the God Machine is not the undermining of the freedom of will or free action – the main problem is the inability to form a will of *one's own* (i.e. an authenticity concern) and the lack of freedom of thought.

A further argument for the importance of a will of one's own, as opposed to freedom of will and action, is the analogy between the God Machine and akrasia (see: Rorty, 1980; Mele, 1982; Pears, 1998). More specifically, there is an analogy between the influence of the God Machine and the examples of akrasia that are not instances of being weak willed. In both the God Machine scenario and the (not weak-willed) akrasia, it is important whether or not the second order volition arose in the right way. Davidson describes akrasia without weakness of will and argues that when people act against their better judgement they temporarily believe that the worse course of action is better, because they have not made an all-things-considered judgment. They have merely made a judgment based on a subset of relevant considerations. This is compatible with Holton's (1999) point that weakness of the will involves revising one's resolutions too easily. Under this view, it is possible to act against one's better judgment (that is, to be akratic) without being weak-willed. How does the analogy apply to the God Machine scenario?

Suppose, for example that, Wilson Sr. judges that taking revenge upon a murderer is not the best course of action, but makes a resolution to take revenge anyway and sticks to that resolution. Such a description of some people's seemingly irrational actions has a high degree of psychological plausibility. I agree with Holton's interpretation that Wilson Sr. behaves akratically but does not show weakness of

[27] See: Locke 1689, Book 2, Chapter 21, Section 12.

the will. I think that the God Machine's interference is best understood in similar terms: not as impairment primarily in the exercise of free will, but rather as an instance of externally enacted (heteronomous) akrasia without being weak-willed. The most important 'break' is not that between thought and action, as Harris (2014a) suggested, but between one's own thoughts and beliefs and resultant effective will.

7.5.3. Separating thought from action and overdetermination

The God Machine is a case of overdetermination. Thus, one could ask about the extent to which the lack of ability to do otherwise in close counterfactual worlds should be the main concern related to the impairment in freedom. In discussing the Rational Persuader Machine, I have argued that overdetermination and the resultant foreknowledge about counterfactuals is less of a problem than other aspects of the God Machine. To consider this point further let us look at the following example of overdetermination:

> **Hall's Brain Implant.**[28] Hall is a convict with a history of violent outbursts and PTSD. He received several prison sentences after reacting with rage and being physically violent after a slight provocation. He has a device implanted that monitors the activity of his brain and would interrupt the activity in cingulate[29] when the implant detects brain markers associated with a high level of fear or rage. He has made an all-things-considered judgement to have the cingulate implant, as he does not endorse his uncontrollable fits of rage on any basis (moral, prudential or epistemic). There are some side effects, but they are all-things-considered acceptable to Hall. The device can be switched off completely but it requires an advance appointment

[28] The development of this case has been inspired by a case of a patient with a brain injury, who despite his best efforts to stay 'on track', faced repeated incarceration after violent outbursts. An analogical case, although more rhetorical in its use of sexual urges leading to offence, is presented in DeGrazia (2014).

[29] Although psychosurgery involving creating lesions in the brain is rare, a small number of patients with intractable aggression are currently subjected to cingulatomy or capsulotomy. See for example: Jiménez-Ponce et al. (2011). DBS has been proposed as a treatment for violent outbursts (Maley et al. 2010).

> with a doctor.[30] Where previously Hall would go into rage, he now does not. In triggering situations, he can feel increasing anger but at some point the device kicks in and the anger diminishes. Often but not always he notices when the device kicks in. The device also has a display that allows him to monitor its activity. However, Hall is free to employ other emotion modulation mechanisms and is learning to regulate his emotions and actions earlier on: by calming himself down, reappraising when he notices his anger rising, noticing triggers and exiting the triggering situation, etc. Sometimes he is successful in regulating his emotions and averting behavioural impacts, sometimes the device kicks in. He can practice self-control with the assurance that he will not harm anyone in the process, as there is a 'safety net' if his self-control fails.

This scenario appears to me to be significantly less problematic than the God Machine scenario. Overdetermination undermining a person's ability to do otherwise seems not to be the main concern here – the difference between those cases does not lie in the ability to do otherwise in any particular situation. Moreover, the action of the device is direct in the sense of acting directly on the brain.

The first difference lies in the fact that the influence of the brain implant is not beyond moral review in the sense of offline reflection and control, although it is beyond the agent's control at the moment of acting. The second difference is that it aids the agent in making the connection between her all-things-considered judgement and actions, although this happens in a somewhat more roundabout way than if Hall exercised self-control. The third difference (which I would like to highlight here) is that the agent's will is not heteronomous to the same extent as in

[30] This condition found its way to the example partly to accommodate a concern raised by Harris in the context of the God Machine related to 'agreeing to enslave oneself,' i.e., agreeing to the diminishment of freedom that cannot be revoked, and partly with an eye on the current practice of DBS in medical contexts, in which the device can be switched off at any time by the user. The criminal justice-related and impulsivity related use would likely offer the ability to switch the device off but not at any point.

the God Machine scenario. Thus, it is not overdetermination (and thus the inability to do otherwise) that is a main issue here.

7.5.4. What is a good state of mind worth?

> Agent Smith: You must be able to see it, Mr. Anderson. You must be able to know it by now. You can't win. It's pointless to keep fighting. Why, Mr. Anderson? Why? Why do you persist?
> Neo: Because I choose to.
>
> *The Matrix Revolutions*

Hall's Brain Implant and the God Machine scenario question the value of a certain state of mind if the actions remain the same – a question raised by Harris (2014a) in *How Narrow the Straight*. Provided that the device functions properly and the neural markers for violent outburst are established with sufficient accuracy in Hall's Brain Implant scenario, whether he exercises self-control or the device acts, the result 'in the outside world' will remain the same. Savulescu and Persson agree with the spirit of Frankfurt's conclusion about the implications of overdetermination cases and claim that '[f]reedom of will or action is not indispensable for moral responsibility' (2012b, p. 115). Recall Harris' response:

> Agents are quintessentially actors; to be an agent is to be capable of action. Without agency, in this sense, decision making is, as I claimed and argue now, both morally and indeed practically barren—literally without issue! Decisions to no effect are pointless from the moral perspective; for what is a good state of mind worth, if it makes no difference to the world?' (2014a, p. 249)

How does this bear on Hall's Brain Implant case and the God Machine? It appears to follow from Harris' position that if there is no difference in the resultant action, there is no moral difference between Hall *a)* exercising self-control, *b)* acting under the influence of the cingulate brain implant, *c)* refraining from harming someone had he been plugged into the God Machine with his knowledge and consent and, *d)* being plugged into the God Machine without his knowledge or consent. I think that such conclusions would be misguided.

Despite the same outcome, the way that the decision is made influences (and I think rightly so) the ascription of moral responsibility. Two actions that appear the same ‘in the world’ acquire a different sense depending on the intentions behind them, whether the action was deliberate and on the context. The fact that unintentional killing is not murder, benefiting someone accidentally is not helping, and making inadvertent mistakes that others subsequently learn from does not amount to teaching *sensu stricto,* demonstrates that the same action may have different moral meanings. In this sense, actions and intentions are joined, and the intention is relevant to the judgement about the character of action.

This link in interpreting what the action consists of transfers to issues of moral responsibility. Thus, the actions that agents are responsible for are not merely actions understood as occurrences in the external world with no regard to agent-causation and intentions. If I have harmed someone while trying to help, I may be guilty of negligence, stupidity or harmful ineffectiveness, but not an act of malice. Consider a hypothetical case in which I have harmed someone while trying to help. Assuming that my action was justified to the best of my knowledge, that I have not been negligent and that the help backfired through no fault of my own, I am not morally responsible for the outcome – although I may be responsible causally.

This line of argument leads to the conclusion that there is a difference between actions that have the same consequences in the world, if the intentions differ – at least for the purposes of moral responsibility. Take I1, I2 to be intentions and A1 and A2 actions in two possible worlds. For the purposes of moral responsibility, A1 does not equal A2 where I1 leads to A1 and I2 leads to A2, even if the external observer cannot discern the difference between A1 and A2. In Frankfurt cases, Jones’s murder of Smith is a different event depending on whether it comes about through Jones acting on his own accord or through the intervention of the Black’s mechanism. Van Inwagen (1978) presents the same argument, referring to intention-action composites as event-particulars and general behavioural outcomes as events-universals, and arguing that we are responsible for event-particulars.[31] A somewhat similar point presented in relation to obligation in Frankfurt cases and immediately applicable to considering the PAP was presented by Rowe (1987, pp. 43-64). Rowe draws on the notion of agent-causation and argues that there is a relevant alternative possibility in Frankfurt cases. An alternative to Jones's agent-causing the murder is Jones's not agent-causing the murder. He further argues that this is precisely the kind of alternate possibility that is relevant to Jones's moral responsibility. Jones' obligation not to agent-cause a volition to murder Smith is discharged when the mechanism pre-empts his powers of agent-causation. Since this alternative is available to him, he is morally responsible for murdering Smith under the terms of PAP.[32] [33] For the purposes of moral responsibility assessment, it makes sense to consider the action caused by the God Machine, the action caused by Hall's brain implant and the action resulting from Hall's self-control as relevantly different actions. Thus, a state of mind is worth quite a lot in moral responsibility currency, and it is relevant not only for freedom of choice (as Harris argues) but moral action properly so called.

Moreover, in thinking about 'what is the good state of mind worth if it makes no difference in the world', the discussion seemed to overlook an important difference – in the agent. Consider Joe, who lives in an oppressive authoritarian regime. He does not resist the regime in any active way, not wanting to endanger his family. Does it make a difference in the world that Joe complies with the Party's policies, disagreeing with them but judging that it is better not to resist? Even if we may argue that Joe should have done something to act against the political system and judge it as a moral failure that he does not do so, I think that it still makes a difference whether he complies willingly or unwillingly. Even if it does not matter for anyone else, it makes a difference for, and to, the agent. Consequently, even if outcomes seem to be the same in the outside world, they are not – agents themselves are not somehow separate from the world and their inner life, the

[31] For a critical discussion see: Hunt (2000).

[32] A similar argument is presented by Naylor (1984, pp. 249-258). For a similar argument in the context of deontic ethics see: Speak, D. (2002).

[33] Admittedly the following argument is subject to objections and the issue of the importance of 'flickers of freedom' is contentious. For the purpose of the current argument, however, it is sufficient to outline the importance of intentions and actions for moral responsibility ascription and moral action

meaning they make of things and the attitudes they take to events makes a difference *in the world*. Perhaps that difference is small in the grand scheme of the utilitarian calculus of all the souls (although do not tell it to poor Joe), nevertheless, it is a morally relevant one.[34] Not counting that difference at all is to treat the agent as if he or she was not there.

One could argue that perhaps it would be better for an agent not to resist and suffer inner struggle. That perhaps depends on individual psychology and the weight given in a moral calculus to well-being vs. maintaining one's values even if one decides not to act on them. Whatever Joe's outcome of that calculation, it is largely irrelevant for the point I am trying to make: that the difference in an agent is morally significant and sufficiently so to make for morally significant alternatives.

An objection could be made that this account stretches the notion of 'moral action' too far. However, in *Moral Enhancement and Freedom*, Harris eloquently argued that what matters (and should matter) in moral enhancement is not only whether people do good, but that they do good from the moral perspective. An event-universal, to use Rowe's language, is either a merely morally relevant action or a moral action, depending on what kind of event-particular it is an instance of. Here I wish to extend this point, and argue that it not only matters that a good action is done from a moral perspective, but that the mechanics of the agent's choices is also important.

7.5.6. Conclusions

On the basis of the Rational Persuader Machine scenario and Hall's Brain Implant case I have argued that the main problem with the God Machine lies not in the fact that the agent lacks meaningful alternative event-universals (overt actions) but rather the lack of the appropriate engagement or control over the influence. Further, taking into consideration the relevant distinction between events-universals and

[34] Another argument can be that the value of a good state of mind comes from the fact that although different intentions may not make a difference in that particular instance but they may impact behaviour in other instances. Here I wish to argue for a stronger position: that the difference in the agent is a morally relevant difference in the world *regardless* of impacts in other instances.

event-particulars, I have proposed that overdetermination cases preserve ‘the ability to do otherwise’ to a sufficient degree to ground moral responsibility. Finally, I have argued that freedom of choice even in the absence of a change in subsequent event-universals makes a morally important difference.

The decoupling of thought and action is not the only problem with the God Machine. What is also problematic about the God Machine is the conjoined diminishment of freedom of thought – even though the agent acts on his free will, it is not free will of his own – and the fact that its influence does not even give the agent a chance to engage with that influence. In my view, those considerations mark a significant difference between the God Machine and imprisonment and propaganda, and the God Machine is much more problematic by comparison. As Harris pointed out in *How Narrow the Straight*, ‘… this is a different level of unfreedom. As with all actions, when we are free, we are only free to do as we like and take the consequences’ (2014a, p. 254).

I am not trying to suggest that freedom of action does not matter at all or that freedom of thought is all that matters. As Harris (2014a) argues, the freedom to act and to make a difference in the world is important for political freedom, self-governance and responsibility understood as accountability for how our actions shape ourselves and the world. I rather want to outline what I consider to be the best interpretation of the God Machine’s influence, an interpretation that elucidates the reasons why the God Machine is *especially* problematic for freedom – significantly more so than living under the rule of law and perhaps even more so than in a despotic regime that minds its own business with regards to the mind.

7.6. Freedom and domination

7.6.1. Introduction

I have argued that a great part of our intuition about the impact of the God Machine on freedom has to do with the inability of the agent, when subject to the God Machine’s intervention, to form the will of their own. Others (Harris 2014a, 2014b; Sparrow, 2014) highlight another reason why the God Machine scenario is troubling: the interference of the state or a powerful state-like entity. In this section, I argue that Sparrow's (2014) argument against the God Machine based on Pettit's (1997) account of non-domination is not convincing.

7.6.2. The God Machine impersonated

Sparrow argues that 'Savulescu and Persson … underestimate the tension between the power of some and the freedom of others' (2014, p. 27). Sparrow attempts to use the concept of freedom as non-domination to argue that the God Machine undermines the political freedom of the members of the God Machine society. He invites us to consider Pettit's (1997) hypothetical case of a slave in the power of a benevolent master:

> 'If he wanted to, this slave-owner could intervene in every part of his slave's life and thwart all their plans and projects. Yet because he happens to be (for the moment, at least) benevolent, he refrains from exercising his power at all and permits his slave to go about their life unconstrained. Pettit points out that we have a strong intuition that slaves ruled over by such a master are not free because they are subject to his power — regardless of whether or not he exercises it.' (Pettit, cf. Sparrow 2014 p. 27)

For Sparrow, the application of this case to the ethics of the God Machine scenario is 'obvious' (2014, p. 27). Although the God Machine only acts to alter an individual's motivations for a narrow subset of intentions – intentions to commit seriously immoral acts – the same power could be exercised to control individuals' other motivations. Hence, Sparrow concludes, the God Machine '"dominates" its subjects' (2014, p.27).

However, the mere possibility of someone's interference with the exercise of my freedom is not sufficient for me to say that I live under their domination, at least not according to the non-domination conception of freedom we are talking about. If it was, than the mere possibility of me killing any person I meet would mean that I dominate everyone I meet (and since the reverse is also true, that everyone I meet dominates me).[35] Analogically, to say that the God Machine 'dominates its subjects' is an overstatement and seems to be missing the point – the point made by those putting forward the conception of freedom as non-domination.

The political idea of freedom as non-domination has been developed to describe certain kinds of structural relations and to account for perceived limitations in the interference view of freedom. The actual exercise of power is not necessary as in the case of the non-interference view, and in this sense the mere possibility is sufficient. However, the non-domination view makes reference to the broader configuration of laws, institutions, and norms that *effectively allow* masters to treat their slaves however they please. Moreover, it is not simply 'power' but rather uncontrolled or arbitrary power that is at issue here (Skinner 1998, 2008; Pettit, 1997). In the example referred to by Sparrow, the slave master who has the institutionally-unrestrained freedom to treat his slaves more or less as he pleases or whose power remains 'unchecked', can be said to 'dominate' his slaves. In contrast, a slave who has the practical ability to kill his master cannot be said to 'dominate' his master. Like non-interference, non-domination comes in degrees: in the republican view of freedom, one is not either free or unfree, but rather more or less free depending on the extent of non-domination one securely enjoys. Citizens in the God Machine society are unfree only to the extent that they are *structurally vulnerable* to the exercise of *arbitrary* power.

Is the God Machine a master over the citizens? Contrary to Sparrow's claim, the application of the non-domination view of freedom to the God Machine is far from straight forward. In so far as domination is seen as the presence of a structural relation in which there is an arbitrary power of the God Machine over citizens, the God Machine does not 'dominate' in the Savulescu and Persson (2012) scenario for two reasons. First, in Savulescu and Persson's (2012a) description of the scenario its power is neither 'arbitrary' nor 'unchecked' (although the latter is less clear in Savulescu and Persson's setup, as I will soon demonstrate). Second, it is a stretch to

[35] This is the difference between the use of 'domination' in which anyone can be dominated by another or a group (See Hobbs' discussion of the status of woman in Leviathan) and the more narrow technical use of 'domination' in the republican conception of freedom as non-domination, which typically refers to a structural relation in which one is vulnerable to others' exercise of arbitrary or unchecked power

consider the God Machine as the kind of entity that can dominate in the sense required by Pettit's non-domination view. A more promising critique rests on the same kind of scrutiny that other instruments of the state would merit and so the God Machine should be seen as the analogy for the law.

7.6.3. Arbitrary and unchecked power

The understanding of what arbitrary power consists of merits some explication. There are two broad approaches to that question. The first approach is to define 'arbitrariness' substantively. According to this approach, power is arbitrary when it fails to track what Pettit called the 'welfare and world-view' (1997, p. 56) of those affected. This substantive definition is open to at least three interpretations, depending on whether we interpret the 'welfare and world-view' of those affected as a) their objectively-defined interests, b) their subjective preferences, or c) their shared ideas as expressed through an appropriate deliberative process. This approach has sparked a complex discussion on the difference between freedom and the common good and the possibility of collapsing them into each other, which I will not examine here (see for example: Larmore, 2004; Costa, 2007; Carter, 2008 for discussion). For now it suffices to say that in the God Machine scenario as stated by Savulescu and Persson (2012a), there is no obvious reason to think that those conditions are not fulfilled.

The second broad approach is procedural. To move away from the difficulties brought on by the discussion of the definition of the 'arbitrary' in relation to the common good, Pettit has proposed abandoning the notion of arbitrary power and focusing instead on whether the power is controlled or uncontrolled (2012, p. 58). On this view, power is arbitrary or unchecked to the extent that it is not reliably constrained by effective rules, procedures, or goals that are common knowledge to all persons or groups concerned (Lovett, 2001, 2010). The appropriate 'check' may come in the shape of control by the persons specifically affected (the democratic view), or control by the society's laws, norms and institutions (the procedural view).

According to Lovett (2012), it is important that the constraints put on the exercise of power are ‘reliably effective.’ In order to be reliably effective, constraints must remain effective over a wide range of possible changes or modifications in the relevant circumstances. In a similar context, Pettit proposed ‘non-manipulability’ as a criterion institutional ‘instruments’ should satisfy:

> ‘Designed to further certain public ends, they should be maximally resistant to being deployed on an arbitrary, perhaps sectional, basis. No one individual or group should have discretion in how the instruments are used... The institutions and initiatives involved should not allow of manipulation at anyone’s individual whim.’ (Pettit, 1997, p. 173)

He then lays out three conditions that a ‘non-manipulable’ system must satisfy. This includes firstly the rule of law, according to which laws ‘should be general and apply to everyone, including the legislators themselves; they should be promulgated and made known in advance to those to whom they apply; they should be intelligible, consistent, and not subject to constant change; and so on’ (Pettit, 1997, p. 174). Secondly, the dispersion-of-power condition requires that ‘powers which officials have under any regime of law should be dispersed’ by mechanisms such as the separation of powers, bi-cameralism, federalism, and international legalism (Pettit, 1997, pp. 177–80). The third condition states that laws should be insulated from ‘excessively easy majoritarian change’ (Pettit, 1997, p. 180).

Does the God Machine satisfy these conditions? There is no reason why it should not be possible. Taking Persson and Savulescu’s (2012a) scenario at face value, there is no reason to suspect that the God Machine society would necessarily fail to institute such protections. Even systems that do what is beyond the ability of one human to do can be reviewed and controlled – for example, the assumptions on which the computations were made can be accessible to further analysis by computer systems independent of the God Machine, predictions generated by the God Machine can be tested and the system’s workings subject to regular audit by a group of auditors immune to the machine’s influence, every intervention of the God Machine might have to be reviewed by an independent computer system and a human, etc.[36] The second question would be about whether the God Machine is vulnerable to abuse by others. This question can only be answered with reference to the political context that the God Machine is placed in. To the extent that this is presented by the authors, one can suspect that in a morally (and politically?) enhanced society appropriate checks on the access to the God Machine will be placed, and the representative system characterised by legitimacy and transparency is likely to be stable.

The most convincing argument in this context (Harris, 2014a) comes from the fact that the God Machine society's citizens would not know when the lawful intervention would take effect, and thus have no opportunity to question, respond to or challenge this intervention. The God Machine would therefore lack an important feature that allows for the protection of liberty. Overall however, if we take the God Machine at face value and thus accept the facts about the possible world it was placed in, the arguments that attempt to establish that the God Machine powers are beyond societies' review largely fails.

The God Machine is not necessarily inimical to appropriate control, at least as presented by Savulescu and Persson (2012a). As a result, Sparrow's account does not support either plausible interpretation of freedom from domination as the right kind of structural relation between the God Machine and the citizens – the God Machine does not obviously fail the 'checks and balances' requirement and it is not clear that the God Machine would serve to establish a relation of domination between some members of the God Machine society and others. A further argument may develop the in-principle problems of auditing and controlling an entity like the God Machine, but in the absence of such an argument, the objection fails.

7.6.4. Ghost in the God Machine: Who is dominating?

> Action without a name, a 'who' attached to it, is meaningless.
>
> Arendt, 1958, *The Human Condition*

[36] We'd assume that the enhanced society is not blind to Clarke's HAL-9000 lessons.

Max Weber held that the existence of domination requires ‘the actual presence of one person successfully issuing orders to others’ (1968, p. 53). On such a narrow conception, the pertinent question will not be about the domination of the God Machine, but rather who the God Machine serves and who controls it. Broader conceptions of domination may not require a specific agent ‘issuing orders.’ For example, domination may arise from the inadvertent and unconscious actions of agents as a by-product of social and economic forces (e.g. Shapiro, 2012). However, even here it is not the social and economic forces that are dominating – as domination presupposes a degree of agency that social, economic and natural forces do not have.

The extent to which the notion of agency is central to the political idea of freedom as non-domination is already indicated in Pettit’s explication of the idea of domination. In Pettit’s words, an agent is dominating when an agent has:

‘1. the capacity to interfere
2. with impunity and at will
3. in certain choices that the other is in a position to make’ (Pettit 1997, pp. 578-581).

To interfere in this sense ‘with impunity’ is to do so without ‘penalty,’ be it resistance by the victim or punishment by some external authority (Pettit, 1997, p. 580). To interfere ‘at will’ is, according to Pettit, to do so at one’s own pleasure or whim. In other words, the interferer has the necessary ‘discretion’ to act as he or she chooses (Pettit, 1997, pp. 580-587). Does the God Machine have a capacity to interfere at one’s own pleasure or whim? I very much doubt that Savulescu and Persson’s (2012a) God Machine experiences much pleasure or is capable of whims in any sense stronger than in an anthropomorphising metaphorical sense, similar to the way in which we are subject to ‘the whims of Nature.’ Similarly, I very much doubt that the ‘penalties’ Pettit refers to would have much impact on the God Machine – for a simple and sufficient reason that the God Machine is not the kind of agent that non-domination freedom theories refer to.

On the non-domination conception of freedom, it is not the ‘laws of slavery’ that dominate the slaves but rather slave owners, who are effectively allowed to do so by the absence of effective laws that would curtail their arbitrary power. Similarly, to the extent that the God Machine is an analogy for state power and law, it is not the God Machine who dominates the citizens of society. In a more technologically-fuelled example, drones flying the sky above the Afghanistan-Pakistan border do not dominate the local people. It is those who steer the drones, those who establish targets, those whose power the drones enable, further and protect.[37]

But perhaps the God Machine is better seen as the analogy for the state. What are we to make of the state, ontologically? According to Hegel, the state was the ‘Divine Idea on Earth’ (1837, p.39). Hobbes (1651) used the metaphor of an ‘Artificial Man’. Nietzsche declared it the ‘coldest of all cold monsters’ (1883, p. 160), although we would be hard pressed to take Nietzsche’s statement in his highly poetic work literally. John of Salisbury (1159) defines the republic as a ‘certain body’ and takes his anatomical metaphor rather far, perhaps sheltered by his reference to Plutarch:[38]

> The place of the head in the body of the commonwealth is filled by the prince, who is subject only to God and to those who exercise His office and represent Him on earth, even as in the human body the head is quickened and governed by the soul. The place of the heart is filled by the Senate, from which proceeds the initiation of good works and ill. The duties of eyes, ears, and tongue are claimed by the judges and the governors of provinces. Officials and soldiers correspond to the hands. Those who always attend upon the prince are likened to the sides. Financial officers and keepers[1] (I speak now not of those who are in charge of the prisons, but of those who are keepers of the privy chest) may be compared with the stomach and intestines, which, if they become congested through excessive avidity, and

[37] I do not wish to make claims about the arbitrariness of the military intervention in Afghanistan or in Pakistan. The presence of drones is a good illustration of the agents-instruments distinction, on which I wish to focus. For the purpose of this argument assume that the intervention would not fulfil the conditions necessary for non-arbitrariness or appropriate control.

[38] Which was most likely what scholars have kindly described as literary device, see: Canning (1996, p. 112).

> retain too tenaciously their accumulations, generate innumerable and incurable diseases, so that through their ailment the whole body is threatened with destruction. The husbandmen correspond to the feet, which always cleave to the soil, and need the more especially the care and foresight of the head, since while they walk upon the earth doing service with their bodies, they meet the more often with stones of stumbling, and therefore deserve aid and protection all the more justly since it is they who raise, sustain, and move forward the weight of the entire body. Take away the support of the feet from the strongest body, and it cannot move forward by its own power, but must creep painfully and shamefully on its hands, or else be moved by means of brute animals.

However, for Wendt, a political scientist and a social constructivist international relations scholar, the state *is* a person. Wendt argues that it is not that the state 'is like' a person, it literally is a person: 'states are people too' (1999, p. 215). It is understandable that the problem of defining state agency emerged in the field of international relations with considerable force, where treating states as agents conferred descriptive ease yet influenced the kind of descriptions, explanations and predictions scholars would make (Wendt, 1987). In positivist explanations of the relation between the citizen and the state, personification of the state was treated as a useful metaphor – it was understood as an instrumental device aimed at facilitating explanation and implied no ontological commitment to the state actually possessing any of the properties assigned to it. To put it in the words of Gilpin, when we talk of 'the state acting,' we engage in a collective illusion (1986, p. 318): we all know that the state does not really act and we also know that in reality there is no such thing as a state.

There are two separate questions here. First is the question of ontology: whether or not the state exists; is the state real, is it a fiction, or is it a theoretical abstraction? The second question concerns the kind of properties that make sense to ascribe to states (and, in our discussion, the God Machine): is a state (and the God Machine) an agent, or, to put it more strongly, a kind of agent that can dominate in a sense that is relevant to freedom by non-domination?[39] I will briefly consider the second question as more relevant to the issue at hand.

What kind of agency is necessary for Sparrow's objection to succeed? To answer that question let us look closer at the notion of agency accommodated by Pettit's (1997) theory. Petitt himself attempts to apply freedom as non-domination to international relations and relations between state and non-state entities (Kukathas and Pettit 1991; Pettit 2010). In response to the agent/structure problem, he explicitly falls on the 'agent' side of the distinction:

> 'while a dominating party will always be an agent – it cannot just be a system or a network or whatever – it may be a personal or a corporate or collective agent' (Pettit, 1997, p. 52).

He recognizes the challenge of normative individualism, which he understands as a position that holds that 'there can be no difference in the value of two institutional arrangements unless there is a difference in the value for individual human beings of those arrangements' (Pettit, 2010). His justification for the extension of the non-domination view of freedom to the business of what he calls 'agencies' (such as states, corporations and non-governmental organisations) and its normative importance, however, explicitly rests on the idea that the agency of collectives is rooted in the agency of the people that constitute them:

> 'the domination of corporate agencies will matter insofar as those agencies are organizations whereby individual human beings combine to act together. If the things that the members do as a corporate entity are subject to the alien control of another agent or agency, then those members are themselves subject to alien control.' (2010, p. 76)

Pettit's (1997) theory is built on a conception of domination as a relation between persons, or groups of agents that are capable of exhibiting collective intentionality. He proposes that a *mens rea* condition is a logically necessary feature of domination: 'the worsening that interference involves always has to be more or less

[39] Wendt himself seems to equate the agency with personhood, a move understandably opposed by other scholars: e.g. Waever (1994).

intentional in character: it cannot occur by accident' (1997, p. 52). There is a strong political purpose for such a concept of 'domination': to call the powerful to account, so that the arbitrary interference that relations of domination enable must be the sort of thing for which the dominating agent can be held responsible (Pettit, 2005, p. 93).

The God Machine is not an 'agent' or an 'agency' in Pettit's understanding of that term, at least not in a straight forward way. Citizens of the God Machine society are not 'members' of the God Machine just like people are not 'members of the criminal law', and so even if we extend the view of agency to organisations, it does not help Sparrow's (2014) argument. In fact, we have stumbled upon a serious limitation of Pettit's (1997, 2001) theory of freedom as non-domination – the limitation related to his view of agency in general and his view of the kind of agency that states enjoy specifically.[40]

Further, Pettit's collective agents are rather odd creatures. On the one hand, Pettit says that collective agents are 'candidates for freedom' insofar as they 'have the capacity to function in their own right as free and responsible agents' (2001, pp. 115-123). On the other hand, Pettit argues that 'it would represent a bizarre normative position to think that [collective agents'] freedom as discursive control mattered in itself, and not just in virtue of the correlated freedom that individuals may enjoy' (2001, p. 126-127). This conclusion is predicated on the fact that collective subjects have a somewhat different status from individual subjects: they 'come into existence in order to serve the interest of individuals' (2001, p. 126). It is worth quoting Pettit's account of the distinctiveness of collective agents at length:

> Although social integrates have to be ascribed personality in the same way as individual human beings, it is worth emphasizing that such collective subjects differ from individuals in as many ways as they resemble them. They are not centres of perception or

[40] The republican notion of agency has received little explicit attention in the debate on freedom as non-domination. One exception is Markell (2008), whose argument about the insufficiency of the ideal of non-domination goes back to questions about agency. Another is a critique put forward by Rigstadt (2011) who argued for a structuralist approach to republican freedom as non-domination.

> memory or sentience. They form their collective minds only on a restricted range of matters, to do with whatever purpose they are organized to advance. And they are artificial creatures whose responses may be governed by reason, not in the spontaneous manner that is characteristic of individual human beings, but only in a painstaking fashion. Their reasoning may be as tortuous as that of the impaired human being who has to work out reflectively, case by case, that in virtue of believing that p and that if p then q, he or she ought also to believe that q. While integrated collectivities are persons and selves in virtue of being conversable and responsible centres of judgment, intention, and action, then, they are persons and selves of a bloodless, bounded, and crudely robotic variety. The most natural way to think of them is as agents to which individuals give life by suspending their own projects, now on this occasion, now on that, in order to serve the collective point of view. (Pettit, 2001, pp. 118–119)

This rather unclear view of agents creates problems for the application of a non-domination view of freedom and raises questions about the relation between full-blown agents, such as people, and agencies (Markell, 2008). I do not aim to develop a full critique of Pettit's view of agency; it suffices to say that the problems for this view arise exactly in considering the relation of individual agents to agencies. Moreover, the examination of Pettit's view of agencies makes one thing clear – the God Machine is not an agent on Pettit's view and thus Sparrow's critique rests on a theory that struggles with conceptualising the issue he wishes to examine.

A serious limitation of Pettit's theory resting on a strictly agential definition of domination is exactly that it is inattentive to the possibility of threats to freedom stemming from impersonal forces: whether systemic or 'agential' but originating in agents who cannot be held responsible and are not 'conversable'. Pettit's focus on agents as opposed to structures is not the only view available in the republican discussion of non-domination. For example, for Hobhouse, the question of how best to minimize arbitrary power should be answered by examining how social structures or systems yield hierarchical or anti-hierarchical effects. Hobhouse

(1911) characterizes the antithesis of liberal freedom as a 'system of rightlessness.' This insight is not new; La Boetie argued in the sixteenth century that 'the mainspring and the secret of domination, the support and foundation of tyranny' is always a hierarchically structured system of patronage (1576, p. 77). Other proponents of structural views of domination include, for example, Bohman who is especially concerned about the global proliferation of displaced people, refugees, asylum seekers and illegal immigrants, because in the dislocation of people from the places in which they can engage in politically effective speech and action, 'globalization has had effects that are structurally similar to modern tyranny' (2007, p. 342). Bohman's view is reminiscent of Hobhouse in that it is concerned about the structurally enacted 'state of rightlessless' that is important, even in the absence of an obvious tyrant. Similarly, Laborde states that on her view 'domination refers not only to interpersonal relationships but to basic, systemic power structures' (2010, p. 54). Given the rich tradition of views more suitable to analysis of the God Machine, it remains a mystery why Sparrow (2014) chose to base his critique on a Pettit's approach.

In so far as the God Machine is the analogy for and a replacement of the criminal law system and, more widely, the power of the state, the God Machine is not an agent, but an instrument that regulates relations between individuals. Thus, Sparrow's (2014) supposedly obvious application of freedom as non-domination rests on a fundamental confusion of *structures such as laws* with *agents allowed to dominate other agents by structures such as laws;* or, to state it another way: *instruments* with *those whose those instruments serve*.

If the God Machine is treated as an instrument as I have argued it should be, the appropriate analysis on the basis of non-domination is twofold. Firstly, it is to examine the context in which its exercise furthers or impedes the freedom as non-domination. Thus, the first set of questions that one should ask include who are the agents behind the God Machine's actions, who are the agents that make and shape the God Machine and what does the God Machine allow others to do should the conditions change? The second set of questions would refer to the features of the instrument itself, and here the analysis returns to the previously discussed conditions of 'non-manipulability' or whatever other framework one wishes to apply for the purposes of a similar analysis.

7.6.5. Conclusions

The non-domination view of freedom that Sparrow (2014) refers to in support of his argument cannot sustain the critique of the God Machine where the God Machine is seen as the dominating agent. Pettit's proposal that we extend the idea of freedom as non-domination to collective agents does not lend much help, as the God Machine is not a collective agent. Pettit's (1997) theory is ill-suited for the kind of critique of the God Machine world that Sparrow (2014) advances. The God Machine neither 'dominates', nor are the citizens of the God Machine society, as Sparrow wrote, 'its subjects' (2014, p. 27).

Although Sparrow's (2014) critique fails because it misconstrues the God Machine as an agent and the facts of the ideal world presented by Savulescu and Persson (2012a), the considerations presented in this chapter are relevant if we wish to use the God Machine scenario as an analogy for the real world. Harris (2014a) rightly raises the question about the limits that should be put on state power, even if the state power is non-arbitrary, democratic and legitimate at the moment. The power can be abused, even the well-crafted laws can be stretched and applied for purposes other than justice, the information leaked intentionally or not; the legitimacy of even democratic governments is open to discussion, and the transparency of governments' actions has to be constantly checked and re-assessed. Any instrument that diminishes such ability unjustifiably impairs the democratic process and undermines the checks and balances necessary for the maintenance of the system of man-made laws that serve those who create and abide by them. Although many readers live in democratic societies and enjoy a considerable amount of political freedom within those states, the degree of liberty even inside of those political systems tends to fluctuate. Finally, those systems sometimes change rapidly or even fall.

7.7. Why is the God Machine an undesirable way of making a better society?

7.7.1. Introduction

In the previous part of this chapter I focused on the question of the powers that can be justifiably given to the state. Given that in our world things can go wrong and protections from the arbitrary and unchecked interference may fail, Harris' critique of Savulescu and Persson's (2012) vision is convincing. The question about the exact limits of state power, however, remains open. In this section I will consider what I think is a crucial foundation of Savulescu and Persson's (2012a) argument: that the state's freedom-impairing interference is justified if it prevents harm. Savulescu and Persson (2012a) refer to Mill's harm principle to support their point. Firstly, I will review aspects of Harris' (2014a) rebuttal. Secondly, I will propose that Savulescu and Persson (2012a) misconstrue the purpose of the harm principle, which weakens their case for the desirability of the God Machine. In the second part of this section I will present an approach which extends the more convincing aspects of Harris' (2014a) argument and, in my view, is more successful in grounding an in-principle critique.

7.7.2. The harm principle

Harris (2014a) first quotes Dworkin, who insists on a distinction 'between the idea of liberty as license, that is, the degree to which a person is free from social or legal constraint to do what he might wish to do, and liberty as independence, that is, the status of a person as independent and equal rather than subservient' (cf. Harris 2014a, p. 258). I have previously suggested that the argument from non-domination (and the same will apply to 'subservience') is not convincing. That leaves us with the importance of liberty understood as independence. I agree with Harris that Savulescu and Persson's invocation of Mill's harm principle seems to overlook the importance of the distinction between liberty as licence and liberty as independence.

Harris writes:

> 'So Mill did not advocate the sort of freedom to do wrong that the law controls. But he recognized, as Savulescu and Perrson do not, that the law is not infallible, and the room, the independence, it leaves citizens to form their own values and choose their own way of life is vital for a free society—a society in which even basic laws may be changed for compelling reasons. The God machine takes away the independence of decision making, of thought that can lead to action; this is why it

> is incompatible with both independence and autonomy, incompatible with both liberty as license and liberty as independence.’(2014a, p. 258)

However, the first problem is that this in itself does not take us far in undermining the Savulescu and Persson (2012a) argument if we take away the support of the other arguments presented by Harris: that we might fare poorly if we believe in any instruments’ infallibility and that the ability to change laws is important (I think this ability is largely preserved in the God Machine scenario). This is because although liberty as ‘independence’ is important, and it too can be restricted – at least insofar as independence includes acting on one’s life plans. For example, Mill argues that freedom of expression, thought and discussion is a fundamental, and fundamentally important, liberty. For Mill, its importance lies in keeping true beliefs from becoming dogmatic, which is necessary if we are to fulfil our nature as progressive beings (see: *On Liberty*, II 20). Yet, even the exercise of basic liberties is limited by the harm principle, which justifies restricting liberty to prevent harm to others. Freedom of expression can be restricted on the basis of harm, as in this well-known passage from *On Liberty*:

> [E]ven opinions lose their immunity when the circumstances in which they are expressed are such as to constitute their expression a positive instigation to some mischievous act. An opinion that corn dealers are starvers of the poor, or that private property is robbery, ought to be unmolested when simply circulated through the press, but may justifiably incur punishment when delivered orally to an excited mob assembled before the house of a corn dealer, or when handed about among the same mob in the form of a placard. (J.S. Mill, ON, III 1)

Second, according to Mill, the harm principle is something that we can apply prospectively to prevent someone from acting in certain ways and causing harm. Although in many cases what we could only reasonably know is that a given action ***risks*** harm, this seems to be all that Mill requires (ON, IV 10). We have previously assumed that the God Machine can assess the danger well enough (infallibility is not needed). Thus, Mill’s harm principle can justify interference even with the liberties that Mill considers basic. As a result, introducing the distinction between liberty as independence and as licence in general cannot carry the weight required to defeat Savulescu and Persson’s (2012a) appeal to the harm principle. A further argument is needed.

There are two arguments I wish to make here. The first argument concerns the nitty-gritty of the application of the harm principle. Mill argued that the harm principle outlines the sphere of self-regarding actions that is protected from ’others meddling’ and *not* to argue for a converse claim that all that brings harm to others is for that reason only open to interference (i.e. harm is a necessary but *not* sufficient condition for a justified interfering action of the society or the state), a mistaken interpretation that seems to be at the core of Savulescu and Persson’s (2012a) confusion. In fact, Mill clarifies that:

> [I]t must by no means be supposed, because damage, or probability of damage, to the interests of others, can alone justify the interference of society, that therefore it always does justify such interference. (J.S. Mill, ON, V 3)

Thus, Mill's position is that causing harm is always *prima facie* and a non-negligible reason in favour of interference, but that this reason might be outweighed by reasons not to interfere. As a result, Savulescu and Persson (2012a) argument based on Mill’s harm principle (at least as it stands) fails.

Secondly, it is highly unclear why in the God Machine scenario the interference is to come in the shape of a direct and surreptitious change of intentions, rather than simply prediction and a last-minute preventative intervention of a SWAT team. Surely, if the God Machine can predict people’s actions with sufficient accuracy to warrant a change of intentions[41], why not simply monitor citizens’ thoughts, yet

[41] Harris’ (2014a) point about the difficulty of inferring the actual quality action looking from the inside at the intention alone is well taken. For the God Machine’s intervention to be sufficiently accurate in changing intentions, it has to have access to the wide knowledge base of the individual whose intentions are to be changed to be able to assess not only the intention to act but also the agent’s interpretation of this action. This would mean that at most the God Machine would know how the actions appear to the agent.

leave the freedom to nearly make the tragic mistake and make the intervention overt? Allowing agents to take the first steps of the harmful action would make the reasons for the intervention clearer, leave the possibility of disputing the interference and allow people to learn from their near-mistakes. In that case, however, the reason for changing peoples' intentions would not be the serious harm of others, but rather the harm to the agent. And this is exactly the kind of interference that would *not* be justified by Mill's harm principle.[42]

Moreover, there are powerful reasons not to allow the interference of the God Machine via changing intentions directly, reasons which I will explore in the next section.

7.7.3. The inner citadel

In his *Two Concepts of Liberty*, Berlin wrote about the retreat to the inner citadel that happens when a person realises he cannot attain his desires:

> 'I am the possessor of reason and will; I conceive ends and I desire to pursue them; but I am prevented from attaining them I no longer feel master of the situation. … I determine myself not to desire what is unattainable. … It is as if I had performed a strategic retreat into an inner citadel - my reason, my soul, my 'noumenal' self - which, do what they may, neither external blind force, nor human malice, can touch. I have withdrawn into myself; there, and there alone, I am secure. (1958, pp. 181-182)

Berlin uses the above passage to argue that the definition of liberty as the ability to do what one wishes to do is not sufficient; it would imply that when one cannot attain what one desires due to an oppressive and unjust social order, teaching oneself not to desire it would be a solution consistent with liberty – and yet, this is the antithesis of political freedom. Perhaps retreating to the inner citadel is not the kind of freedom that we wish to pursue and create via our social institutions, but I wish to make another point here. The terrifying fate that Berlin describes is

[42] There is much more to be said here about Mill's view of paternalism (See Mill 1859, chapter V) and alternative views, but it is sufficient to say here that Savulescu and Persson's (2012a) appeal to the harm principle is insufficient.

surpassed by the God Machine world where there is not even a citadel to retreat to. The God Machine, in contrast to even highly coercive measures such as imprisonment, undermines the very basic independence – the safety of the citadel itself.

The inner citadel, the ability to engage with the world and interpret it in our own way, engage with and take a stance towards the societal values and conditions we found ourselves in, lies at the root of the value of liberty. This ability is not everything, but it remains important and foundational – whether or not there are obstacles to living according to our endorsed choices and values. On the one hand, such conceived 'inwardly focused' autonomy is only one of the many facets of freedom. The idea of a 'chain of freedom' (Harris, 2014a), a chain from thought to action is, I think, compelling. The chain leading to free and effective action in the thickest sense may be severed in many places and by various influences. Our beliefs may be erroneous; our choices marked by the values of the society we would not reflectively endorse; the incentives, nudges and prods of policy makers, the law and circumstance may change the architecture of rationally endorsable options, and the exercise of our wills may be subject to obstacles, whether rooted in nature, social arrangements or intentional actions of other people. On the other hand, the God Machine does something more than break the chain: it attacks the chain at the very source of it, at the spot on which the meaningfulness of the whole chain depends.

What would justify the crucial importance of freedom of thought that I am advocating? Mill gives two reasons that could provide grounding for this idea in his defence of expressive liberties and arguments against censorship. The first is the utility of public discussion to promote the generation of knowledge. In the second argument, and the one that is more appealing in the context of the God Machine, Mill argues that freedoms of thought and discussion are necessary for fulfilling our natures as 'progressive beings' (ON, II 20). It is the exercise of our higher or deliberative capacities, Mill argues, that make a human life good (ON, I 11, 20; ON, III 1–10); the capacities to form, revise, assess, select, and implement our own ideas and plan of a good life. The God Machine, and heteronomy in general, undermines the exercise of those capacities in fundamental ways.

Another reason why Mill found liberty as independence from external influence important is because it is important that a person leads his life on his own terms and develops his capacities and faculties according to ‘his own mode of laying out his existence’ (Mill, 1859, p. 64), and to deny him that opportunity via interference is profoundly insulting – treating another not as an equal being capable of developing her own ideas of the good (see also: Quong, 2011, pp. 101–106). Raz puts forward a similar point:

> It is commonplace to say that by coercing or manipulating a person one treats him as an object rather than as an autonomous person. But how can that be so even if the consequences of one's coercion are negligible? The natural fact that coercion and manipulation reduce options or distort normal processes of decision and the formation of preferences ***has become the basis of a social convention, loading them with meaning regardless of their actual consequences***. They have acquired a symbolic meaning expressing disregard or even contempt for the coerced or manipulated people. … [S]uch conventions are not exceptionless. There is nothing wrong with coercion used to stop one from stepping into the road and under a car. Such exceptions only reinforce the argument for the conventional and symbolic or expressive character of the prohibition against coercion and manipulation, at least to the extent that it transcends the severity of the actual consequences of these actions. (Raz, 1986, p. 378)

I find the argument from the ‘expressed attitude’ unconvincing. It is no insult to a person's status as a moral equal to treat him in ways that presume that his rational capacities are not perfect, but subject to error. It does not follow from that attitude that they lack, or are deficient in, the capacities that give raise to their equal moral status. It is moreover unclear to me why we should be that strongly worried about other people’s (or even the state’s) insults – if the issue ends on insults, that is. The importance of the God Machine’s interference does not lie in the insulting character of it, but rather in the fact that it undermines our independence on a very basic level. It attacks the inner citadel, the safety of which is a precondition necessary for the political freedom of individuals to be meaningful. Thus, the ‘freedom in our mind,’ and the ability to trust that the making-up of our mind is truly ours, is foundational.

Having sketched out the reasons for the importance of the freedom of thought, the ‘inner citadel’, one has to ask what exactly is the scope and limit of the protection of such freedom (Blitz, 2010), what does that freedom consists of, how does the protection or promotion of such freedom conflict with other valuable kinds of freedom and what freedom is threatened by. There is insufficient space in the present work to develop a satisfactory account that can answer all these questions, but such an account would have to reflect on the ***ways*** in which desires are formed, endorsed and put into action — whether as a result of rational reflection on all the options available, or as a result of pressure, manipulation, ignorance or brainwashing (Christman, 1991; Dworkin, 1988).

Such influences have not been created equally – non-coercive measures such as structuring options or providing incentives lie on the one side of that spectrum, while the influence of the God Machine lies on the other. It not only prevents the exercise of – and impairs – the persons’ very capacities that are the centre of autonomy (my worry is thus similar to Harris’ 2014a, 2014b), but goes a step further. Whatever its exact scope and shape, the God Machine is clearly inimical to the safety of the inner citadel. It not only constitutes an influence that clearly comes from outside and unbeknownst to the agent, but also denies the agent the certainty of independence in the narrow area that is the last bastion of freedom when other freedoms might have been taken away. In Harris’ words, ‘… the independence … [of people] to form their own values and choose their own way of life is vital ...’ (2014a, p. 258).

7.8. Why the God Machine is a poor analogy for real-world MB

The God Machine thought experiment has invited much commentary. It gives us an opportunity to consider and flesh out the importance of different aspects of freedom, their respective value and importance for moral responsibility and moral life of agents. However, it suffers serious limitations insofar as it is taken to be an analogy for the real world or a basis for an argument about MB in the real world.

First, although the God Machine was proposed in the context of the discussion about the potential desirability of moral bioenhancement as a solution to society's great problems (Savulescu and Persson, 2012a; DeGrazia, 2014), only a small percentage of great harms are likely to result from intentional immoral action on a large scale (Fenton, 2010; Harris, 2010). Even for the morally enhanced population of the God Machine world, more harm would likely result from dangerous driving, negligent action or carelessness about safety than from deliberate harm. If the God Machine is programmed with utilitarian principles in mind, it should either dismantle itself, or put its computing power towards an aim that provides more overall benefit or better strategies of harm prevention. If it is Savulescu and Persson's (2012a) intention to argue for the desirability of moral bioenhancement, it is a poor omen that the God Machine is unconvincing even on the level of a thought experiment.

Second, the God Machine and pharmacological direct emotion modulation present different considerations in relation to PAP and overdetermination. While the issue of overdetermination may arise clearly in some specific cases of brain-state triggered brain modulation such as Hall's Brain Implant case from section 7.5, generally speaking direct emotional modulation is unlikely to be best understood as a clear-cut case of overdetermination. For example, while pharmacological attempts at MB may challenge our ability to ascribe responsibility and causation ('do you love me or is it your pill?') the issues raised here will more likely be related to authenticity rather than overdetermination.

Third, although the God Machine scenario is a good opportunity to reflect about the kinds of means we want states to have at their disposal in assuring the safety and security of citizens, the idealised scenario that we are presented with is very far removed from the possible worlds we are likely to experience. Thus, the kinds of issues we are likely to grapple with in the context of political freedom are not adequately captured by the scenario. Specifically, the justified worries about various abuses of power raised by Harris (2014a) do not readily apply to the God Machine society. Additionally, the fact that Hall's Brain Implant scenario is in my view significantly less troubling than the God Machine world means that the main problem lies not in the mode of intention change (i.e. via biomedical direct brain modulation) but rather with the fact that the modulation is not initiated by the agent – the agent is not aware of the external interference and cannot take a stance towards it, the agent cannot modify it and that the source of the influence is both external and externally controlled, thus undermining the process of creating a 'free will of one's own.' In reality that amount of ignorance on the part of the affected agent is highly unlikely.

Moreover, I agree with Harris (2010, 2014a) that the problems with specificity and the strength of biomedical interventions, together with the fact that large-scale harm is at least as likely to result from negligence, means that the application of MB for the originally proposed purpose of preventing large-scale harm is a red herring. The effects of biomedical modifications may be freedom-subverting or promoting, but I am hard pressed to see how beyond the God Machine scenario, MB could provide effects strong enough to give raise to a strong argument for compulsory use of MB on a population-wide basis (this is an empirical point, and I am open to being proven wrong). Yet, only such mandated use would overcome the collective action problem that the MB was originally conceived to remedy.

The God Machine scenario is symptomatic of a head-spinning mix of thought experiments which are many 'what ifs…' removed from the real world (and possible worlds that are actually possible for us). The practical conclusions drawn from such experiments are limited. To a large extent this chapter engaged with this mode of discussion and so some of the conclusions presented here will be susceptible to the same criticism. There is nothing in principle wrong with in principle arguments. Thought experiments that involve logically possible but practically impossible or unlikely worlds have some utility both in philosophical scholarship and in applied ethical brain-teasing: they can be helpful in elucidating and separating important aspects of a concept or issue, and this is how I intended the examples and arguments presented in this chapter to be taken.

However, we run into a serious problem when the optimistic 'what ifs…' are immediately followed by a sudden jump back to the reality. An example here is a passage from DeGrazia who swiftly follows the section containing an invitation to '[i]magine further that, as a result of MB, there were no more wars or starvation and everyone in the world had access to the basic necessities of life' with a radically more practically oriented approach: '[i]n the absence of a deity who will give us this better world, it is up to us human agents to attain it. Without a substantial improvement in moral behaviour, we are highly unlikely to do so. … In view of what is at stake, we should open-mindedly consider this non-traditional means of moral enhancement' (2010, p. 367). Nordmann (2007) critique of such approach pinpointed the mechanics and effects of such displacement of the hypothetical with the actual:

> 'An if-and-then statement opens by suggesting a possible technological development and continues with a consequence that demands immediate attention. What looks like an improbable, merely possible future in the first half of the sentence, appears in the second half as something inevitable. And as the hypothetical gets displaced by a supposed actual, an imagined future overwhelms the present.' (p. 32)

Although I am open to 'open-mindedly consider' MB, I do not think that the optimism rooted in taking a thought experiment like the God Machine as a close analogy for our world is the best justification for such consideration.

The context in which the biomedical modification is likely to be applied if it makes its way into our world is much less rosy than the God Machine society. The extent to which Savulescu and Persson (2008, 2012a, 2012b) advocate compulsory use is troubling, given that biomedical ME will likely be first applied in the context of the criminal justice and mental health systems of our morally unenhanced world. I have argued that, from the point of view of freedom, and as applied to the prevention of serious harm, the God Machine scenario fails to offer a compelling case for compulsory use even in its imaginary setting, and this conclusion is likely to be even stronger if we tried to more practically imagine it applied in a world more prone to abuses of power.

Finally, the God Machine scenario, with its focus on infallibly changing very specific intentions, has very little to say about the cases when (hopefully) voluntary, narrowly applied biomedical interventions can prevent harm or promote (or impair, if we are convinced by Harris 2014a, 2014b) taking a moral stance and moral action, as well as the kind of trade-offs and dilemmas such use would involve. In the next chapter, I will discuss some of the important ethical aspects of MB in the real (and our) world that could not be considered in the discussion shaped by the God Machine scenario.

7.9. Conclusions

In this chapter I aimed to contribute to the analysis of the ethical God Machine thought experiment considered the extent to which issues raised in relation to freedom in the discussion of Savulescu and Persson's (2012a) thought experiment calls the desirability of MB into doubt. Using a series of thought experiments to tease out exactly in what way the God Machine could endanger freedom (various cases of overdetermination of agent's actions).

In sections7.3-7.5 I have argued that the main problem with the God Machine is that it breaks the link between agents own reasons for action and the outcome in the world. Section 7.3 explicated the issues in relation to moral luck and prise and blame, section 7.4 used Frankfurtian analysis in order to argue that an important aspect of the God Machine's threat to freedom lies not in endangering free will generally but rather by undermining specifically the ability to form a *will of our own*. Section 7.5 I argued that the problem does not necessarily lie in the fact that the God Machine is a case of overdetermination, and that the more plausible uses of MB that involve overdetermination would be significantly less problematic.

Section 7.6 examined Sparrows (2014) objection in the context of Pettit's freedom as non-domination. After outlining Sparrow's (2014) objection and discussing the freedom as non-domination account, I have argued that Sparrow's (2014) critique fails because it misconstrues the God Machine as an agent and the facts of the ideal world scenario presented by Savulescu and Persson (2012a). The freedom as non-domination theory might be particularly ill-suited to the ethical analysis of the God Machine, due to problematic aspects of the notion of agency in that theory as well as an unclear agential status of the God Machine. Thus, I have concluded that that non-domination theory of freedom as interpreted by Sparrow (2014) is ill fitted to ground a robust critique.

In section 7.7 I used an analysis of the application of Mill’s harm principle and argued that the God Machine would be an undesirable way of achieving a morally better world because it might adversely affect the way the desires and inclinations for action are formed. This discussion shed light on the factors that are to be considered when evaluating the impact of MB on moral agency.

However, I have argued that the conclusions taken from the consideration of the God Machine thought experiment can only bring our attention potentially important aspects, but due to the degree of abstraction and important differences between the God Machine and MB, the arguments related to the God Machine should be transferred with much caution to the ethical assessment of real-world MB. I have concluded that the arguments raised in relation to the God Machine thought experiment fail to call the desirability of real world, hopefully voluntary and agent led MB, into doubt.

CHAPTER 8. Beyond the God Machine

(…)
In the general mess of imprecision of feeling,
Undisciplined squads of emotion. And what there is to conquer
By strength and submission, has already been discovered
Once or twice, or several times, by men whom one cannot hope
To emulate—but there is no competition—
There is only the fight to recover what has been lost
And found and lost again and again: and now, under conditions
That seem unpropitious. But perhaps neither gain nor loss.
For us, there is only the trying. The rest is not our business.
(…)
T.S. Eliot

8.1. Introduction

Harris argues that MB would likely eliminate a significant measure of our freedom and that this loss would be unacceptable because it diminishes truly moral action. Harris proposes that MB bypasses morality and that MB ‘far from raising consciousness, may well dull it to the point where the individual is no longer choosing’ (Harris, 2014b, p. 372). On one plausible interpretation, this would seem to posit that the best MB can do is to create compulsions to act in a morally good way or compulsions to refrain from immoral acts. As a result, Harris (2011) argues, MB is not a moral enhancement properly so called (moral enhancement that results in bettering of moral agency) as there is no virtue in doing what you must.

Before considering this argument further, let us clarify some misconceptions. One proposed interpretation of this argument is to refer to non-deterministic views of free will. For example, DeGrazia (2014) considered Harris’ worry by outlining and then arguing against Chisholm’s (1966) incompatibilist conception of free will (DeGrazia, 2014, p. 365).

However, Harris' argument about bypassing morality can be made sense of within a compatibilist view of free will and free action such as Frankfurt's. Indeed, Harris states that 'no assumptions about the viability of a non-deterministic account of free will and, for what it is worth, think some version of compatibilism is probably right' (Harris, 2014b).[43] Even though in his reflections on the God Machine, Harris seems to argue that we need to keep PAP in order to preserve the 'ought implies can' principle, I agree with Harris that this line of the debate is not the most relevant.[44] According to compatibilists, what matters is whether the decision was arrived at in the right way, with the right sort of relationship between one's preferences and actions – thus, acting on some – and not other – kinds of causal influences. If one acts with the ability to act in accordance with good reasons (Wolf, 1990) or if one acts with 'guidance control' which consists in part of acting on a reasons-responsive mechanism for which one has taken responsibility (Fischer and Ravizza, 1998), one can be responsible for one's actions. This approach is consistent with Harris' emphasis on the importance of the all-things-considered judgement (Harris 2011, 2012, 2013a, 2014a, 2014b). Thus, according to a compatibilist account, freedom can be impaired if action is produced by causal influences that do not satisfy the right criteria.

Moreover, at this point of the debate I think it is important to move beyond the following line of reasoning. Harris' argument can be seen as addressing only the cases brought forward by Persson and Savulescu (2008). Harris sees Persson and Savulescu (2008) proposing the use of biomedical means in such a way as to necessarily produce morally good outcomes via modulation of emotions, sentiments or attitudes. When talking about serotonin and oxytocin's influence on what psychologists and neuroscientists call moral judgement, Harris (2014b) states that: '[t]he presence of these molecules in particular doses is indeed "freedom-

[43] Although in another article Harris defends the Principle of Alternative Possibilities (Harris, 2014a).

[44] I do not see it as 'irrelevant,' as our commitments in the compatibilism/incompatibilism domain may influence the kinds of arguments we are going to find more or less compelling. However, I do not think that taking hard positions on those questions is necessary for continuing a meaningful debate about moral enhancement. Accepting the 'ought implies-can' maxim should lead one to accept that an agent must have genuine access to the world which renders the 'can-claim' true. Even if unconvinced by my argument that the agent has such access in overdetermination cases, it remains an open question to what extent this would apply to more worldly incarnations of MB.

subverting" – if it were not, it is unlikely they would have the effects vaunted by their advocates, that is, effects that operate independently of the will or of judgement ...' (p. 373). Although the interpretation of the original Savulescu and Persson paper (2008, and latter proposals) is justified and Harris' argument works in this context, this line of discussion seems to me to be a dead-end. For one, oxytocin and serotonin are very unlikely to produce the type and magnitude of effects that Savulescu and Persson seem to be after in their early paper – an effect strong enough to preclude large scale harm. Secondly, as psychopharmacological means are used more and more to modify mood, cognition and our social functioning, more plausible uses of MB warrant a careful examination. In considering biomedical emotion modulation and its potential for moral enhancement understood as making better moral agents, God lies in the details.

In this chapter I will first examine whether Harris' worry about the freedom-subverting effects of psychopharmacological means that act on attitudes or desires is likely to apply to more plausible attempts at MB. I will ask whether biomedically created 'compulsions to do what is good' preclude moral review, addressing the question of whether morality is indeed bypassed by MB. Using the example of obsessive-compulsive disorder (OCD) to illustrate the issue and considering biomedical mood modification, I will suggest that even strong impulses to act in a certain way are not beyond meaningful moral review.

The second part of the chapter addresses issues related to moral control. I consider MB in the context of a discussion of Aristotelian concepts of virtue, enkrateia, akrasia and vice. I suggest that inverse-akrasia is the best analogy for Persson and Savulescu's (2008) MB and argue that even where the agent acts akratically, compulsory MB would not be desirable. In further sections I argue that enhancing individual moral agency is a valuable pursuit and that modulation of inclinations (including biomedically-aided modulation) not only has a legitimate place in this pursuit, but is a crucial part of it.

8.2. Are ‘compulsions for the good’ beyond moral review?

8.2.1. Introduction

Shaftesbury asserted that:

> ‘every reasoning or reflecting Creature is, by his Nature, forc'd to endure the Review of his own Mind, and Actions; and to have Representations of himself, and his inward Affairs, constantly passing before him, obvious to him, and revolving in his Mind.’ (Shaftesbury, 1699 II.1)

MB could be detrimental to freedom and rational moral agency if it resulted in the creation of what I am going to call ‘compulsions for the good’ – strong inclinations that are more likely than not to result in the agent acting consistently with the good.[45] Here, the agent would be as unfree with regards to her desires as a (willing or unwilling) addict. Harris (2011) argued that insofar as MB results in creating such compulsions, MB bypasses reasoning and is beyond moral review. However, I will argue that such compulsions *are* subject to moral review. In this section I will distinguish between offline and online moral review and offline and online moral control and use an example of an obsessive-compulsive disorder (OCD) to argue that even where online control is affected, it does not follow that the online and offline moral reviews are fatally diminished.

8.2.2. Online and offline moral review vs online and offline moral control

I propose that we introduce two distinctions when discussing biomedical emotion modulation. First, it is important to distinguish between ‘moral review’ and ‘moral control’. I will understand *moral review* of actions as an evaluation of actions

[45] One could object that without the guidance of reason we are not able to reliably achieve outcomes consistent with the good. This may be likely, so I have set the bar much lower (see also the penultimate section of this chapter for further justification of such lower requirements). At this point it is not crucial to demonstrate that the notion of ‘compulsions for the good’ makes sense. Since I will consider here the statement that ‘if they are possible, they would preclude moral review,’ we can for now simply agree that the argument is conditional on the existence of such compulsions. The conclusions of the argument will later be applied to cases when MB does not involve compulsions, and here I set out to demonstrate that even in the more troubling case of MB involving compulsions, the issue of moral review is not a strong argument against MB.

undertaken from a moral perspective.[46] In this discussion I will assume that moral review will be rational and deliberative in nature, although the extent of the deliberation may vary from very little to the writing of philosophical tractata. In contrast, *moral control* refers to the exercise of a capacity to modify actions within the parameters of what is prescribed by moral review. A second distinction helpful in the current discussion is a distinction between what I will call 'online' and 'offline' control and review. I will understand *online review and control* to take place as the behaviour or emotion unravels; *offline review and control* will refer to appraisal and control that takes place either *post-factum* (e.g. reflecting on the appropriateness of an emotional reaction after the emotion subsided) or before the event (e.g. stimulus avoidance, contingency planning).

Strong desires, emotions and the resultant compulsions to act are subject to review in two ways. Firstly, even if we cannot make our will effective at a given time, we can take a position towards that compulsion in a similar way that an addict may or may not endorse their effective desire for a drug (offline review). This may not make any difference in that particular instance or while an agent acts on that desire, but provides a background for an effort to modify such desires or mitigate the impact of actions flowing from those desires (offline control). If a certain dose of oxytocin resulted in me being so sensitive to witnessed suffering that I feel an inescapable compulsion to help, this does not preclude my reflection on whether that was an all-things-considered good thing *post factum* and making later adjustments. The mere presence of a drug-induced pro-social compulsion does not abolish that judgement.

Second, the capacity for online moral review might be preserved, even if the action is automatic, despite the fact that traditionally automatic actions have been associated with lack of awareness (Norman and Shallice, 1986) and contrasted with willed action (James, 1891). For James, automatic action happens 'wherever movement follows unhesitatingly and immediately the notion of it in the mind … We are then aware of nothing between the conception and the execution,' while

[46] See: Oxford Dictionary, http://www.oxforddictionaries.com/definition/english/review (Accessed: 27.08.2014). Although in some uses of the concept, 'review' may include an intention to make necessary changes, the action guiding aspect is not a necessary component of the concept and I will use 'review' to denote 'critical assessment' or 'evaluation.'

acts that require exercise of will, include ‘an additional conscious element of a fiat, mandate, or expressed consent’ (p. 522). While automatic actions may happen without awareness of the action taking place, we can typically bring our conscious awareness to observe and monitor those actions. Taking a sip of water, breathing or blinking might happen without conscious awareness when we are engrossed in a conversation, but we might chose to direct our attention and examine these processes and actions more closely. The ability to consciously attend can be further trained. For example, mindfulness meditation may involve the systematic practice of attending to actions performed without conscious awareness, such as breathing and walking, and attending to otherwise unnoticed sensations, emotions and thought processes. The ability to be aware of automatic actions and processes suggests that the preservation of review in general and moral review specifically is at least in principle plausible even when the actions cannot be said to be ‘willed’ in James’ (1891) sense.

To more closely examine this possibility, let us examine whether conscious review is preserved in the case of uncontrollable behaviours in OCD. In his article on the phenomenology of compulsions, Denys (2011) begins with an illustrative case. A patient is young mother, who is terrified by the thought of killing her daughter:

> ‘When I’m alone at home and I see my daughter sleeping in her crib then I can see myself strangling her. I’m terribly shocked by the thought and I am very frightened by it. If nobody holds me back, I could murder my daughter. I don’t want to harm her, but there is no guarantee that I never will. I can’t control myself any longer. I thought I was a good mother, but the fact that I think about it says something about who I really am. It shows that perhaps I don’t love my daughter enough. I don’t want to think about it but I’m not able to keep the thought out of my mind. The harder I resist, the stronger the thought is. In the beginning I occasionally thought about it, but now I think about it all the time. Though I realize that the thought is absurd, I can’t stop it.’ (in: Denys, 2011, p.1)

In this example, Denys' (2011) patient takes an active evaluative stance[47] towards compulsive thoughts and the resulting behaviours.[48] OCD is typically considered to involve 'insight' about the 'senselessness' of the behaviours and lack of endorsement of the intrusive thoughts. Such 'insight,' in the language of psychiatry, is somewhat different from the concept of 'moral review' I proposed earlier. It is usually defined and established on the basis of more than procedural criteria. Rather, 'insight' is understood in content-laden terms, encompasses more than just moral review specifically and is often established on the basis of the *outcome* of a review. Despite these dissimilarities between 'insight' and 'moral review,' the example serves our discussion well in so far as the presence of 'insight' indicates a preserved ability for epistemic, pragmatic and moral review – even when strong compulsions are present. As Denys' (2011) points out,

> 'With OCD there is always a moment of subjective reflection as a result of which a viewpoint will be taken against the contents and the form of the symptoms. The patient with OCD is engaged in a dialogue with his or her disease and constantly reviews his or herself with respect to the contents and the form of the thoughts and acts. … The obsessions … will be denied, resisted, avoided, doubted, smoothed over, compared, balanced.' (p.5)

The first lesson worth considering in discussing MB is that the presence of strong compulsions and intrusive thoughts[49] need not abolish moral review.

[47] OCD is typically regarded as involving 'insight' about the 'senselessness' of the behaviours and thoughts. Such 'insight' is usually defined not in procedural terms, such as an ability for review, but in more content-laden terms. Nevertheless, it demonstrates the ability.

[48] Psychologists and psychiatrists call such thoughts 'obsessions,' while they refer to behaviours that relieve anxiety by obsessions, situational avoidance and covert behaviours such as thought suppression and prayers as 'compulsions.' See for example: Abramowitz (1998). For the purpose of this work, however, what matters more is the property of 'compulsiveness' (uncontrollability leading to some effects) whether as applied to intrusive thoughts or to uncontrollable behaviours.

[49] One could object that intrusive thoughts in OCD thoughts are 'mere imaginings' in the sense that they do not have a motivational component. Analogically, even though I could visualize shaving off my partner's hair while he is asleep, I may have little or no motivation to actually do it. I do not think that such an objection is strong – it does not follow from the

A further question that can be raised is whether or not the presence of strong compulsion allows for the presence of moral review of sufficient quality: whether such review takes into consideration the right factors in the right way, whether it relies on well-justified beliefs, and so on. This doubt is *prima facie* warranted. The OCD symptoms are thought to be maintained by faulty beliefs, e.g., the belief that having a thought is morally no different from doing a bad thing (what psychologists call 'thought-action fusion')[50] and that having a thought about harming one's child means that there is a significant actual danger of doing harm. Another concern about the review of actions in the presence of a strong compulsion is related to the question of whether or not such review is all-things-considered. A more specific question to be asked in the context of MB-induced compulsions is whether or not, and to what extent, the presence of strong impulses *precludes* an appropriate all-things-considered review.

Although the example of OCD is useful in demonstrating the possibility of preserving review in the presence of strong compulsions, the example is of limited utility in considering whether or not the presence of MB-induced compulsions would impair the quality of moral review. Since OCD is treated as a disorder, something did go wrong at some point.[51] Even if the quality of all-things-considered review would be impaired in this population, it would be unclear whether or not this impairment is due to the presence of compulsions or some other characteristics of that population.

fact that OCD patients do not usually act on their imagining that the thoughts are 'motivation-less' as the fact that the desire is not effective does not mean that there is no motivational component at all; moreover, I have not encountered any indication of lack of such motivational component of the intrusive thoughts in the literature on OCD.

[50] Note that although I have earlier argued that motivations on which we act should be considered together with actions in assessment of moral responsibility (and are thus 'joined' for the purposes of such assessment) this does not commit me to problematic positions entailed in 'though-action fusion.' Accepting actions and their motivational underpinnings as the unit of moral responsibility assessment entails neither that we are responsible for our thoughts without any regard to actions, nor that we are responsible for every single one of our thoughts, nor that thinking about doing something is as morally good or bad as actually doing it.

[51] This point is related to the discussion of medicalization in Chapter 5. The construction of a certain way of thinking or acting as a medical condition often includes the assessment of it as in some way harmful or undesirable, thus making it a value-laden process. This assessment may include the reference to ideas about agency, moral agency including.

However, let us suppose that the above problem does dissuade us from attempting to make an argument that OCD indicates that strong compulsions preclude all-things-considered judgement moral review of a sufficient quality. There are further problems with such a line of argument. The first problem relates to the influence of faulty beliefs. It is a matter of discussion whether the *process* of the all-things considered judgement is faulty or whether the process is comparable to non-OCD population, with the problem here lying in the influence of faulty *beliefs*. There is some indication that faulty beliefs are an important factor. CBT treatment for OCD focuses in the first instance on changing faulty beliefs by examining them in talking therapy, confronting the beliefs with reality by exposure and habituating the patient to the anxiety that the thoughts evoke. It would seem that the main target in CBT is change of beliefs and integrating them into the process of emotional responses, rather than enhancing all-things-considered judgement. The conclusion we can justifiably take away from this is a modest one – that the review in general and moral review specifically is influenced by the beliefs that a person holds. Little can be concluded about the quality of the process of review in comparison to those not afflicted by strong compulsions.

The second problem is related to the direction of the causal influence. Post-partum OCD often arises in patients with previous OCD characteristics. This indicates that it is not the occurrence of a particular strong impulse or inclination that leads to the development of OCD; the causal influence is more likely to go the other way round. Many new parents may experience a thought or impulse to harm their child (similar to the impulse to hit when we are very angry), which when not acted upon, unendorsed and not given much weight, simply diminishes with time or is not a problem. It seems that the way OCD patients respond to such impulses contributes to their maintenance. As a result, although OCD is a good example of a moral review present *alongside* compulsions, it is not a good example of *compulsion-induced impairment* in moral review.

Assessing the extent to which the ability for *online* review is preserved in the presence of strong compulsions is more difficult. I am not aware of phenomenological data that clearly assesses the kind and quality of awareness and moral review while the patient experiences intrusive thoughts and performs ritualistic compulsive actions, nor data from other compulsive behaviours such as binge eating.[52] An additional problem is that for the argument about the impairing effects of compulsions on online moral review, we would have to have similar data about a matched population not afflicted by compulsions. Since neither are present, we have to work here on best guesses.

Patients' reports such as the one cited above provide some indication that online review is to some extent present at all times and that online review is strengthened by therapeutic interventions (Denys, 2011). Moreover, the reports of the feeling of 'loss of control' during compulsive or impulsive episodes indicate the presence of online review (O'Guinn and Faber, 1989). Such feeling presupposes a degree of online review happening, even if the ability for online control is impaired. The possibility of the presence of online moral review is also implicitly indicated in cases of successful attempts for online control that is a common experience for many of us – even when we are very angry we might judge it inappropriate to act on that anger, realise that our anger is disproportionate to the stimulus and go for a head-cooling walk instead of acting on the anger-related impulses. Moreover, even if we fail to resist following angry impulses, we might act knowing 'at the back of our head' that we should not. This provides support for the practical plausibility of the dissociation between moral review and moral control both online and offline. Moreover, it seems that ability for online moral review can be preserved even in the presence of strong compulsions or impulses and is separable from online control.

8.2.3. Biomedical emotion modulation and moral review

A further question to be raised is the question of whether biomedically-induced strong inclinations specifically abolish moral review. I do not think that the mere presence of a strong impulse or inclination does so, and so one cannot justifiably infer from the presence of a strong and even behaviourally effective impulse that the moral review is impaired. I have argued that for those MB attempts that would indeed induce strong compulsions, the argument that MB abolishes moral review *in virtue of* its compulsion-inducing effects fails.

[52] Despite my best and prolonged effort to find such data.

However, there are two further arguments to consider before we conclude that biomedical MB does not threaten moral review to an unacceptable degree. The first argument may be derived especially from the observation that strong passions profoundly distort our view of things if they are strong. In anger or jealousy, for example, when the red mist comes down over the eyes, and we can feel the blood pulsing in our temples, things look other than the way they are and our emotions can mislead us profoundly. In his treatise on anger, *De Ira,* Seneka warned about the potential outcome of intense emotions in general and anger specifically:

> 'Anger, I say, has this evil: it refuses to be governed. It rages at truth itself, if truth appears to conflict with its wishes. With shouting, turmoil and a shaking of its entire body, it makes for those whom it has earmarked, showering them with abuse and curses.' (Seneca, 1995, 19.1).

But perhaps it is not the kind of emotion, but rather its strength, that clouds judgement and precludes the agent from acting according to reason. Could then MB lead to agents who, although moved to action by a benevolent emotion, are in the red (or perhaps… pastel pink?) mist of empathy? Chan and Harris, for example, have argued that if oxytocin or serotonin induced strong feelings, they would impair judgement (Chan and Harris, 2011).

But what are we concerned about here exactly? We regularly experience emotions clouding our judgement in the moment. In Wordsworth's (1815) *Surprised by Joy,* the narrator recalls emotions evoked by a memory of his deceased child:

> Surprised by joy – impatient as the wind
> I turned to share the transport – Oh! With whom
> But thee, long buried in the silent tomb,
> That spot which no vicissitude can find?
> Love, faithful love, recalled thee to my mind –
> But how could I forget thee? - Through what power,
> Even for the least division of an hour,
> Have I been so beguiled as to be blind
> To my most grievous loss? – That thought's return
> Was the worse pang that sorrow ever bore,

Save one, one only, when I stood forlorn,
Knowing my heart's best treasure was no more;
That neither present time nor years unborn
Could to my sight that heavenly face restore.

As the narrator is experiencing 'the worse pang of sorrow ever bore,' we probably would not ask him to make decisions that require calm reasoning at that moment. But he is not incapacitated in his life for having experienced that feeling once or even regularly. In fact, the feelings of joy, guilt and sorrow reveal something about what is important to the narrator. Save for the times when grief is overpowering over a long period, the presence of temporary strong emotions simply speaks to what we care about and does not make us generally incapable of rational action or pursuing our life plans. It might lead to re-evaluation of priorities but does not preclude self-governance.

Even if MB produced pangs of temporarily incapacitating empathy as its side effect, that would not be to the absolute peril of rational agency. Perhaps we would ask ourselves 'through what power … have I been so beguiled as to be blind' to the suffering of the starving and dying millions. I do not say that such pangs are what morality is all about, nor even that it would necessarily produce the behavioural effects that Savulescu and Persson (2008) hope for. In the end, as Rousseau pointed out, the pity aroused by a tragic drama can be nothing more than a 'transitory and fruitless emotion, which lasts no longer than the emotion producing it. … A barren compassion indulging itself in a few tears but never productive of any act of humanity' (p. 34) and so can be the pangs of biomedically induced sympathy or compassion. But should we decide for other reasons that emotion modulation is something we wish to pursue, perhaps the impairment of rationality coming as a result of occasionally 'clouded judgement' as side effect is not something to be gravely concerned about, and, to the extent that the feelings are endorsed, perhaps it would bring our attention to something we do – or perhaps even ought to – care about.

MB cannot have the strong 'vice abolishing' effects that Savulescu and Persson (2008, 2012) are after, in part because most commonly even strong emotions can be regulated and acted against. But this also means that the arguments against the ethical permissibility of MB lose some of their edge – the effects of MB are also subject to the regulation of affect and even strong affect and presence of compulsions can leave the ability for offline moral review and rational agency intact. If MB is proposed to indeed act though by-passing judgment and regulation *altogether* to produce impossible-to-regulate affect and behaviour, this indeed would be a problem – not only for freedom but for the basic ability to act as the agents we are used to being. I have, however, argued that this is not a necessary outcome. The ability for offline review can be preserved, and it is the ability for offline review that matters for crafting institutional level solutions that Harris (2013b) proposes we should prioritise.

But perhaps I am wrong – biomedically induced emotions and inclinations could also significantly impair the process of offline moral review itself. Harris appears to me to assume as much at least in some of his writings (2013b)[53] and the psychologist Bloom recently made a similar argument in a *Boston Globe* article aptly titled *'Against Empathy'* since, as he writes, he is 'against empathy.' Bloom argues that feeling strong empathy does a disservice to both the empathizer and the individual on the receiving end and fails to make the world a better place by impacting negatively on what I have called offline review. Bloom argues that:

> 'Our policies are improved when we appreciate that a hundred deaths are worse than one, even if we know the name of the one, and when we acknowledge that the life of someone in a faraway country is worth as much as the life a neighbor, even if our emotions pull us in a different direction. Without empathy, we are better able to grasp the importance of vaccinating children and responding to climate change. These acts impose costs on real people in the here and now for the sake of abstract future benefits, so tackling them may require overriding empathetic responses that favor the comfort and well-being of individuals today.' (Bloom, 2014)

[53] Although this interpretation might be overblown given that he mainly responds to the rather extreme proposals in Persson and Savulescu 2008.

But beyond the first appeal this is puzzling – what has empathy to do with it? We could empathise with a vaccinated child or imagine the ones who would die if not for the vaccination. If we are to support a military intervention by our own country in a faraway but oil-rich land, it does not appear to me that we do it because we personally know the soldiers, or *empathise* with our neighbours' benefits more than the other people's sorrow. There might be a bias in terms of being more easily able to empathise with people more like us, but most of us do not find it hard to empathise with those who had their loved ones killed or are starving – those arouse rather vivid pictures and it does not take a lot of cultural sensitivity to be able to react to them with empathy. We may suffer from a problem of imagination, care about our own interests more than citizens of other nations, be more often exposed to the images in the media that enhance our ability to empathise with 'our side' (considering that there is an 'our side' is probably an indication of a certain way of perceiving and understanding the events), and care more about the interests of our family than others or simply be too busy with our own lives to care. Those plausible explanations have little to do with *empathy* as an affective or cognitive ability. Just as anger is not *the problem* to be eliminated (contra Seneca), as it can equally serve to fuel the fight against injustice and unfairness or lead to unjust violence; similarly empathy can be conducive or not to moral outcomes. Especially when we are talking about offline moral review, such as in deliberation about a moral course of action and thinking about political solutions to matters requiring collective action, the quality and process of our reasoning about policy measures seems to hardly be rooted in 'more or less empathy.'[54] Thus, 'more empathy' is not a solution to world's ills, but neither is 'more empathy' perilous to the moral discourse. The ability for offline moral review, by and large, is preserved regardless of the strength of emotions and is thus accessible to people with a range of emotions and degrees of empathy – bioenhanced or not bioenhanced.

I do not see a clear indication that serotonin or oxytocin-induced changes would necessarily have such an perilous effect and thus should not gravely worry us,

[54] Within certain parameters – some empathetic ability may be necessary for understanding what others' interests are, for example. But I assume that we are talking about enhancement in the population that does not, for the most part, suffer from Autistic Spectrum Disorder. The ability to *understand* that to starve or to be poor is not good for others does not require a whole lot of empathy.

although given the side effects of most drugs and their typical impacts on the outcomes of reasoning, the matter is up for debate. This, however, leaves us with a difficult question – since there exists a degree of variability in the presence of those neurotransmitters in drug-free populations anyway, the drug-induced changes might be seen as not more biasing that our non-modified condition. A successful critique would have to make sense of and account for that.

Notice that I have argued that the ability for moral review would be likely to be present 'by and large.' This is because the question about more subtle effects on reasoning (as well as motivation and action) remains open. If we were convinced that any kind of biomedical emotion modulation would have a detrimental effect on moral agency, we would have to accept that people who take SSRIs are somehow made worse moral agents for that. For Harris' and Blooms' argument to be strong, such impairment needs to happen in a specific way – by impairing the ability for appropriate moral review and seriously impairing their ability for political participation. This conclusion seems counterintuitive.

I do not think that the *counterintuitiveness* of such a conclusion is a good reason to abandon doubts about MB altogether. In fact, it might be worth looking more closely at the influence of SSRIs on our moral action and reasoning. This is especially important given that millions take those drugs and that the effects of those drugs on moral agency and reasoning specifically were not evaluated to a great extent. The few conducted studies demonstrated a weak effect. Those results evoked considerable discussion (Crockett et al., 2010, for discussion see for example: Harris and Chan, 2010; Chan and Harris, 2011) but a more careful examination of the effects is needed to make strong conclusions about the effects of SSRIs on moral reasoning and action. When SSRIs are used to alleviate strong and debilitating depression, it is plausible to assume that that the overall obvious effects in relieving depression and the related increase in the ability for moral review and action outweighs the less noticeable and speculative impairments. This argument becomes weaker and weaker as the drugs are taken by people with lesser and lesser severity of mood disturbance. As a result, I think that there is merit in, and a need for, further investigating the effects of SSRIs on moral and social reasoning and action.

However, given the scarcity of current evidence I think that the claim that SSRIs substantially impair moral review is currently unwarranted. Even if SSRIs influence cognition as well as emotionality and general mood, people who take them seem generally able to reflect on their stance towards the effects both online and offline (Prince et al. 2009). There is no obvious reason why such review should selectively cease to apply to moral actions and pro-social inclinations, and those taking SSRIs at least appear as perfectly able to care about the world's great ills and engage in political debate as others.

However, the matter merits further discussion and empirical investigation, especially as we are talking about the impact on everyday moral agency. Everyday and face-to-face moral agency may be more affected by the emotional dispositions of the agent, and there are some reports indicating that some patients taking SSRIs experience emotional blunting. Opbroek at al. study suggests that a subpopulation of SSRIs users experience emotional side effects, including less ‘ability to cry, irritation, care about others' feelings, sadness, erotic dreaming, creativity, surprise, anger, expression of their feelings, worry over things or situations’ (Opbroek et al., 2002, p. 147) More recently, Sansone and Sansone (2010) reviewed the evidence for what they called ‘SSRI-induced indifference,’ an SSRI-related state that includes both emotional blunting and behavioural apathy and Marazzati et al.’s (2014) research suggests that antidepressants significantly affected patients’ feelings of love towards their partners.

Although there is a need to further investigate the morally-related effects of currently used biomedical emotion modulation, in the absence of further argument the objection that biomedical direct emotion modification impairs moral review does not apply to biomedical ways of influencing morality any more than it does to our moral life without such influence. As a result it cannot ground a decisive objection against the use of biomedical emotion modulation.

8.2.4. Conclusions

I have argued that there is no good reason to see MB as necessarily impairing moral review. Even if in effect MB makes people experience compulsions to act pro-socially, this does not undermine the ability to take a moral stance; it does not preclude the ability for online and offline moral review. Moreover, even if such compulsions are irresistible – an agent has to act on them – they cannot be said to impair moral review. They are better conceptualised as impairing freedom by decreasing the amount of effective moral *control*. Thus, the argument against Persson and Savulescu (2008)-style MB needs to be qualified and sharpened to refer primarily to moral control.

Moreover, the distinction between moral review and moral control makes a difference for the ethical assessment of MB. MB needs not preclude taking a moral stance, i.e. considering actions from a moral perspective, nor the ability to assess one's actions from a moral perspective. Moreover, if MB's influence is known to the agent, this can be accounted for in the process of review, just as the impact of SSRIs on person's mood is accounted for when people reason about the influence of SSRIs on their lives and decide whether or not to continue the medication. The influence on the control of one's actions, and its moral importance, is a separate issue.

The fact that moral review does not appear to be significantly impaired constitutes a big difference between the likely real-world pharmacological or brain-stimulation enhancements and the God Machine scenario – the God Machine impairs the ability for moral review significantly more as it precludes any resistance or meaningful reasoning about the sources of one's actions. I have argued that the God Machine does not simply introduce the irresistible compulsion that one may view as such and take a stance towards. The loss of the ability to govern one's actions independently and rationally is covert, unknown to agent and achieved exactly by affecting the beliefs and structures that makes the moral review the *agent's own* moral review. Contrary to the Rational Persuader Machine, in the God Machine scenario the changes in the agent's beliefs do not come in a way that the agent can appropriately engage with, in a way that would constitute a process of learning and revising one's view on the basis of reasons. Voluntary MB does not suffer from similar pitfalls and, as a result, MB does not diminish moral review to the extent that the God Machine does. In other words, the real-world MB does not necessarily undermine the crucial precondition for freedom and the meaningful talk about freedom as the God Machine does.

This conclusion has to be qualified to account for cases where the MB-induced emotions overwhelm the agent to the point of precluding moral review and seriously and negatively affect judgement. We have strong reasons to avoid such outcomes. At times, however, we experience strong emotions, but they need not render us incapacitated with regard to the ability for moral review, especially insofar as we are talking about moral agency that involves offline moral deliberation such as talking about collective, state and supra-state solutions to the world's problems. As a result, we have a strong reason to use biomedical ME with continued evaluation of their effects and reflective guidance. This justified caution, however, does not mean that there is a strong reason to abandon the pursuit of MB.

8.3. Moral control

8.3.1. Introduction

Recall that Baron-Cohen (2011) argued that a high level of empathy such as 'being continually focused on other people's feelings ... in a constant state of hyperarousal, such that other people are never off their radar,' is conducive to making people good and creating good societies. Those ideas met with resistance (for more discussion see: Chapter 3, section 3.3). One of the problems highlighted (Bloom, 2011; Harris, 2011, 2012, 2013a, 2013b, 2014a, 2014b) is that such person's drive to empathize is unstoppable and does not necessarily stem from moral concern. While Bloom argues on practical grounds why high levels of empathy do not make for nicer or better people, here I am more concerned with the impairment to moral agency – specifically, moral control – that aiming to create such individuals may bring.

When we talk about MB as introducing strong pro-social inclinations, we run the risk that instead of creating virtuous individuals, we create akratic individuals. In the following sections, I will examine the ethicality of biomedical modification of inclinations in relation to Aristotle's account of virtue, enkrateia and akrasia (with some revisions). I will first consider inverse-akrasia and, using the example of Huckleberry Finn, argue that even in a weak willed action it does matter whether the actions flow from reasons that are a part of the agent's moral worldview. This constitutes a strong reason against compulsory MB, but not against voluntary and agent-led biomedical modification. Further, I will argue that although creating weak-willed individuals who do good is hardly the ideal moral enhancement (understood as making morally better agents), due to limits in our ability for self-control, a fair deal of *de facto* automatic action is inevitable in our everyday lives. This means that having the inclinations conducive to the good is an important factor in enabling us to be good people and effective moral agents. Moreover, even if we consider enkratic actions as being as morally good as virtuous ones,[55] the limits of self-control mean that enkrateia will not suffice – to maximally enhance moral agency, we need to aim for virtue. Biomedical emotion modulation, when embedded in an appropriate process of reflection, can add to our ways of shaping our inclinations so that they more often than not lead us to act consistently with the good.

8.3.2. Aristotle's akrasia, enkrateia and virtue

In *Nichomean Ethics*, Aristotle discusses six moral states: heroic virtue, virtue, enkrateia, akrasia, vice and brutishness. In this and latter sections I will use four states (virtue, enkrateia, akrasia, vice) as a conceptual scaffolding to aid our discussion about the desirability of MB. Aristotle defines virtuous character in *Nicomachean Ethics* as follows:

> 'Excellence [of character], then, is a state concerned with choice, lying in a mean relative to us, this being determined by reason and in the way in which the man of practical wisdom would determine it.' (NE, II.7)

Aristotle thinks that appropriate inclinations are part of virtue and ascribes marked importance to the creation of habit: 'moral excellence [i.e. virtue] comes about as a result of habit' (NE, 1103a16-17). In the first chapter of Book II Aristotle presents his analogy of virtue to the arts largely in order to argue for virtues as sets of skills gradually developed over time through practice. By calling virtue a state of character, Aristotle does not mean that it is a feeling nor a capacity nor a mere tendency to behave in specific ways. In chapter 4 of Book II, Aristotle notes the incompleteness of the analogy with the arts, and argues that virtues additionally

[55] 'Morally good' in the thicker sense used in previous chapters: involving being both good and motivated by appropriate moral concern.

require a person to be in a particular internal state. While ‘it is possible to do something grammatical either by chance or under the guidance of another’ (NE, 1105a22-23) virtue requires more – and that more includes the right inclinations and an appropriate cooperation of inclination and reason (NE, 1105a28-30). Right inclinations, even if sufficiently fine-tuned, are not sufficient for virtue on their own. As a result, MB on its own will never be able to create a virtuous person on Aristotle’s account, and thus ‘manufacturing virtue’, is not an option – at least as far as Aristotelian virtue in concerned.

However, the right habits and inclinations do contribute to virtuous character. A virtuous person knows the good, acts according to the good and her reason is in harmony with inclinations. In Aristotle's words, the non-rational part of a virtuous person's soul ‘speaks with the same voice’ (*homophônei*, NE, 1102b28) as the rational part. In contrast to a virtuous person, an enkratic or continent person knows the good, acts according the good but needs to conquer the passions that nudge towards the bad. Akrasia involves acting against one's better judgement.[56] An akratic person is someone who because of his feelings abandons himself against correct reason.

According to Aristotle, emotion challenges reason in three ways. In both the akratic and the enkratic, it competes with reason for control over action. Second, in the akratic, it temporarily robs reason of its full acuity, thus handicapping it as a competitor for control over actions – it keeps reason from fully exercising its power.[57] Third, passion can make someone impetuous; here victory over reason is so powerful that the decision does not enter the arena of conscious reflection until it is too late to influence action.

[56] This is the standard view of akrasia. It is not universally shared, however. Perhaps the earliest philosophical discussion of akrasia is in Plato’s (1996) *Protagoras*, where Socrates argues, in effect, that akrasia is impossible, since no one ever knowingly chooses to do wrong. All apparent cases of akrasia are in fact cases of weakness of will. Another account is that of Watson (2004) who questions the clarity of the distinction between (blameworthy) akrasia and (blameless) compulsion.

[57] According to Aristotle, Socrates argued that there is no akrasia understood as weakness of will. A similar position was put forward by Hare (1952). When reason remains unimpaired and unclouded, its dictates will carry us all the way to action, save for practical obstacles. It is only the clouded judgement that makes a person akratic. I follow Aristotle in disagreeing with Socrates and Hare: I think that the example of OCD convincingly demonstrates that weak willed akrasia is possible.

Aristotle’s emotions and appetites (*pathos*) do not necessarily translate into strong psychological forces: anger is a *pathos* whether it is weak or strong. Aristotle clearly indicates that it is possible for an akratic person to be defeated by a weak *pathos* – the kind that most people would easily be able to control. Thus, it is not the strength of emotion or inclination itself, but rather the ratio between the reasoned control and the behaviour resulting from passions that matters – this latter type of akrasia involves the passions ‘overtaking’ reason.

Aristotle’s account considers virtue, enkrateia, akrasia and vice as properties of character. Consistently with this, Aristotle describes an akratic individual, an individual who more commonly than others succumbs to passions and appetites instead of following reason. Following more recent discussions on akrasia, I will focus on those concepts as applying to *actions* instead of *agents*. I will often talk about ‘desires’ and ‘inclinations’, but the same considerations apply to biomedically modifiable undepinnings of action in so far as they give rise to desires.

8.3.3. Is inverse akrasia any good?

Aristotle’s akratic agents know how they ought to act, and yet are being led astray by their desires and passions. In this section I will consider an example more suited to considering Persson’s and Savulescu’s proposals for MB: inverse akrasia (Arpaly and Schroder, 1999; Holton, 1999; Doucet, 2014). Although Aristotle did not outline such an option,[58] it seems at least possible that akrasia could be similar to virtue in its actions. In this case, the akratic agent would abandon herself to emotions, against the dictates of *in*correct reason. Inverse akrasia would require doing good while displaying the same pattern of practical reasoning as standard akrasia. A good example of inverse akrasia can be found in Mark Twain’s novel in

[58] It is possible to define morally good actions in such a way so as to rule out akrasia, thus making inverse-akrasia impossible. Aristotle’s strategy seems to be to say that since morally right action is always rational and akrasia involves irrationality, morally right action cannot be akratic: “not everyone who does something because of pleasure is… incontinent, but only someone who does it because of a shameful pleasure.” (Aristotle, *Nicomachean Ethics*, 1151b24). I am not convinced by this argument, and I think that it makes sense to talk of inverse akrasia given the structural similarity (agents' lack of control and not acting on their best judgement).

the character of Huckleberry Finn. This is how Doucet summarises the relevant details of Huck's story:

> 'Huck is an uneducated boy from antebellum Missouri with many of the values and beliefs common to that place. He does not question the moral justifiability of slavery, and he believes that slaves should be treated as property. During the course of Twain's story, Huck befriends Jim, a slave, and helps him escape. This action goes against Huck's strong belief that he ought to turn Jim in, since, as a slave, Jim is someone's lawful property. On two separate occasions, however, Huck is faced with an opportunity to turn Jim in, and on both occasions, he finds that he cannot, despite his belief that it would be the right thing to do. This causes him to feel intense regret; he berates himself for aiding in what he considers to be "theft," and believes that he has a duty to return Jim to Miss Watson, Jim's owner. Far from believing that he acted rightly, his conviction that he has repeatedly acted both weakly and badly convinces him that he is destined to remain a "bad boy".' (c.f. Doucet, 2014, pp. 3-4)

Twain's description of Huck's psychology seems coherent and can serve as an analogy for MB-induced inclinations. I would like to consider two plausible interpretations of what happens in Huck's case, and the conclusions they offer for thinking about biomedical emotion modulation.

On the first interpretation Huck acts merely on the basis of unreasoned emotional 'pull' of sympathy. Bennett (1974) argues that although in Huck's case the outcome of his action is consistent with what is morally good, his actions cannot be considered to be properly 'moral'. He emphasises that '*feelings* must not be confused with *moral judgments*' (p.124). This interpretation is supported by the fact that Huck's conviction that Jim is rightfully considered property and thus should be returned to his owner seems to be strong; moreover, Huck does not entertain reasons that would question this conviction. On this interpretation, a feeling of sympathy leads Huck to override his moral judgement. The moral review of his actions is preserved, yet moral control is not.

This appears closely analogical to Persson and Savulescu's (2008) proposals of how MB would look. Since Persson and Savulescu want to eliminate vice leading to great evils and do not mind creating permanent Ulysses' (Savulescu and Perssons, 2014a), it follows they would endorse creating Huck Finns – agents whose endorsed belief has been overridden by a strong, uncontrolled pro-social motivation. As such, their interest does not lie in creating virtue but rather substituting inverse-akrasia for vice. Although moral review may be preserved, both online and offline moral control would *ideally* be diminished. If that indeed is the rationale for their MB project, I agree with John Harris (2011, 2014a, 2014b) that it would diminish freedom. Moreover, making agents act against their better judgement and not being able to modify their behaviour seems indeed like a recipe for decline in moral agency, and as such, is not desirable where MB is to aim at creating better moral agents.

However, the difference between Huck Finn and anti-vice MB is that Huck does act on an emotional pull that is both consistent with and stems from his moral worldview, or at least part of it. This brings us to the second interpretation of Huck's case. Doucet (2014) argues that Huck's case is better considered to be a case of conflict between competing moral reasons.[59] The first time Huck decides to turn Jim in, two things Jim says cause him not to follow on his resolve: Jim calls Huck the best friend he ever had and 'the only white gentleman to ever keep a promise to him' (Twain, chapter 16). Huck's emotions and attitudes are rationally grounded as they depend on his having a series of beliefs about Jim, friendship, promises, and loyalty. Doucet argues that even though Huck acts irrationally from his own point of view, the problem is less due to the fact that he has been overtaken by an uncontrollable pang of sympathy and more to do with the fact that he failed to consider all of the reasons he has for acting. Thus, Huck sees his judgement as a 'better judgement' but in fact he fails to make an all-things-considered judgement, and the ignored reasons catch up with him; since Huck failed to consider *any* reasons for helping Jim escape when he was deliberating about what to do, yet those reasons came to his attention when he was about to act, he certainly did not

[59] A similar argument was made in Audi (1990).

consider *all* his relevant beliefs. Rather, he considered only his beliefs about slavery and property.

What should we make of this interpretation of inverse-akrasia when considering MB? The first relatively uncontroversial points are that a) we are not perfect deliberators, sometimes failing to consider all the moral reasons we have for action, b) that sometimes we act against our better judgement and c) that sometimes acting against our better judgement is conducive to the good. That is not much of a surprise. The more important point here is that although akratic action may be irrational in the sense of acting against one's best judgement, it does not mean that it is 'bizarre' in the sense of lack of responsiveness to any reasons. Akratic action, even though irrational overall, may be *more or less* irrational depending on what set of reasons it is based on. We could simply wave away the issue since the action is involuntary, but I think that this would be missing something important. Even when we act akratically and fail to consider all the relevant reasons, it *does* make a difference whether or not we act on the basis of attitudes, emotions and beliefs that are part of and consistent with our world view.

The difference in ethical significance is twofold. Firstly, given that we regularly act akratically (for support of this claim see 8.3.6.), it is better from the point of view of moral agency that those akratic actions are at least partially justified within what the agent finds important and valuable and has some actual connection to the moral reasons that the agent would endorse. This is why it makes a difference whether Huck's akratic actions stem from reasons related to the value of friendship, keeping promises and loyalty (a moral stance) or Huck's good act is motivated by reasons altogether non-moral (e.g., Jim is a good fisherman and Huck is hungry) or unendorsed considerations (e.g., Huck thinks friendship, loyalty and keeping promises is for moral weaklings and true morality consists of moral egoism unpolluted by attachment to any particular individuals). Secondly, although whether or not akratic acts are connected to moral reasons the agent has might not make a moral difference in that particular instance – insofar as the agent is acting akratically anyway – it *is* important for the development of moral agency that the akratic actions are more, rather than less integrated with our world view. We remain imperfect moral agents, and the process of developing appropriate and endorsed emotional responses, flexible control of those responses and doing so with reference to moral reasons is, well, a process. However, we know we are further from the goal when our acts are based on desires and inclinations that we in no way endorse from the moral perspective and which go against the moral reasons we find compelling.

I do not intend to suggest that akratic action constitutes the ideal moral agency we want to seek or promote. It further diminishes moral agency, and especially moral development, however, to more often act akratically on desires which we do not endorse and which give very little compelling (to us) reasons for action. Compulsory moral MB as proposed by Persson and Savulescu (2008), in contrast to voluntary agent-led MB, widens the gap between the 'motivational pulls' and the reasons we have *even in the akratic*. In doing so, it puts the agent further away not only from continent action, but also from virtue. By contrast, there is lesser *prima facie* danger of this kind from voluntary and agent-led biomedical modifications.

8.3.4. Deliberation to action

In his response to Harris' (2011) concerns, Douglas (2013) elaborates on what he calls the Kantian objection to MB. Kant held that if any action is to be morally good, it is not enough that it should conform to the moral law – it must also be done for the sake of the moral law. This requires deliberation – a deliberative review from the moral standpoint. Agents' actions after biomedical modification of emotions, insofar as MB affects conative states (what Douglas calls 'brute conformity' enhancements), were produced through non-deliberative means. As a result, the argument goes, the resultant conduct is not 'moral' and even if MB makes agents act according to the good, it is not truly 'moral' conduct.

Douglas (2013) argues that some technologically plausible enhancement might 'operate precisely by facilitating the sort of deliberation that the Kantian … takes to be necessary for moral worth' (p. 7).[60] Discussing a series of examples of biomedical and non-biomedical enhancements, he suggests that removing the influence of non-endorsed bias, such as unconscious racist attitudes (the case of

[60] A similar proposal was brought forward in Douglas' 2008 paper, in which he proposed that enhancing moral motives through eliminating known biases might be a biomedical enhancement that escapes much of the bioconservative criticism.

Andrew) or being insufficiently moved by moral motives (the case of Bryoni) could improve moral deliberation.

> ‘Andrew is a doctor working in multi-racial area. He was brought up in a racist environment and emotional responses introduced during his childhood still have a biasing influence on his conduct. For example, they incline him to take more care in treating White patients than Black patients. Andrew is aware of this aspect of his psychology and suspects it to be morally problematic. Hoping to mitigate his bias, he embarks on a new programme developed by neuroscientists. He first observes stimuli that elicit racial aversion (such as photos of mixed race couples and civil rights protests) while undergoing high resolution brain scanning to determine which neural connections mediate the aversion. Those connections are then selectively attenuated via regular sessions of transcranial electrical brain modulation. This programme significantly weakens his disposition to racial aversion and does indeed lead him to treat his Black and White patients more equally.’ (Douglas, p. 8)

In response, Harris (2011) argues that, although racist beliefs are still present in many parts of the world, traditional means of influencing moral agency such as education, legislation and public disapproval has greatly reduced the prevalence of racist behaviours. As a result, the non-biomedical means seem to be effective in reducing the racist bias and there is no need to resort to biomedical means, which are likely to be less specific in their actions and may come with side effects. Moreover, Harris (2011) points out that prejudices such as racism, sexism or homophobia are unlikely to be simple, visceral aversive responses – such as, for example, an aversion to spiders might be. Rather, prejudicial attitudes are linked to and rely on cognitive content. Thus, one may conclude, changing false beliefs and prejudices is best achieved by a combination of rationality and education and possibly biomedical cognitive enhancement.

Although Harris (2011) correctly pointed out that prejudicial attitudes are more complex than simple aversions and may be linked to, sustained by and necessarily include beliefs about facts, I think it is questionable to claim on this basis that beliefs are at the centre of prejudicial behaviour and are corner-stones of prejudicial reactions. Rather, prejudice is likely to involve beliefs about the world as well as emotions and behavioural habits all linked in the web that influences the way we perceive events and people around us, along with the ways in which we react to them. This multifactorial way of seeing moral agency especially applies insofar as we are concerned with agents' *actions* and not only professed beliefs. As a result, there is no in principle reason why changing beliefs or enhancing deliberative ability, whether by education, biomedical cognitive enhancement or policy, should be the sole or even best way of decreasing the prevalence of racist actions. Moreover, it seems that what Andrew primarily needs is not a change in beliefs – that he embarks on the neuroscientist-led examination and change of his implicit biases already suggests that he is not fatally missing in the area of knowing the good. Rather, he struggles to make the belief 'sink from top to bottom': to the level of implicit beliefs and automatic emotional reactions that to a large extent guide our everyday actions. This gap between his rationally endorsed moral beliefs and an ability to translate them into everyday behaviours is Andrew's chief problem. Harris made a similar comment about a different case in Douglas (2013), and it applies well to Andrew too – according to Harris, such agents do not lack moral goodness.

Two functions of moral deliberation need distinguishing when we talk about change in moral conduct. One has to do with a distal reason motivating MB, the ability to review its effects and justify the intervention as indeed conducive to the good. What we are talking about here belongs to what I have previously called offline moral review and control. The second function has to do with online moral control and creating conditions in which the agent is more likely to act according to the good. This second function includes the ability to revise ones attitudes, implicit beliefs and behavioural schemas in accordance with such deliberatively examined and endorsed values and general beliefs. Harris (2012) argues that deliberation, and whatever enhances deliberation, is most conducive to morality – not only in its distally motivating and justificatory role, but also in fulfilling the second function. I do not deny that deliberation oriented towards practical action, revising one's beliefs and developing new ones is also immensely valuable insofar as it translates into action. However, I doubt that it is the only (or perhaps even the most) effective way of making our rationally endorsed beliefs 'sink in.'

This empirical claim needs some more support. Psychology in its therapeutic aspect is perhaps the discipline most concerned with the kind of behavioural change we are talking about here (although usually more aimed at increasing well-being). The history of psychology reflects the above tension between the importance of insight into the sources, causes, functions and mechanics of particular behaviours (including the associated beliefs) and its aim of changing problematic behaviours and alleviating distress (Friedman, 2011). There is no need to discuss this controversy in detail. For the purpose of our discussion it is sufficient to simply refer to the gap familiar to psychotherapists: the gap between insight and behavioural or emotional change.

Admittedly, the concept of 'insight' as used in psychology and psychiatry is somewhat vague and used to describe a variety of phenomena. However, even with this limitation in mind, it seems to encompass the belief change that results from deliberation and reflection and is thus applicable to the matter discussed here. Measures of 'insight' used in cognitive-behavioural therapy, for example, include 'becoming aware of one's beliefs' or 'identification of errors in thinking' (Tang and DeRubeis, 1999). In turn, the measurement of 'insight' used in more psychodyamically oriented approaches relies on the estimate of the patient's understanding of his internal conflicts, associated problems, reoccurring behaviours and associations with previous experiences and includes the awareness of the connected beliefs (Johansson et al., 2010).

The psychoanalytical tradition typically attributed a considerable potential of therapeutic change to insight. For a long time, behaviour change was seen as a natural consequence and an integral part of a so called 'true' insight (Sandler, Dare, and Holder, 1973). This way of thinking is difficult to uphold in light of current empirical examinations of the effects of therapy, as an observable therapeutic change often occurs only some time after achieving insight – and sometimes does not occur at all (see Høgland et al., 1994). Additionally, it is unclear to what extent the change was brought about by insight alone, and to what extent by other, mediating factors, initiated by the insight or co-occurring with it. In fact, empirical research comparing the treatment outcomes of cognitive behavioural therapy with more behaviourally oriented approaches indicates that the 'cognitive' aspects of cognitive-behavioural therapy are less obviously effective than the behavioural components. Moreover, psychologists of various traditions often distinguish between what they call 'intellectual insight' and 'emotional insight', with only the latter thought to be accompanied by change in behaviour (for review see: Elliott et al., 1994). Although the concepts employed here admittedly are not very sharp, I think that even this crude distinction gets at an important observation repeatedly made by those professionally in the business of reflectively-embedded behavioural change: that there is a gap between consciously held beliefs and their impact on everyday behaviour. This gap is present in our everyday lives, but becomes especially vivid in the context of change.

There are limitations to drawing inferences from psychotherapy in discussing bettering moral agency and MB. The goal of therapy is different than moral enhancement aimed at making better moral agents. One would have to look more closely at the exact constructs evaluated to draw any strong conclusions. Discussion of particular therapeutic tools would ground a stronger argument. The assessment of particular aspects of therapy contributing to change is in its infancy, etc. As a result, the data from empirical evaluations is not sufficient to provide a basis for a strong claim that methods other than belief change should be prioritized. However, empirical evaluations of various methods of therapy still provide the best and most ecologically valid evidence we have for the effect of change in beliefs specifically – insofar as we expect the change of moral beliefs to be followed by change in action. As it stands today, the research supports the weaker claims I wish to make – that there is a gap between the reflectively endorsed beliefs and everyday behaviour, that this gap is common enough (as I suggest) to warrant attention, and that it is not best addressed by *more* deliberation.[61]

[61] One might object that this gap might be best addressed by means of *better practical* deliberation. However, I would think that therapy is, generally speaking and permitting differences in approaches, as close to good practical deliberation as we have. Moreover, patients are most often motivated to gain from therapy, and are in therapy because something in their lives needs to change. Even if the method could be improved, I think the

This problem is not new – as correctly Harris pointed out, weakness of will is a perennial problem. Harris (2013a) argues that:

> ‘Where there is weakness of will, (akrasia), the problem is not one that requires moral enhancement but something akin to ‘stiffening the sinews’ and ‘summoning up the blood’… Socrates was surely close when he saw it as a combination of knowing the good and doing the good. If and when there is a gulf between these two there may be no reliable way of filling it. Weakness of will seems to be a perennial problem but it is not the same as absence of moral emotions and no one has yet shown that emotional enhancement has any greater likelihood of bridging the gap between thought and action that [sic] anything else. Feeling the good is no closer to doing the good than is knowing the good.’ (p. 172)

I agree with Harris that merely ‘feeling the good’ does not bring us any closer to doing the good than knowing the good – for the sufficient reason that on Harris’ account (and the one I would agree with contra moral intuitionists) *feeling* the *moral* good is not possible. However, I think that much of the population has little problem with knowing the good. As argued in Chapter 4, we might disagree on some of the goals, priorities and means of achieving the good, but the scope of agreement is also substantial. Yet, despite such agreement and the awareness of the vivid presence of preventable suffering, many local and global problems that could eventually be addressed by simply doing what we already know we should remain undone. Although weakness of will is not the same as absence of moral emotions, the issue of what means are necessary to address the often quite literally fatal results of the widespread and perennial akrasia remains open. I have suggested that enhancing deliberation is not enough if we want to support moral agency in action.

discussed results are a strong indication that it is not deliberation or belief change that many of patients are lacking in order to make the needed change *even if* change in beliefs is a necessary precondition for that change.

8.3.5. Is enhancing individual moral agency self-indulgent?

What should be the place of enhancing individual moral agency in addressing problems such as global poverty and racism? Harris argues that

> ‘Douglas and indeed Savulescu and Persson … are taking an excessively (one might say ‘obsessively’) individualist view of the way to solve, not Chloe’s problem that is indeed a problem, but rather the way to solve or help solve global poverty. … Addressing the problem of global poverty … is, I suggest, insane to leave to personal altruism. We should not worry too much about Chloe’s weakness of will! (…) In a real sense it is gross self-indulgence, not to mention self-defeating, to try to address these big problems at the level of individual morality. Let’s leave poor Chloe alone and think about addressing these important problems at the level of policy and indeed of government or better, at a combined governmental, truly international, level. ’ (pp. 289-290)

The present thesis, with its focus on the individual dimensions of biomedical emotion modification, is open to Harris’ charge of self-indulgence. I am neither suggesting that issues such as global poverty, climate change and the provision of healthcare are to be left to individual altruism nor up to individual morality. The focus on individual agency need not happen to the exclusion of collective solutions.[62] The collective effort would be aided by lessening the hold of akrasia on morally good action – whether as a result of more active participation in the creation of the collective solutions Harris is talking about or even less active agreement and support for measures such as taxation. Even if governmental or supra-governmental action is what is required, someone has to make the collective solutions happen – and it will be the Chloe’s of the world who already comprise – and will continue to comprise – the bulk of the public support for such measures,

[62] Savulescu and Persson also propose that MB happens alongside collective and political solutions.

who campaign and pressure the governments, who create easier ways for the less concerned to add to the collective effort and who spearhead that change.

MB alone will not solve the great global problems, but facilitating individuals acting on the moral beliefs they already hold is a valuable way of working towards that change. The means that facilitate moral agency are varied: increasing political participation and awareness, creating systems that direct and compound the individual efforts, increasing motivation for acting using public campaigns, making MB available, facilitating a society in which basic needs are satisfied so individuals do not have to focus on achieving basic security and freedom for themselves. MB via emotion modulation is only one of those methods.

One could argue that perhaps it is self-indulgent to focus on biomedical means where the traditional means of moral enhancement provide, in Harris' (2011) words, 'a blueprint' for effective interventions aimed at making better moral agents. Perhaps we only need more of the same. However, I think that the reasons to do the kind of research that lays the foundation for MB do not only come from the response to great moral issues. The very idea of MB via emotion modulation came[63] *after* the booming neuroscience finally extended to investigations of the social aspects of emotions, cognition and behaviour and provided us with some new insights into its workings.

The research will likely continue to be fuelled by the need for addressing the economic and human impacts of emotion-related conditions such as clinical depression and anxiety, as well as other conditions such as Autism, the interest in addressing problems related to crime, and scientific curiosity. We might as well use this research to aid us in understanding the necessary and motivationally effective underpinnings of pro-social and moral action and apply the gained knowledge and tools in support of autonomy and moral agency. Biomedical emotion modulation as applied to aiding moral agency is no different here than biomedical cognitive enhancement. We also have reliable non-biomedical means of enhancing cognition, but we do not expect cognitive enhancement to either replace traditional education

[63] Or re-surfaced with renewed force.

or to provide knowledge or wisdom. Rather, we expect that it helps individuals to do what they do anyway, or want to do, more effectively. I suggest that we extend the same reasonable (if modest) expectations to MB.

.

8.3.6. Why enkrateia will not suffice

In the last two sections I have outlined some of the reasons why more deliberation is perhaps not the only thing that we need to enhance moral agency, and that enhancing individual moral agency is valuable. In this section I will outline another reason why emotion modification may be of interest in enhancing moral agency. Specifically, I will argue that self-control alone *cannot* deliver outcomes *as good as* self-control together with emotion modification – that if we are thinking about enhancing moral agency, we need to think about modifying emotions, desires and habits because self-control will not suffice.

In his paper comparing various putative modes of enhancing moral agency, Douglas (2014) asks us to consider a series of cases that increase what he calls ‘moral conformity’, including the case of Bryony and Chloe:

> ‘Bryony is a student from a wealthy family. She suspects she ought to do more to help the global poor. She does occasionally do something to help, for example, giving small amounts to support famine relief when approached by charities, but most of the time, the world’s most unfortunate are far from her thoughts, and when they do cross her mind, she has trouble drumming up the sort of sympathy that might motivate greater sacrifices on her part. In an attempt to remedy this, she sets up her television so that it regularly displays disturbing and graphic images of the effects of poverty, though for such brief periods that she does not consciously recognise them. Nevertheless, through subliminal effects, the images do increase her feelings of sympathy, and these feelings stimulate her to make a large donation to Oxfam.
>
> Like Bryony, Chloe is a student who suspects she ought to do more to help the global poor, but has trouble drumming up much sympathy for them. In an attempt to remedy this, she goes to her local library and

> borrows a number of books containing first-hand accounts of life in poverty. Reading and reflecting on this literature augments her feelings of sympathy, and these feelings stimulate her to make a large donation to Oxfam.' (p.5)

In considering cases such as those previously mentioned by Andrew, Bryony and Chloe, Douglas (2014) draws an analogy between biomedical and non-biomedical means of modulating affective and conative states. He contrasts 'brute conformity enhancements' with 'deliberative conformity enhancements' (p.78). While the former modify conative and affective states without deliberation as the proximal cause of modifying the said states, the latter might involve moral reasoning, introspective reflection on one's moral failures, or calm moral discussion. Douglas uses this conceptual groundwork to more closely consider the following claim:

> '(The Moral Worth Claim) For all brute conformity enhancements likely to be developed in the medium-term future, whenever an agent has a choice between pursuing that conformity enhancement or achieving the same increment in moral conformity via a typical deliberative conformity enhancement, adopting the brute conformity enhancement will result in less morally worthy conduct.' (Douglas, 2014, p. 80).

In his response, in *Moral Progress and Moral Enhancement*, Harris (2013b, 2014b) clarifies that he does not consider 'brute conformity enhancement' to be impermissible and if such enhancement would be effective it would be welcomed. Rather, he argues that such conformity enhancements are not specifically 'moral'. Chloe, according to Harris 'needs … determination, not goodness.' (p. 290) As I indicated before, I agree with Harris. Unfortunately, the space between knowing the good and doing the good is not *entirely* inhabited by freedom – insofar as it relates to the ability to put our better judgement into practice, that space is also inhabited by lack of freedom. Harris recognises the problem of akrasia, yet seems not to see emotion modulation as a valuable response and advocates institutional level solutions, cognitive enhancement and self-control ('summoning up the blood' and 'determination') instead. In previous sections I have outlined why enhancing individual agency via increasing agents' abilities to deal with akrasia is valuable. In the next paragraphs I will consider what I think is a strong reason why cognitive enhancement and self-control are not enough.

Recall that according to Aristotle, enkrateia involves having inclinations that are not conducive to the good, but doing good nevertheless while exercising self-control. This involves what I have previously called effective online control. However, our ability for effective online control to change the course of action is limited. This ability is limited by various constraints, including time constraints (many of the decisions for actions are and need to be made quickly), by the efficiency trade-offs (even if I had time to reflectively consider whether to give money to a panhandler every time, spending two hours giving it deep consideration means I am not doing other valuable things), the limited degree to which we can exercise self-control over specific desires (for example, a dieting person may be able to overcome her craving for chocolate on most but not all occasions, we might be able to inhibit acting on a craving for chocolate more than the craving for crisps, etc.) and the limited amount self-control resources. For the purpose of this argument it is sufficient to consider the latter limitation.

A line of research pursued by Baumeister and colleagues indicates that the type of self-control involved in resisting temptation requires effort, and that exerting such effort diminishes the ability to resist further temptations. For example, Muraven, Tice, and Baumeister (1998) demonstrated that that when a situation demands two consecutive acts of self-control, performance on the second act is frequently impaired. The impairment is present even if quite different spheres of self-control are involved (e.g., an initial act of stifling or amplifying one's emotional response led to a subsequent reduction in ability to work through pain and fatigue while squeezing a hand grip, and a brief thought suppression task weakened subsequent persistence on a task involving solving a puzzle). Such research suggests that many widely different forms of self-control draw on a common resource and that such a resource might be depleted. Researchers suggest that the metaphor of a muscle well describes the effects demonstrated in research on self-control; although repeated practice increases the available self-control resources, effort diminishes the resources available.

The conclusions from Baumeister et al's (1998) research also apply to moral agency. Andrew, the doctor with implicit racist attitudes from Douglas' paper, might in principle be able to control his racist inclinations by noticing when his behaviour is impacted and inhibiting acting on them. However, spending his self-control resources on mitigating this bias means that he cannot allocate his self-control to other pursuits, including moral deliberation and action. Exercising online all-things-considered judgement is a very resource-demanding way of making decisions. Similarly, action guidance that relies on self-control is also effortful and depletes the scarce self-control resources. Even if acting enkratically is otherwise as morally good as acting virtuously, the limited self-control resources mean that we would always have strong reasons to adjust fast heuristic processes that give rise to inclinations to act (such as habits, and emotions) in a way that is most often aligned with moral outcomes. This is not to say that moral review and the ability to adjust our actions online is unimportant. Rather, the limited nature of self-control resources means that adjusting automatic reactions is a necessary part of effective agency.

The point I am making is not philosophically sophisticated. However, if we are interested in agents that are able to act according to their assessment of what is good and have more cognitive resources to spend on deliberating about the good, we have a strong *prima facie* reason to adjust emotions, inclinations and habits. Biomedical emotion modification is one way of making such adjustment. Moreover, the fact that the adjustment does not require effortful deliberation to *produce* the change in conative and affective states (although deliberation remains necessary in establishing what are the inclinations conducive to the good and fine-tuning) is in this context a clear advantage over more laborious habit-formation. Since the justificatory and action guiding role of deliberation remains intact, my argument evades Harris' concern about Douglas' position that since '"once the enhancement has been initiated, there is no further need for cognition", then the morally enhanced action is effectively automatic, unconscious and therefore unintended.' (Harris, 2013b, p. 179) Thus, the limitations of self-control resources are a strong *prima facie* reason against a view that inclinations that require frequent overcoming are not a problem if the effortful control over them is effective.

Enkrateia might be as good as virtue all other things being equal – it is just that 'all other things' are never equal.

Aristotle seems to take a similar stance when he emphasises the importance of habits. But it is not only that we cultivate virtue by simply practicing virtue; rather we cultivate both the internal states of virtue as well as the skills necessary for moral action by practicing the external actions of virtue. In cultivating those internal states, we make the external actions of virtue easier to perform:

> '…by abstaining from pleasures we become temperate, and it is when we have become so that we are most able to abstain from them; and similarly too in the case of courage; for by being habituated to despise things that are terrible and to stand our ground against them we become brave, and it is when we have become so that we shall be most able to stand our ground against them.' (*Nicomachean Ethics*, 1104a33-b3).

However, the cultivation of the internal states that allow and make virtue is not easy, and the thus enkratic and akratic actions (which interests me) and characters (which interests Aristotle) abound.

Whether and to what extent the biomedically produced approximation of virtuous inclinations can be achieved remains to be seen. Douglas (2014) expressed a further concern about the moral worth of moral conformity enhancements achieved by modifying the underpinnings of inclinations – a concern related to reliability. The first concern is that 'brute conformity enhancements' will be more contingent on the circumstances than their deliberatively achieved counterparts. This suggestion is, I think, very likely to be true. However, as I have previously argued (Pacholczyk 2011), context-sensitivity should be taken as a given – while greater inclinations to feelings of sympathy might be more conducive overall to moral outcomes in some circumstances and not others. Some of those circumstances will be rare or not foreseeable, such as the case of the amateur emergency surgery in an example brought forward by Harris (2014a). In many cases however, the effects are foreseeable, and often agents have a good indication of which of their current inclinations are not conducive to the good. Douglas (2014) gives the example of an emergency medic 'surrounded by severe pain and suffering' (p. 14). Other examples include a nurse who suffers from burnout (Pacholczyk 2011), a person who cares for a chronically and severely ill family member and cannot cope with the witnessed suffering, or a surgeon. In fact, medical education, especially of some specialties, may be seen to include selectively impairing the exercise of empathetic ability, with the goal of the increasing ability to act in the presence of suffering and sights that are difficult to see for the unaccustomed eye and mind. This cognitive and emotional 'skill' is developed over years of education and practice in an Aristotelian fashion – by exposure coupled with deliberation. When circumstances change, however – for example, if the surgeon decides to change his profession to one requiring sustained sensitivity to others' suffering or a soldier finishes military service – the profile of emotional reactions may need changing and the transition may be aided or at least eased by biomedical means. As a result, the context-sensitivity of what inclinations are conducive to the good is not a damning a problem for a voluntary agent-led emotion modulation.

The second worry raised by Douglas (2014; see also Pacholczyk, 2011) is that 'brute enhancements' are more sensitive to the magnitude of the transformation than the deliberative alternative:

> 'Whether tendencies towards impulsive violence and indifference to the suffering of strangers impede moral conformity depends on the degree to which those tendencies are present. For example, though a strong tendency towards impulsive violence is unlikely to be conducive to moral conformity, a milder tendency of the same kind may well be conducive to it, for example, because it helps to prevent excessively submissive conduct.' (p. 14)

This second consideration, although resting on a correct observation, is also not a strong objection against MB if it is led by an agent and subject to post-intervention review and modification. Moreover, seen from the perspective of limited self-control resources, biomedical emotion modification does not need to result in inclinations that *reliably* lead to the good – it only has to produce the effect that is *somewhat more likely* to lead to morally good outcomes than in the case of non-enhancement.

Insofar as we see moral enhancement as aimed at making better moral agents, we are concerned with moral action. Risking stating the obvious, if we are concerned about bettering moral action, we should be concerned with both parts of 'moral action': the 'moral' and the 'action'. As Harris correctly points out, akrasia is not a specifically moral problem. However, a problem that afflicts rational agency is also a problem that afflicts our moral agency, and those who are concerned with moral agency also need to be concerned with making it more possible for agents to act on their moral beliefs.[64] Many of the ways that aid both achieving the good and moral agency are already here and were listed by Harris and include moral education, making it easier for agents to participate in collective efforts of making substantial change, legislation and policy, systems of incentives and disincentives, etc. However, if we are concerned with moral agency specifically and agency in general, the 'cognitive' means of enhancement alone are not going to deliver better solutions to the problem of weak-willed akrasia than cognitive and emotional modulation together.

8.3.7. Conclusions

In this section (8.3) I have suggested that MB, insofar as it is agent-led and embedded in appropriate reflection, can be a desirable way of contributing to enhancing moral agency. Agent-led biomedical modifications in their inclinations are best seen to be analogical to a set of emotion modulation mechanisms used in everyday life, as were considered by Douglas (2014) and are best ethically assessed in that context. MB is different from some of those mechanisms insofar as it might involve more stable changes in inclinations than case-by-case emotion modulation, and so this aspect resembles creating right habits and right internal states.

We often rely on our habits and there is a good reason for that – our self-control is limited. The fact that self-control resources are limited is a good reason to look for ways which would help to adjust inclinations so that they have to be overcome less,

[64] Unless one is of an opinion that most peoples' moral beliefs are mistaken and facilitating peoples' autonomy would lead to more evil than good. I do not share this pessimism. Usually facilitating individual autonomy is seen as a desirable thing, even if some morally relevant actions are restricted or punished by law. It leaves the burden of argument on the critic to demonstrate that their pessimism is justified.

rather than more often. Although pharmacological means of biomedical enhancement are blunt tools and their impact will not amount to virtue, perhaps they could assist us in achieving the inclinations conducive to the good. In contrast to developing habits conducive to the good, which is effortful and time-consuming, MB would likely require less time and cognitive resources. The considerations related to comparative cost-effectiveness of specifically biomedical interventions discussed in Chapter 2, taken together with the fact that traditional habit change is effortful and time consuming, make biomedical emotion modification a tool to consider in aiding moral agency.

As discussed in chapter 2, whether or not moral modification amounts to moral enhancement understood as making a better moral agent is context dependent and is best assessed in reference to a particular agent and their situation. Consequently, such moral enhancement is best pursued via the process of voluntary agent-led moral modification. In this process, the agent deliberates from a moral standpoint (as discussed in chapter 3) and with reference to their goals and life plans (as discussed in Chapter 7) decides what changes in inclinations would facilitate acting according to the good. As argued in section 8.2. of this chapter, the modifications continue to be subject to moral review. Although we can disagree about what ends to pursue and means to choose (see: Chapter 4), ultimately we continue to be moral agents that have to make those choices. Agent-led biomedical modification that facilitates the ability of moral agents to act according to their endorsed moral beliefs would in my view constitute improvement in moral agency and thus moral enhancement, in Harris' words, properly so called.

Moreover, finite self-control resources limit our ability to do good enkratically, which means that changing our inclinations to be more conducive to the good than the unmodified ones could be considered to be a legitimate way of moral enhancement, These considerations lower the threshold at which emotion modulation can meaningfully aid moral agency and make for a prima facie attractive option all-things-considered – resulting in a more modest, but also more achievable goal for emotion modulation as applied to enhancing moral agency.

There is no 'magic pill' that would make people act in a more moral way and the effects of every enhancement are going to be dependent on already held beliefs, the agent’s existing ability for self-control and context. This, however, is not a strong argument against biomedical attempts at emotion modulation, once we consider it to best be agent-led, voluntary and embedded in deliberation. Such biomedical emotion modulation can promote self-governance and moral agency. In contrast, compulsory MB looks much less desirable in this context. I have argued that in many cases we act in a *de facto* weak willed way, but even in the akratically acting agent there are ethical reasons against the compulsory use of MB. Even though the problem of akrasia is not a specifically moral one, it is one to be considered when we talking about pursuing moral enhancement understood as making better moral agents. I have argued that such moral enhancement can include helping people in circumstances similar to speculative cases of Chloe and Andrew in this chapter, people who know the good but experience problems with making their endorsed moral belief ‘sink in’ to the level of action. Contrary to some commentators, in section 8.3.5 I argued that such conceived moral enhancement is a valuable addition to our toolkit.

Conclusions

The recent boom in neuroscience investigation of sociality stimulated a great deal of philosophical and ethical debate. Public intellectuals, psychologists and philosophers began to examine the possibility of enhancing our moral faculty and behaviour to address the problems that humanity faces (Persson and Savulescu, 2008; Rifkin, 2010a; Baron-Cohen, 2011; Churchland, 2011; Savulescu and Persson, 2012b; Harris, 2011). The capacity for empathy has enjoyed much attention, with several thinkers proposing that an increase in empathy will pave the way for a better tomorrow.

Those concerns were reflected in the ethical debate on the desirability of MB and ethical issues raised by its potential application. Persson and Savulescu (2008) argued that since cognitive enhancement and increased access to technology means that the society is increasingly exposed to the risk of large scale harm perpetuated by a small minority, we have a reason to complement cognitive enhancement with MB aimed at preventing such harm. They proposed enhancing social sentiments and capacities such as empathy and sense of fairness and argued that MB could help to solve the great problems facing humanity, such as climate change and poverty. Douglas (2008; 2013; 2014) argued that direct emotion modulation using biomedical means could be a permissible way of moral enhancement understood as making better moral agents.

However, the idea of MB via direct emotion modulation has also attracted criticism. Among others, critics raised concerns about the plausibility of MB related to low specificity or weak effects of the proposed interventions, questioned whether the modification of pro-sociality translates into enhanced moral agency and raised doubts related to moral disagreement. Moreover, critics pointed out that MB could have a negative impact on our identity and freedom, and so might not be desirable. The aim of the current work was to contribute to the ethical discussion of MB by examining the claims and arguments put forward both by its proponents and critics.

This work aimed to explore what ethical issues should one consider when deciding whether and how to use biomedical means of moral modification. Chapters 1-4 focused on the conceptual issues and the plausibility of MB. Chapters 5-7 addressed objections and doubts raised about the ethical desirability of MB. Chapters 7 and 8 examined arguments related to the potential impact of MB on freedom and moral agency.

Chapter 1 addressed the question of what can be considered to be moral enhancement by examining the ambiguity of the phrase 'moral enhancement' and clarifying what can be meant by 'moral' and 'enhancement'. In section 1.2 I have outlined the three uses of 'moral' and proposed that 'moral enhancement' can refer to any enhancement that is morally desirable, interventions aimed at making morally better agents and enhancement in the moral sphere that is beneficial to the agent. In section 1.3 I have outlined the objections to the normative strength of the treatment-enhancement distinction and, finding the arguments convincing, proposed to understand 'enhancement' widely as 'improvement' and proceeded on this basis. This meant that the ethical assessment of MB might include MB utilized both in medical-therapeutic and non-therapeutic contexts. In section 1.4 I have argued that the additional level of moral consideration includes the assessment of the overall moral permissibility of the specific way MB is used.

Thus, in this first conceptual chapter I have proposed that we consider 'moral enhancement' widely, as biomedical interventions that modify the moral sphere in addition to those aimed at improving moral agency specifically. I have also proposed that in the debates sprang out by the prospects of modifying the underpinnings of morality and sociality, we should also attend to cases where the modification does not aim at creating morally better people, but rather is done on the basis of other, e.g. prudential, considerations. This means that an enhancing intervention in the sphere relevant to moral function may bring a dis-enhancement in moral agency and questions about what is and is not conducive to morality in given context and for certain occupations (e.g. military personnel, doctors, care professionals). I have proposed that the use of the term 'biomedical moral modification' (MB) might be more appropriate, given various possible goals of such modifications and pending the assessment of the numerous, often context-dependent, factors that influence the assessment whether a given intervention is morally desirable.

In Chapter 2, I have asked whether specifically biomedical means of modifying the moral sphere are likely to be effective, and what kinds of effects can be expected after MB and thus what goals are in this context reasonable. I argued that MB is plausible, but that we should revise our goals and expectations in discussion MB. The aims such as eliminating large scale harm, reducing poverty or addressing climate change (Rifkin, 2010; Persson and Savulescu, 2008) mean that we expect too much of MB. Instead, I proposed that we see MB similarly to biomedical cognitive enhancement: as a small but significant improvement in underlying cognitive or affective processes that will help us to better do what we want. With our goals revised and expectations adjusted, the examination of the effects of oxytocin on social function revealed that sometimes the effects of biomedical modification can be significant. This gives a *prima facie* reason to discuss the potential for biomedical modification for the purposes of making better moral agents.

Chapter 3 considered whether or not biomedically increasing pro-social emotions and attitudes such as empathy would likely better moral agency. I argued that biomedical modification of affective capacities and reactions, such as increasing empathy or decreasing anger, is not sufficient to make us better moral agents. The equivocation of the moral and the pro-social is unjustified for three main reasons. Firstly, anger and empathy are multi-purpose and a modification in each can lead both to morally desirable and undesirable outcomes. For example, modification of empathy could be conducive to moral outcomes but could also lead to empathetic distress, a state rarely conducive to good outcomes and harmful to the agent. Secondly, even where biomedical emotion modulation would lead to better moral outcomes, it does matter that the behavioural change relates in a right way to the moral reasons we have. As a result, creating moral agents requires that the change of abilities is appropriately embedded in agents' reflection about the good.

Chapter 4 enquired about the way in which the presence of moral disagreement affects the application of MB. If moral enhancement aimed at aiding moral agency needs to be connected to agent's moral reasons and ideas of the good, and since agents' beliefs of what is good may differ, one can raise an objection that moral disagreement undermines the moral enhancement project. Section 4.1 explored the limitation of the scope of the argument that MB may be implausible in the presence of moral disagreement. In section 4.3 I examined the implications of fundamental moral disagreement for MB and argued that although moral disagreement may pose a challenge for evaluation of MB applications, there is no reason to favour the *status quo* in the outcome of this deliberation.

After looking at the sources of moral disagreement and implications of fundamental moral disagreement for the MB project, I argued that even the presence of fundamental moral disagreement does not mean that we should abandon out moral beliefs and pursue our moral projects. The axiological difference may create some problems in evaluation of whether a purported improvement in moral agency indeed achieves this goal, but on practical level such differences are usually accommodated by the political process. Moreover, there is no good reason to favour the *status quo,* as the same disagreement can exist about the current level of traits potentially modifiable by MB. I concluded that although moral disagreement may pose a challenge for evaluation of MB applications, it does not give us a strong reason to forgo MB generally, and enhancing moral agency using biomedical means specifically,

Further chapters explored concerns that even if effective MB is plausible, it is not desirable. Chapter 5 examined objections related to medicalization and asked whether using specifically biomedical means of moral modification gives rise to a strong ethical reason to forgo using MB. After examining arguments brought forward by critics of medicalization, I argued that the process of medicalization is in itself normatively neutral, and only acquires meaning on the basis of what medicalization allows us to do and what costs it brings with it. To provide some counterweight to the outlined criticism, I discussed the benefits of seeing a problem as medical and argue that medicalization of an issue is desirable where it better allows us to get what we want. I concluded that the general critique of medicalization fail to give us a strong reason to forgo MB, and that the assessment of whether medicalizing a certain trait of function should be done on case-by-case basis.

Chapter 6 explored the concerns raised in relation to identity and aimed to examine whether or not narrative identity theories can ground a strong ethical objection to MB. I examined Schechtman's objection to deep brain stimulation and argued that her account is insufficient to ground the critique she makes. In search of another account that better outlines what is specifically 'narrative' about narrative identity, I examined the potential of Ricoeur's theory to ground ethical evaluation of MB. I concluded that the same what makes Ricoeur's narrative identity specifically narrative, the criterion of homophony, problematically relates to claims about what we should (ethically speaking) do. I extended this examination to a more general critique of the *strong ethical narrative* thesis and concluded that narrative identity theories face serious problems in providing ethical action-guiding reasons.

The last two chapters discussed the impact of MB on freedom and agency. Chapter 7 asked to what extent issues raised in relation to freedom in the discussion of Savulescu and Persson's (2012a) thought experiment called the God Machine call the desirability of MB into doubt. Using a series of thought experiments to tease out exactly in what way the God Machine could endanger freedom (various cases of overdetermination of agents actions), in sections 7.3-7.5 I have argued that the main problem with the God Machine is that it breaks the link between agents own reasons for action and the outcome in the world. Section 7.3 explicated the issues in relation to moral luck and prise and blame, section 7.4 used Frankfurtian analysis in order to argue that an important aspect of the God Machine's threat to freedom lies not in endangering free will generally but rather by undermining specifically the ability to form a *will of our own*. In section 7.5 argued that the problem does not necessarily lie in the fact that the God Machine is a case of overdetermination, and that the more plausible uses of MB that involve overdetermination would be significantly less problematic. In section 7.6 I examined Sparrow's (2014) objection in the context of freedom as non-domination and argued that non-domination theory of freedom is ill fitted to ground a robust critique. In section 7.7 I used an analysis of the application of Mill's harm principle and argued that the God Machine would be an undesirable way of achieving a morally better world because it might adversely affect the way the desires and inclinations for action are formed. This discussion shed light on the factors that are to be considered when evaluating the impact of MB on moral agency. However, I have argued that the conclusions taken from the consideration of the God Machine thought experiment can only bring our attention potentially important aspects, but due to the degree of abstraction and important differences between the God Machine and MB, the arguments related to the God Machine should be transferred with much caution to the ethical assessment of real-world MB. I have concluded that the arguments raised in relation to the God Machine thought experiment fail to call the desirability of real world, hopefully voluntary and agent led MB into doubt.

Chapter 8 discussed issues arising in plausible applications of MB in order to examine whether real-world MB would endanger or could also facilitate moral agency. In order to answer this question, I critically examined Harris' (2011) objection that MB would be beyond moral review. Using the example of obsessive-compulsive disorder, I argued that even biomedically induced compulsions would not necessarily be beyond online and, more importantly, offline moral review. Consequently, real-world MB would likely allow further moral deliberation about the effects of MB. Further, I considered MB in the context of an Aristotelian framework, and argued that the limitations of self-control mean that effective moral agency could be aided by the modification of dispositions. In so far as MB offers the possibility to modify dispositions and emotions and would be embedded in appropriate moral reflection, it can result in enhancing moral agency.

References

Abramowitz, J. (1998). Does cognitive-behavioral therapy cure obsessive-compulsive disorder? A meta-analytic evaluation of clinical significance. *Behaviour Therapy, 29*, 339-355.

Adams, R.M. (1985). Involuntary sins. *The Philosophical Review*, *94,* 3-31.

Adler, J. M., Skalina, L. M., & McAdams, D. P. (2008). The narrative reconstruction of psychotherapy and mental health. Psychotherapy Research, 18(6), 719-734.

Agich, G. (1983). Disease and value: A rejection of the value-neutrality thesis. *Theoretical Medicine, 4,* 27–41.

Agid Y, Schupbach M, Gargiulo M, Mallet, L., Houeto, J. L., Behar, C., Maltete, D., Mesnage, V., & Welter, M. L. (2006). Neurosurgery in Parkinson's disease: The doctor is happy, the patient less so? *Journal of Neural Transmission* (s70), 409-414.

Alpert, H. (1938). Durkheim's functional theory of ritual. *Sociology and Social Research, XXIII*, 103-108.

Angus, L. E., and McLeod, J. (2004). Toward an integrative framework for understanding the role of narrative in the psychotherapy process. In L. E. Angus and J. McLeod (Eds.), The handbook of narrative and psychotherapy: Practice, theory, and research (pp. 367-374). Thousand Oaks, CA: Sage Publications, Inc.

Arendt, H. (1958). *The human condition*. Chicago: Chicago University Press.

Aristotle. (2004). *Nicomachean ethics*, ed. Crisp, R. Cambridge: Cambridge University Press.

Aronson , E. (2007). *The social animal*, 10th ed. New York, NY: Worth/Freeman.

Arpaly, M. and Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies, 93*(2), 161-188.

Arrington, R. (1982). Advertising and behaviour control. *Journal of Business Ethics, 1,* 3-12.

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

Avdi, E. and Georgaca, E. (2007).Narrative research in psychotherapy: A critical review. *Psychology and Psychotherapy: Theory, Research and Practice*, 80, 407–419.

Avorn, J. (2012). Two centuries of assessing drug risks. *The New England Journal of Medicine. 367*(3), 193-197.

Ayer, A. (1936 [1952]). *Language, truth and logic*. New York, NY: Dover.

Baker, L. R. (2000). *Persons and bodies: A constitution view*. Cambridge, England: Cambridge University Press.

Baron-Cohen, S. (2011). *The science of evil*. New York, NY: Basic Books.

Barraza, J. and Zak, P. (2009). Empathy toward strangers triggers oxytocin release and subsequent generosity. *Annals of the New York Academy of Sciences, 1167,* 182-189.

Barsky, A. and Boros, J. (1995). Somatization and medicalization in the era of managed care. *Journal of the American Medical Association, 274,* 1931-1934.

Bartov, O. (1992). *Hitler's army*. Oxford: Oxford University Press.

Bartz, J., Zaki, J., Bolger, N., Hollander, E., Ludwig, N., Kolevzon, A. and Ochsner, K. (2010). Oxytocin selectively improves empathic accuracy. *Psychological Science, 21*(10), 1426-1428.

Batson, C. and Weeks, J. (1996). Mood effects of unsuccessful helping: Another test of the empathy-altruism hypothesis. *Personality and Social Psychology Bulletin, 22*(2), 148-157.

Batson, C., Dyck, J., Brandt, J., Batson, J., Powell, A., McMaster, M. and Griffitt, C. (1988). Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis. *Journal of Personality and Social Psychology, 55*(1), 52.

Batson, C., Klein, T., Highberger, L. and Shaw, L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology, 68*(6), 1042-1054.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U. and Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58*(4), 639-650.

Baylis, F. (2011). "I am who I am": On the perceived threats to personal identity from deep brain stimulation. Neuroethics. doi: 10.1007/s12152-011-9137-1.

Bell, J., Reed, K., Ashcroft, R., Witton, J. and Strang, J. (2012). "Treating opioid dependence with opioids: exploring the ethics." In *Addiction Neuroethics* (Ed. Carter, A., Hall, W. and Illes, J.). London: Academic Press.

Ben-Ari, E. (1998). *Mastering soldiers: Conflict, emotions, and the enemy in and Israeli military unit.* New York, NY: Berghahn Books.

Bennett, J. (1974). The conscience of Huckleberry Finn. *Philosophy, 49*, 123-134.

Berlin, I. (1958). "Two concepts of liberty," In *Four Essays on Liberty*. Oxford: Oxford University Press.

Berton, O. and Nestler, E. (2006). New approaches to antidepressant drug discovery: Beyond monoamines. *Nature Reviews Neuroscience, 7,* 137-151.

Blackburn, S. (1998). *Ruling passions*. Oxford: Clarendon Press.

Blitz, M.J. (2010). Freedom of thought for the extended mind: Cognitive enhancement and the constitution. *Wisconsin Law Review, 4*, 1049-1117.

Bloom, P. (Sept 10 2014). "*Against empathy*." Boston Review. Accessed Sept 27 2014 at http://www.bostonreview.net/forum/paul-bloom-against-empathy.

Blumenthal, J.A. (2005). Does mood influence moral judgment-an empirical test with legal and policy implications. *Law & Psychology Review*, *29*(1).

Bohman, J. (2007). *Democracy across borders: from Demos to Demoi*. Cambridge, MA: MIT Press.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science, 44*(4), 542-573.

Bordo, S. (1998). "Braveheart, Babe and the contemporary body." In *Enhancing human traits: ethical and social implications* (Ed. Parens, E.).: Washington, DC: Georgetown University Press.

Bos, M.W., Dijksterhuis, A., and van Baaren, R.B. (2008). On the goal-dependency of unconscious thought. *Journal of Experimental Social Psychology, 44,* 1114-1120.

Boyd, R. (1988). "How to be a moral realist." In *Essays on moral realism* (Ed. Syre-McCord, G.) Ithaca and London: Cornell University Press.

Brewster, M. (Nov 26 2014) "Vets needing PTSD benefits face dizzying paperwork, wait: auditor." *The Hamilton Spectator*, Accessed Nov 26 2014 at http://www.thespec.com/news-story/5156316-vets-needing-ptsd-benefits-face-dizzying-paperwork-wait-auditor/.

Brickman, P., Rabinowitz, V., Karuza, J., Coates, D., Cohn, E. and Kidder, L. (1982). Models of helping and coping. *American Psychologist, 37*, 368-384.

Brody, H. (1994). "My story is broken; can you help me fix it?": Medical ethics and the joint construction of narrative. *Literature and Medicine, 13*(1), 79-92.

Brown, P., Zavestoski, S., McCormick, S., Mayer, B., Morello-Frosch, R., & Altman, R. (2004). Embodied health movements: New approaches to social movements in health. *Sociology of Health & Illness, 26*(1), 50–80.

Burke, M. (2011). "Resisting pathology: GID and the contested terrain of diagnosis in the transgender rights movement." In *Sociology of Diagnosis* (Eds. McGann, P. and Hutson, D.). Bingley: Emerald Group Publishing.

Campbell, A. (2008). Attachment, aggression, and affiliation: The role of oxytocin in female social behavior. *Biological Psychology, 77*(1), 1-10.

Canning, J. (1996). *A History of Medieval Political Thought: 300–1450*. London, UK: Routledge.

Carritt, E. (1947). *Ethical and political thinking*. Oxford: Oxford University Press.

Carter, A. and Hall, W. (2008). Informed consent to opioid agonist maintenance treatment: Recommended ethical guidelines. *International Journal of Drug Policy, 19*(1), 79-89.

Carter, A. and Hall, W. (2012). *Addiction neuroethics: The promises and perils of neuroscience research on addiction.* Cambridge: Cambridge University Press.

Carter, I. (2008). "How are power and unfreedom related?" In in *Republicanism and Political Theory*, Laborde, C. and Maynor, J. (eds.). Malden, MA: Blackwell Publishing.

Chan, S. and Harris, J. (2011). Moral enhancement and pro-social behaviour. *Journal of Medical Ethics, 37*(3), 130-131.

Chisholm, R. (1966). "Freedom and action." In *Freedom and determinism* (Ed. Lehrer, K.) New York, NY: Random House.

Christman, J. (1991). Autonomy and personal history. *Canadian Journal of Philosophy 20*, 1-24.

Churchland, P. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton, NJ: Princeton University Press.

Clarke, A., Shim, J., Mamo, L., Fosket, J. and Fishman, J. (2003). Biomedicalization: Technoscientific transformations of health, illness, and U.S. biomedicine. *American Sociological Review*, *68*, 161-194.

Cohen , J.B. and Goldberg , M.E. (1970). The dissonance model in post-decision product evaluation. *Journal of Marketing Research* , *11*, 315 – 321.

Cole-Turner, R. (1998). "Do means matter?" In *Enhancing human traits: ethical and social implications* (Ed. Parens, E.). Washington, DC: Georgetown University Press.

Conrad, P. (1975). The discovery of hyperkinesis: Notes on the medicalization of deviant behavior. *Social Problems, 23*, 12–21.

Conrad, P. (1992). Medicalization and social control. *Annual Review of Sociology, 18*, 209-232.

Conrad, P. (2000). "Medicalization, genetics, and human problems." In *The Handbook of Medical Sociology, 5th ed* (Eds. Bird, C., Conrad, P. and Fremont, A.). Upper Saddle River, NJ: Prentice-Hall.

Conrad, P. (2005). The shifting engines of medicalization. *Journal of Health and Social Behavior, 46,* 3-14.

Conrad, P. (2007). *The medicalization of society: On the transformation of human conditions into treatable disorders.* Baltimore, MD: John Hopkins University Press.

Conrad, P. and Potter, D. (2000). From hyperactive children to ADHD adults: Observations on the expansion of medical categories. *Social Problems, 47,* 559-382.

Conrad, P. and Schneider, J. (1980). *Deviance and medicalization: From badness to sickness.* Philadelphia, PA: Temple University Press.

Conrad, P., Mackie, T. and Mehrotra, A. (2010). Estimating the costs of medicalization. *Social Science and Medicine, 70,* 1943–1947.

Costa, M.V. (2007). Freedom as non-domination, normativity, and indeterminancy. *Journal of Value Inquiry, 41,* 291-307.

Crisp, R. (1987). Persuasive advertising, autonomy, and the creation of desire. *Journal of Business Ethics, 6*(5), 413-418.

Croft, A. (2007). A lesson learnt: the rise and fall of Lariam and Halfan. *Journal of the Royal Society of Medicine, 100*(4): 170–174.

Churchland, P.S. (1986). *Neurophilosophy: Toward a unified science of the mind/brain*. Cambridge, MA: MIT Press.

Custers, R. and Aarts, H. (2005). Positive affect as implicit motivator: on the nonconscious operation of behavioral goals. *Journal of Personality and Social Psychology, 89*(2), 129-142.

Custers, R. and Aarts, H. (2007). In search of the nonconscious sources of goal pursuit: Accessibility and positive affective valence of the goal state. *Journal of Experimental Social Psychology, 43,* 312-318.

Custers, R. and Aarts, H. (2010). The unconscious will: how the pursuits of goals operates outside of conscious awareness. *Science, 329*, 47-50.

Dackis, C. and O'Brien, C. (2005). Neurobiology of addiction: treatment and public policy ramifications. *Nature Neuroscience, 8,* 1431-1436.

Damasio, A. (2005). Human behaviour: Brain trust. *Nature, 435*, 571-572.

Daniels, N. (1985). *Just health care*. Cambridge: Cambridge University Press.

Daniels, N. (1996). *Justice and justification: Reflective equilibrium in theory and practice*. Cambridge: Cambridge University Press.

Davidson, D. (1970). "How is weakness of the will possible?," in D. Davidson (1980), *Essays on actions and events*, Oxford: Clarendon Press, pp. 21-42.

Davies, J. (2011). Positive and negative models of suffering: An anthropology of our shifting cultural consciousness of emotional discontent. *Anthropology of Consciousness, 22*(2), 188-208.

DeGrazia, D. (2005). *Human identity and bioethics*. Cambridge, England: Cambridge University Press.

DeGrazia, D. (2014). Moral enhancement, freedom, and what we (should) value in moral behaviour. *Journal of Medical Ethics, 40*(6), 361-368.

Denys, D. (2011). Obsessionality and compulsivity: A phenomenology of obsessive-compulsive disorder. *Philosophy, Ethics, and Humanities in Medicine, 6*(3).

Dethlefs, D. (2007). *Chemically enhanced trust: Potential law enforcement and military applications for oxytocin.* Master's thesis, Naval Postgraduate School, Monterey, CA.

Dickens' *A Christmas Carol,*

Dijksterhuis, A. and Aarts, H. (2010). Goals, attention, and (un)consciousness. *Annual Review of Psychology, 61*, 467-490.

Dimaggio, G., and Semerari, A. (2001). Psychopathological narrative forms. *Journal of Constructivist Psychology*, 14, 1–23.

Dimaggio, G., and Semerari, A. (2004). Disorganized narratives: The psychological condition and its treatment. In L. E. Angus and J. McLeod (Eds.), *The handbook of narrative and psychotherapy: Practice, theory and research* (pp. 263–282). London: Sage.

Dimaggio, G., Salvatore, G., Azzara, C., & Catania, D. (2003). Rewriting selfnarratives: The therapeutic process. *Journal of Constructivist Psychology*, 16, 155–181.

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U. and Henrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological Psychiatry, 65*(9), 738-731.

Domes, G. Heinrichs, M., Michel, A., Berger, C. and Herpertz, S. (2007). Oxytocin improves "mind-reading" in humans. *Biological Psychiatry. 61*(6), 731-733.

Doucet, M. (May 15 2014). "In praise of akrasia." 2nd Annual Northwestern Society for Ethical Theory and Political Philosophy Conference. Evanston, Il.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy, 25*(3), 228-245.

Douglas, T. (2013). Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics, 27*(3), 160-168.

Douglas, T. (2014). Enhancing moral conformity and enhancing moral worth. *Neuroethics, 7*(1), 75-91.

Dow, J. (2013). "Ethics of Advertising." In *The International Encyclopedia of Ethics* (Ed. LaFollette, H.). Oxford: Blackwell Publishing.

Durkheim, E. (1915). *The elementary forms of the religious life* (Trans. Swain, J.). London: Allen & Unwin.

Dworkin, G. (1988). *The theory and practice of autonomy*, New York: Cambridge University Press.

Eddie. (Sep 2 2007). Re: "Death and destruction." Message posted to http://airborneparainf82.blogspot.com. Accessed 28 Oct 2014.

Elliott, R., Shapiro, D., Firth-Cozens, J., Stiles, W., Hardy, Llewelyn S., and Margison, F. (1994). Comprehensive process analysis of insight events in cognitive-behavioral and psychodynamic-interpersonal psychotherapies. *Journal of Counselling Psychology, 41*(4), 449-463.

Engelhardt, H. (1986). *The Foundations of bioethics*. Oxford: Oxford University Press.

Eyler, J. (2003). Smallpox in history: The birth, death, and impact of a dread disease. *Journal of Laboratory and Clinical Medicine, 142*(4), 216-220.

Farkas, C.-A. (2013). Potentially harmful side-effects: Medically unexplained symptoms, somatization, and the insufficient illness narrative for viewers of mystery diagnosis. *Journal of Medical Humanities*. doi: 10.1007/s10912-013-9234-8.

Feinberg, J. (1970). *Doing and deserving: Essays in the theory of responsibility*. Princeton, NJ: Princeton University Press.

Feldman, R., Weller, A., Zagoory-Sharon, O. and Levine, A. (2007). Evidence for neuroendocrinological foundation of human affiliation: Plasma oxytocin levels across pregnancy and the postpartum period predict mother-infant bonding. *Psychological Science, 18*(11), 965-970.

Fenton, E. (2010). The perils of failing to enhance: A response to Persson and Savulescu. *Journal of Medical Ethics*, *36*(3), 148-151.

Ferguson, J., Young, L., Hearn, E., Matzuk, M., Insel, T. and Winslow, J. (2000). Social amnesia in mice lacking the oxytocin gene. *Nature Genetics, 25*, 284-288.

Ferguson, J., Aldag, J., Insel, T. and Young, L. (2001). Oxytocin in the medial amygdala is essential for social recognition in the mouse. *Journal of Neuroscience, 21*(20), 8278-8285.

Feshbach, N. (1990). "Parental empathy and child adjustment/maladjustment." In *Empathy and its development* (Eds. Eisenberg, N. and Strayer, J.). Cambridge: Cambridge University Press.

Fetchenhauer, D. and Dunning, D. (2010). Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. *Psychological Science, 21*(2), 189-193.

Feynman, R.P. (2010). *Surely you're joking, Mr. Feynman!": Adventures of a curious character*. London: W. W. Norton & Company.

Figley, C. (1995). *Compassion fatigue: Coping with secondary traumatic stress disorder in those who treat the traumatized.* Philadelphia, PA: Brunner / Mazel, Inc.

Fischer, A. and Roseman, I. (2007). Beat them or ban them: The characteristics and social functions of anger and contempt. *Journal of Personality and Social Psychology, 93*(1), 103-115.

Fischer, J. and Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. New York, NY: Cambridge University Press.

Focquaert, F., and DeRidder, D. (2009). Direct intervention in the brain: ethical issues concerning personal identity. Journal of Ethics in Mental Health 4: 1–7.

Fox, D. (2005). Safety, efficacy, and authenticity: the gap between ethics and law in FDA decision making. *Michigan State Law Review, 1135.*

Fox, P. (1989). From senility to Alzheimer's disease: the rise of the Alzheimer's disease movement. *The Milbank Quarterly, 67*(1), 58-102.

Francis, A. (2012). Stigma in an era of medicalisation and anxious parenting: how proximity and culpability shape middle-class parents' experiences of disgrace. *Sociology of Health and Illness, 34*(6), 927-942.

Frank, A. (1995). *The wounded storyteller: body, illness, and ethics.* Chicago, IL: University of Chicago Press.

Frankfurt, H.G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*(3), 829-39.

Frankfurt, H.G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5-20.

Freidson, E. (1970). *Profession of medicine: A study of the sociology of applied knowledge*. Chicago: University of Chicago Press.

Friedman, R. (Jan 17 2011). "When self-knowledge is only the beginning." *The New York Times*. Accessed Sept 27 2014 at http://www.nytimes.com/2011/01/18/health/views/18mind.html

Fulford K.. (1989). *Moral theory and medical practice*. Cambridge: Cambridge University Press.

Gabe, J. (2004). "Medicalization." In *Key concepts in medical sociology* (Eds. Gabe, J., Bury, M. and Elston, M.). London: SAGE.

Gartner, C. and Partridge, B. (2012). “Addiction neuroscience and tobacco control.” In *Addiction Neuroethics* (Ed. Carter, A., Hall, W. and Illes, J.). London: Academic Press.

Gibbard, A. (1990). *Wise choices, apt feelings*. Cambridge, MA: Harvard University Press.

Giddens, A. (1991). *Modernity and self-identity: Self and society in the late modern age*. Cambridge: Polity Press.

Gilpin, R. (1986). “The richness of the tradition of political realism,” in Keohane, R. (ed.), *Neorealism and its critics*, New York: Columbia University Press, 1986.

Goodman, R. (2010). Cognitive enhancement, cheating, and accomplishment. *Kennedy Institute of Ethics Journal*, *20*(2), 145–160.

Graham, L. (2007). Countering the ADHD epidemic: a question of ethics? *Contemporary Issues in Early Childhood, 8*(2), 166-169.

Greenspan, S., Loughlin, G. and Black, R. (2001). “Credulity and gullibility in persons with mental retardation: A framework for future research.” In *International review of research in mental retardation, Vol. 24* (Ed. Glidden, M.). San Diego, Academic Press.

Guastella, A., Mitchell, P. and Mathews, F. (2008). Oxytocin enhances the encoding of positive social memories in humans. *Biological Psychiatry, 64*(3), 256-258.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814–834.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Vintage Books.

Hall, W., Gartner, C.. and Carter, A. (2008). The genetics of nicotine addiction liability: ethical and social policy implications. *Addiction, 103,* 350–359.

Halperin, E. (2008). Group-based hatred in intractable conflict in Israel. *The Journal of Conflict Resolution, 52*(5), 713-736.

Halperin, E. (2011). Emotional barriers to peace: Emotions and public opinion of Jewish Israelis about the peace process in the Middle East. *Peace and Conflict: Journal of Peace Psychology, 17*(1), 22-45. Harris, J. (1980). *Violence and responsibility*. London: Routledge and Kegan Paul.

Halpern, S. (1990). Medicalization as a professional process: Postwar trends in paediatrics. *Journal of Health and Social Behavior, 31,* 28-42.

Hare, R., (1952). *The language of morals*. Oxford: Clarendon Press.

Harman, G. (1975). Moral relativism defended. *Philosophical Review, 84*, 3-22.

Harris, J. (2001). One principle and three fallacies of disability studies. *Journal of Medical Ethics, 27,* 383-387.

Harris, J. (2007). *Enhancing evolution: The ethical case for making better people*. Princeton, NJ: Princeton University Press.

Harris, J. (2009). “Enhancements are a moral obligation.” In *Human Enhancement,* (Eds. Savulescu, J. and Bostrom, N.). Oxford: Oxford University Press.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25,* 102–11.

Harris, J. (2012). What it’s like to be good. *Cambridge Quarterly of Healthcare Ethics, 21*(3), 293-305.

Harris, J. (2013a). ‘Ethics is for bad guys!’ Putting the ‘moral’ into moral enhancement. *Bioethics, 27*(3), 169-173.

Harris, J. (2013b). Moral progress and moral enhancement. *Bioethics, 27*(5), 285-290.

Harris, J. (2014a). "… How narrow the strait!" *Cambridge Quarterly of Healthcare Ethics, 23*, 247-260.

Harris, J. (2014b). Taking liberties with free fall. *Journal of Medical Ethics, 40*(6), 371-374.

Hattenstone, S. and Allison, E. (Oct 18 2014). "'You don't ever get over it': meet the British soldiers living with post-traumatic stress disorder." *The Guardian*, Accessed Oct 28 2014 at http://www.theguardian.com /society /2014/oct/18/collateral-damage-ex-soldiers-living-with-ptsd.

Hayry, M. (2010). *Rationality and the genetic challenge: Making people better?* Cambridge: Cambridge University Press.

Hegel, GWF. (1837[1956]). *The Philosophy of history*, trans. Sibree, J. New York: Dover.

Heine, H. (1840 [2006]). *Ludwig Borne: A memorial* (Trans. Sammons, J.). Rochester, NY: Camden.

Henry, M., Fishman, J. and Youngner, S. Propanolol and the prevention of post-traumatic stress disorder: Is it wrong to erase the 'sting' of bad memories? *American Journal of Bioethics, 7*(9), 12-20.

Hermans, H. (2003). The construction and reconstruction of a dialogical self. *Journal of Constructivist Psychology, 16*, 89–130.

Hermans, H. and Dimaggio, G. (Eds.) (2004). *The dialogical self in psychotherapy*. New York, NY: Brunner-Routledge.

Hildt, E. (2006). Electrodes in the brain: Some anthropological and ethical aspects of deep brain stimulation. *International Review of Information Ethics, 5*, 33-39.

Hobbes, T. (1651[1962]). *Leviathan*. London: Oxford University Press.

Hobhouse, L. (1911). *Liberalism*. Oxford: Oxford University Press.

Holton, R. (1999). Intention and weakness of will. *The Journal of Philosophy, 96*(5), 241-262.

Hume, D. (1739 [1975]). *A treatise of human nature* (Ed. Sleby-Bigge, L.). Oxford: Clarendon Press, 1975.

Hume, D. (1751 [1983]). *An enquiry concerning the principles of morals* (Ed. Schneewind, J.). Indianapolis, IN: Hackett.

Hume, D. (1757 [1828]). "Dissertation on the passions", in *The philosophical works of David Hume*, vol. 4, Boston: Little, Brown and Company.

Hunt, D.P. (2000) Moral responsibility and unavoidable action. *Philosophical Studies*, 97(2), 195-227.

Huskamp, H. (2003). Managing psychotropic drug costs: Will formularies work? *Health Affairs, 22*(5), 84-96.

Hutcheson, F. (1725 [2004]). *An inquiry into the original of our ideas of beauty and virtue in two treatises.* Indianapolis, IN: Liberty Fund.

Hutcheson, F. (1728a [2002]). "Essay on the nature and conduct of the passions and affectations." In *Essay on the nature and conduct of the passions with illustrations on the moral sense* (Ed. Garrett, A.). Indianapolis, IN: Liberty Fund.

Hutcheson, F. (1728b [2002]). "Illustrations on the moral sense." In *Essay on the nature and conduct of the passions with illustrations on the moral sense* (Ed. Garrett, A.). Indianapolis, IN: Liberty Fund.

Hyman, S. (2014). I hope that we are not living in a post-fact world. *AJOB Neuroscience, 5*(3), 1-2.

Illich, I. (1975 [1976]). *Medical nemesis: The exploration of health*. New York, NY: Pantheon.

Inlander, C. (1998). Consumer health. *Social Policy, 28*(3), 40-42.

James, W. (1891). The principles of psychology, vol. 2. London: MacMillan and Co. Johansson, P., Hoglend, P, Ulberg, R., Amlo, S., Marble, A. … and Heyerdahl, O. (2010). The mediating role of insight for long-term improvements in psychodynamic therapy. *Journal of Consulting and Clinical Psychology, 78*(3), 438-448.

Jiménez-Ponce, F., Soto-Abraham, J.E., Ramírez-Tapia, Y., Velasco-Campos, F., Carrillo-Ruiz, J.D. and Gómez-Zenteno, P. (2011). Evaluation of bilateral cingulotomy and anterior capsulotomy for the treatment of aggressive behavior. *Cirugia y Cirujanos, 79*(2), 107-113.

Johanson, R., Newburn, M. and Macfarlane, A. (2002). Has the medicalization of childbirth gone too far? *British Medical Journal, 324*(7342), 892-895.

Jotterand, F. (2011). "Virtue engineering" and moral agency: Will post-humans still need virtues? *AJOB Neuroscience, 2*(4), 3-9.

Joyce, R. (2005). "Moral fictionalism." In *Fictionalism in metaphysics* (Ed. Kalderon, M.). Oxford: Oxford University Press.

Juengst, E. (1998). "What does enhancement mean?" In *Enhancing human traits: Ethical and social implications* (Ed. Parens, E.). Washington, DC: Georgetown University Press

Keats, J. (1816 [1847]). "Sleep and poetry." In *The poetical works of Coleridge, Shelley, and Keats*. Philadelphia: Crissy & Markley.

Key, J. (2007). *The deserter's tale: The story of an ordinary soldier who walked away from the war in Iraq*. Toronto: House of Anansi.

Keyes, D. (1966 [2004]). *Flowers for Algernon*. Orlando, FL: Houghton Mifflin Harcourt.

Kirmayer, L. (2000). Broken narratives: clinical encounters and the poetics of illness experience. In C. Mattingly and L. C. Garro (Eds.), *Narrative and the cultural construction of illness and healing* (pp. 153-180). Berkeley, CA: University of California Press.

Klerman, G. (1972). Psychotropic hedonism versus pharmacological Calvinism. *Hastings Center Report, 2*(4), 1-3.

Klimecki, O., and Singer, T. (2011). Empathic distress fatigue rather than compassion fatigue? Integrating findings from empathy research in psychology and social neuroscience. *Pathological altruism*, 368-383.

Kramer, P. D. (1994). *Listening to Prozac*. London, England: Fourth Estate Paperbacks.

Kroll-Smith, S. and Floyd, H. (1997). *Bodies in protest: Environmental illness and the struggle over medical knowledge.* New York, NY: New York University Press.

Kukathas, C. and Pettit, P. (1991) *Rawls: A theory of justice and its critics*. Cambridge and Stanford, CA: Polity Press and Stanford University Press.

La Boétie, É., (1576 [2008]). *The discourse of voluntary servitude*. Auburn, AL: Ludwig Von Mises Institute.

Laborde, C. (2010). Republicanism and global justice: A sketch. *European Journal of Political Theory, 9*(1), 48–69.

Larmore, C. (2004). 'Liberal and republican conceptions of freedom' in *Republicanism: History, theory, and practice*, Weinstock, D. and Nadeau, C. (Eds.), London: Frank Cass.

Lehrer, J. (Feb 25 2010). “Depression’s upside.” *New York Times Magazine*.

Lem, S. (1974 [1985]). *The cyberiad* (Trans. Kandel, M.). Orlando, FL: Harcourt.

Leshner, A. (1997). Addiction is a brain disease, and it matters. *Science, 278*(5335), 45-47.

Leukefeld, C., Gullotta, T. Gregrich, J. (Eds.). (2011). *Handbook of evidence-based substance abuse treatment in criminal justice settings*. New York, NY: Springer.

Link, B. and Phelan, J. (1995). Social conditions as fundamental causes of disease. *Journal of Health and Social Behavior, 35(special issue),* 80-94.

Locke, J. (1689[1996]). *An essay concerning human understanding,* Winkler, K.P. (Ed.). Indianapolis, IN: Hackett.

Lovett, F. (2001). Domination: A preliminary analysis. *Monist, 84*, 98-112.

Lovett, F. (2010). *A general theory of domination and justice*. Oxford: Oxford University Press.

Lovett, F. (2012). What counts as arbitrary power? *Journal of Political Power*, *5*, 137-152.

Maan, A. K. (2010). *Internarrative identity: Placing the self* (2nd ed.). Lanham, MD: University Press of America.

MacIntyre, A. (1981). *After virtue*. Notre Dame, IN: University of Notre Dame Press.

Mackenzie, C. and Poltera, J. (2010). Narrative Integration, Fragmented Selves, and Autonomy. *Hypatia 25*(1), 31-54.

Mackie, J. (1977). *Ethics: Inventing right and wrong.* Harmondsworth: Penguin.

MacManus, D., Dean, K., Jones, M., Rona, R., Greenberg, N., Hull, L. … and Fear, N. (2013). Violent offending by UK military personnel deployed to Iraq and Afghanistan: A data linkage cohort study. *The Lancet, 381*(9870), 907-917.

Maley, J.H., Alvernia, J.E., Valle, E.P., and Richardson. D. (2010). Deep brain stimulation of the orbitofrontal projections for the treatment of intermittent explosive disorder. *Neurosurgical Focus, 29*(2), E11.

Manuel, D., Lim, J., Tanuseputro, P., Anderson, G., Atler, D., Laupacis, A. and Mustard, C. (2006). Revisiting Rose: strategies for reducing coronary heart disease. *British Medical Journal, 332*(7542), 659-662.

Marazziti, D., Akiskal, H. S., Picchetti, M., Baroni, S., Massimetti, G., Albanese, F., & Dell’Osso, L. (2014). Dimorphic changes of some features of loving relationships during long-term use of antidepressants in depressed outpatients. *Journal of Affective Disorders*,166, 151-155.

Markell, P. (2008). The insufficiency of non-domination. *Political Theory, 36*(1), 9-36.

Marx, K. (1843 [1970]). *Critique of Hegel’s philosophy of right*. Oxford: Oxford University Press.

Maslow, A. (1969). *Psychology of Science: A reconnaissance*. Washington, DC: Regnery.

Mayberg, H., Lozano, A., Voon, V., McNeely, H., Seminowicz, D., Hamani, C. … and Kennedy, S. (2005). Deep brain stimulation for treatment resistant stimulation. *Neuron, 45*(5), 651-660.

McGrath, S. (2007). “Moral disagreement and moral expertise.” In *Oxford Studies in Metaethics Vol. 4* (Ed. Shafer-Landau, R.). Oxford: Oxford University Press.

McKinlay, J. (1982). Toward the proletarianization of physicians. In *Professionals as workers: Mental labor in advanced capitalism* (Ed. Derber, C.) Boston, MA: G.K. Hall.

McKinlay, J. and Marceau, L. (2002). The end of the golden age of doctoring. *International Journal of Health Services, 32,* 379–416.

McLellan, T., Lewis, D., O'Brien, C. and Kleber, H. (2000). Drug dependence, a chronic mental illness: Implications for treatment, insurance, and outcomes evaluation. *Journal of the American Medical Association, 284*(13), 1689-1695.

McMahan, J. (2002). *The ethics of killing: Problems at the margins of life*. Oxford, England: Oxford University Press.

Mele, A.R. (1982). *Irrationality: An essay on akrasia, self-deception, and self-control*. Oxford: Oxford University Press.

Merkel R., Boer G., Fegert J., Galert T., Hartmann D., Nuttin B., & Rosahl S. (2007). *Intervening in the brain: Changing psyche and society*. Berlin, Germany: Springer.

Merkl, P. (1986). *Political violence and terror: motifs and motivations*. Berkeley: University of California Press.

Mikolajczak, M., Pinon, N., Lane, A., de Timary, P. and Luminet, O. (2010a). Oxytocin not only increases trust when money is at stake, but also when confidential information is in the balance. *Biological Psychology, 85*(1), 182-184.

Mikolajczak, M., Gross, J., Lane, A., Corneille, O., de Timary, P. and Luminet, O. (2010b). Oxytocin makes people trusting, not gullible. *Psychological Science, 21*(8), 1072-1074.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology, 67*(4), 371-378.

Mill, J. (1859 [2003]). *On liberty*. New Haven, CT: Yale University Press.

Moran, D. (2000). *Introduction to phenomenology*. New York, NY: Routledge.

Morris, D. B. (2012). Narrative and pain: towards an integrative model. In R. J. Moore (Ed.), *Handbook of Pain and Palliative Care* (pp. 733-751). New York, NY: Springer.

Moses, (2012). *Child parent relationship therapy for parents of children with disruptive behaviour*. Unpublished dissertation, Kalamazoo: Western Michigan University.

Mullen, E. and Skitka, L. (2006). Exploring the psychological underpinnings of the moral mandate effect: Motivated reasoning, group differentiation, or anger? *Journal of Personality and Social Psychology, 90,* 629–643.

Muller, S., and Christen, M. (2011). Deep brain stimulation in Parkinsonian patients: Ethical evaluation of stimulation-induced personality alterations. *AJOB Neuroscience,* 2(1), 3–13.

Muraven, M., Tice, D. and Baumeister, R (1998). Self-control as limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology, 74,* 774-789.

Muzak, J. (2007). "They say the disease is responsible": Social identity and the disease concept of drug addiction. In V. Raoul, C. Canam, A. D. Henderson & C. Paterson (Eds.), *Unfitting stories: narrative approaches to disease, disability, and trauma* (pp. 255-264). Waterloo, Canada: Wilfred Laurier Press.

Nadelson, T. (2005). *Trained to kill: Soldiers at war*. Baltimore, John Hopkins University Press.

Nagal, T. (1979). *Mortal questions*. Cambridge: Cambridge University Press.

National Institute on Drug Abuse. (2012). *Principles of drug addiction treatment: A research-based guide*, 3rd ed. National Institute of Health.

Naylor, M.B. (1984). Frankfurt on the principle of alternate possibilities. *Philosophical Studies, 46*, 249-258.

Nietzsche, F.(1883[1976]). "Thus Spoke Zarathustra" in *The Portable Nietzsche*, Kauffman, W (trans.). New York: Penguin.

Nordmann, A. (2007). If and then: A critique of speculative nanoethics. *Nanoethics*, 1(1), 31–46.

Norman, D. and Shallice, T. (1986). "Attention to action: Willed and automatic control of behavior." In *Consciousness and self-regulation: Advances in research and theory* (Eds. Davidson, R., Schwartz, R. and Shapiro, D.). New York: Plenum Press.

Novalis, (1997). *Philosophical writings* (trans. and ed. Stoljar, M.). New York, NY: State University of New York Press.

Nozick, R. (1974). *Anarchy, state, and utopia*. New York, NY: Basic Books.

O'Guinn, T. C., and Faber, R. J. (1989). Compulsive buying: A phenomenological exploration. *Journal of consumer research*, 16(2), 147-157.

Olsen, J. (2006). Depression, SSRIs, and the supposed obligation to suffer mentally. *Kennedy Institute of Ethics Journal, 16*(3), 283–303.

Olson Eric, (2003). "Personal identity". In *The Blackwell guide to philosophy of mind* (Eds. Stich, S. and Warfield, T.). Malden, MA: Blackwell.

Opbroek, A., Delgado, P., Laukes, C., McGahuey, C., Katsanis, J., Moreno, F. and Manber, R. (2002). Emotional blunting associated with SSRI-induced sexual dysfunction. Do SSRIs inhibit emotional responses. *The International Journal of Psychopharmacology, 5*(2), 147-151.

Osborne, E. (1996). "Volcano's gift." In *Eruption: Montserrat versus volcano* (Ed. Fergus, H.). Montserrat: University of the West Indies School of Continuing Studies.

Pacholczyk, A. (2011a). DBS makes you feel Good! – Why some of the ethical objections to the use of DBS for neuropsychiatric disorders and enhancement are not convincing. *Frontiers in Integrative Neuroscience, 5*, 14.

Pacholczyk, A. (2011b). Moral enhancement: What is it and do we want it? *Law, Innovation and Technology, 3*(2), 251-277.

Pacholczyk, A. (2015). "Ethical objections to deep brain stimulation for neuropsychiatric disorders and enhancement: A critical review." In *Handbook of neuroethics* (Eds. Clausen, J. and Levy, N.). London: Springer.

Pacholczyk, A., Harris, J. (2010). *Dignity and Enhancement*, unpublished paper.

Parens, E. (2011). On good and bad forms of medicalization. *Bioethics, 27*(1), 28-35.

Parfit, D. (1984). *Reasons and persons*. Oxford, England: Clarendon.

Parfit, D. (1988). *What we together do*. Unpublished manuscript, Oxford.

Parsons, T. (1951). Illness and the role of the physician: A sociological perspective. *American Journal of Orthopsychiatry, 21*, 452–460.

Pawluch, D. (1983). Transitions in paediatrics: A segmental analysis. *Social Problems, 30*, 449-465.

Pears, D. (1998). *Motivated irrationality*. South Bend, IN: St. Augustine's Press.

Pedersen, C. (2004). Biological aspects of social bonding and the roots of human violence. *Annals of the New York Academy of Sciences, 1036,* 106-127.

Pedersen, C., Ascher, J., Monroe, Y. and Prange, A. (1982). Oxytocin induces maternal behaviour in virgin female rats. *Science, 216*(4546), 648-650.

Pennebaker, J. (1997). *Opening up*. New York: Guilford.

Persson, I. and Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral characters of humanity. *Journal of Applied Philosophy, 25*(3), 162-167.

Pescosolido, B. (2006). Professional dominance and the limits of erosion. *Society, 43*(6), 21–29.

Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government*. Oxford: Clarendon Press.

Pettit, P. (2001). *A theory of freedom: From the psychology to the politics of agency*. Cambridge: Polity Press.

Pettit, P. (2005). "The domination complaint." In: Williams, M.S., (Ed.), *Nomos 46: political exclusion and domination.* New York: New York University Press.

Pettit, P. (2010). A Republican law of peoples. *European Journal of Political Theory, 9*(1), 70–94.

Pettit, P. (2012). *On the people's terms: A republican theory and model of democracy*. Cambridge: Cambridge University Press.

Phillips, J. (2003). Psychopathology and the Narrative Self. *Philosophy, Psychiatry and Psychology 10*(4), 313-328.

Phillips, M. (1994). The inconclusive ethical case against manipulative advertising. *Business & Professional Ethics Journal, 13*(4), 31-64.

Pierre, J. (2013). "Overdiagnosis, underdiagnosis, synthesis: A dialectic for psychiatry and the DSM." In *Making the DSM-5: Concepts and controversies* (Eds. Paris, J. and Phillips, J.). New York, NY: Springer.

Plato. (1996). *Protagoras* (Trans. Taylor, C.). Oxford: Oxford University Press.

President's Council on Bioethics. (2003). *Beyond therapy: biotechnology and the pursuit of happiness*. Washington, DC.

Prince, J. Cole, V. Goodwin, G.M. (2009). Emotional side-effects of selective serotonin reuptake inhibitors: qualitative study. *British Journal of Psychiatry, 195*(3), 211-117.

Quong, J., (2011). *Liberalism without perfection*. Oxford: Oxford University Press.

Raz, J. (1986). *The Morality of freedom*. Oxford: Oxford University Press.

Read J, Haslam N, Sayce L, and Davies E. (2006). Prejudice and schizophrenia: a review of the 'mental illness is an illness like any other' approach. *Acta Psychiatrica Scandinavica, 114*(5), 303–318.

Read, J. Catwright, C. and Gibson, K. (2014). Adverse emotional and interpersonal effects reported by 1829 New Zealanders while taking antidepressants. *Psychiatry Research, 216*(1), 67-73.

Ricoeur, P. (1984). *Time and narrative,* vol 1. Chicago, IL: University of Chicago Press.

Ricoeur, P. (1985). *Time and narrative,* vol 2. Chicago, IL: University of Chicago Press.

Ricoeur, P. (1988). *Time and narrative,* vol 3. Chicago, IL: University of Chicago Press.

Ricoeur, P. (1992). *Oneself as another*. Chicago, IL: University of Chicago Press.

Rifkin, J. (2010a). *The empathic civilization: The race to global consciousness in a world in crisis.* Cambridge: Polity Press.

Rifkin, J. (2010b). *The empathic civilization*. Address to the British Royal Society for the Arts.

Rigstadt, M. (2011). Republicanism and geopolitical domination. *Journal of Political Power, 4*(2), 279-300.

Ritchie, E., Block, J. and Nevin, R. (2013). Psychiatric side effects of mefloquine: Applications to forensic psychiatry. *Journal of the American Academy of Psychiatry and the Law, 41*(2), 224-235.

Robben, A. (2010). Chaos, mimesis and dehumanisation in Iraq: American counterinsurgency in the global War on Terror. *Social Anthropology, 18*(2), 138-154.

Robben, A. (2012). From dirty war to genocide: Argentina's resistance to national reconciliation. *Memory Studies, 5*(3), 305-313.

Romeo, D. (Jun 25 2012). "The victims of Montserrat's volcano tragedy deserve an investigation." *The Guardian*. Accessed Aug 24 2014 at http://www.theguardian.com/commentisfree/2012/jun/25/ victims-montserrat-volcano-tragedy-investigation.

Rorty, A. (1980). Where does the akratic break take place? *Australian Journal of Philosophy, 58*(4), 333-346.

Rose, G. (1981). Strategy of prevention: Lessons from cardiovascular disease. *British Medical Journal, 282*(6279), 1847-1851.

Rose, G. (2001). Sick individuals and sick populations. *International Journal of Epidemiology, 30*(3), 427-432.

Rosenblatt, R. (1989). The perinatal paradox: Doing more and accomplishing less. *Health Affairs, 8*(3), 158-168.

Rosenstein, P. (1995). Parental levels of empathy as related to risk assessment in child protective services. *Child Abuse & Neglect, 19*(11), 1349-1360.

Rousseau, J.J. (1767). *The Miscellaneous Works of Mr. J. J. Rousseau, vol.3*. Charleston, SC: Nabu Press.

Rowe, W.L. (1987). Two concepts of freedom. *Proceedings and Addresses of the American Philosophical Association, 61*, 43-64.

Russell, B. (2010). *Why men fight*. Oxon: Routledge Classics.

Sabin, J. and Daniels, N. (1994). Determining "medical necessity" in mental health practice. *Hastings Center Report, 24*(6), 5-13.

Sagan, C. (1996). *The demon-haunted world: Science as a candle in the dark*. New York, NY: Random House.

Sandler, J., Dare, C. and Holder, A. (1973). *The patient and the analyst: The basis of the psychodynamic process.* London: George Allen & Unwin.

Sansone, R.A. and Sansone, L.A. (2010). SSRI-induced indifference. *Psychiatry, 7*(10), 14-18.

Savulescu, J. and Persson, I. (2012a). Moral enhancement, freedom and the God Machine. *Monist, 95*(3), 399-421.

Savulescu, J. and Persson, I. (2012b). *Unfit for the future: The need for moral enhancement*. Oxford: Oxford University Press.

Schechtman, M. (1996). *The constitution of selves*. Ithaca, NY: Cornell University Press.

Schechtman, M. (2009). Getting our stories straight: Self-narrative and personal identity. In D.J.H. Mathews, H. Bok, & P.V. Rabins (Eds.), *Personal identity* *and fractured selves* (pp. 65-92). Baltimore, MD: Johns Hopkins University Press.

Schechtman, M. (2010). Philosophical reflections on narrative and deep brain stimulation. *Journal of Clinical Ethics, 21*(2), 133-139.

Schermer, M. (2007). The dynamics of the treatment-enhancement distinction: ADHD as a case study. *Philosophica 79*, 25-37.

Schermer, M. (2008). Enhancements, easy shortcuts, and the richness of human activities. *Bioethics, 22*, 355–363.

Schermer, M. and Bolt, I. (2011). "What's in a name? ADHD and the gray area between treatment and enhancement." In *Enhancing human capacities* (Eds. Savulescu, J., ter Meulen, R. and Kahane, G.). Chichester: Blackwell Publishing.

Schupbach M, Gargiulo M, Welter ML, Mallet, L., Behar, C., Houeto, J. L., Maltete, D., Mesnage, V., & Agid, Y. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology 66*(12), 1811-1816.

Scott, R., Ruef, M., Mendel, P. and Caronna, C. (2000). *Institutional change and healthcare organizations: From professional dominance to managed care*. Chicago, IL: University of Chicago Press.

Scott, W. (1990). PTSD in DSM-III: A case in the politics of diagnosis and disease. *Social Problems, 37*(3), 294-310.

Selsam, H. (1963). *Reader in Marxist philosophy*. New York, NY: International Publishers.

Seneca. (1995). *Moral and Political Essays,* (Eds. Cooper, J. and Procope, J.). Cambridge: Cambridge University Press.

Shaftsesbury, A. (1699 [2000]). "An inquiry concerning virtue or merit." In *Characteristics of men, manners, opinions, times,* (Ed. Klein, L.) Cambridge: Cambridge University Press.

Shamay-Tsoory, S., Fischer, M., Dvash, J., Harari, H., Perach-Bloom, N. and Levkovitz, Y. (2009). *Biological Psychiatry, 66*(9), 864-870.

Shapiro, I. (2012). On non-domination. *University of Toronto Law Journal, 62*(3), 293-335.

Shenk, J. (2005). *Lincoln's melancholy: How depression challenged a president and fuelled his greatness*. Boston, MA: Houghton Mifflin.

Sidgwick, H. (1907 [1981]). *The methods of ethics*. Indianapolis: Hackett.

Skinner, Q. (1998). "Liberty before liberalism," Cambridge: Cambridge University Press. 2008.

Skinner, Q. (2008). "Freedom as the absence of arbitrary power," in *Republicanism and Political Theory,* Laborde, C. and Maynor, J. (eds.). Malden, MA: Blackwell Publishing.

Smith, A. (1759 [2009]). *The theory of moral sentiments* (Ed. Hanley, R.). London: Penguin.

Smith, M. (2011). "Empathy, expansionism, and the extended mind." In *Empathy: Philosophical and psychological perspectives* (Eds. Coplan, A. and Goldie, P.). Oxford: Oxford University Press.

Sontag, S. (1978). *Illness as metaphor*. New York, NY: Farrar, Straus and Giroux.

Sparrow, R. (2014). Better living through chemistry? A reply to Savulescu and Persson on 'Moral Enhancement.' *Journal of Applied Philosophy, 31*(1), 23-32.

Speak, D. (2002). Fanning the flickers of freedom. *American Philosophical Quarterly, 39*(1), 91-105.

Specker, J., Focquaert, F., Raus, K., Sigrid, S. and Schermer, M. (2014). The ethical desirability of moral bioenhancement: a review of reasons. *BMC Medical Ethics*, 15, 67

Starr, P. (1982). *The social transformation of American medicine: The rise of a sovereign profession and the making of a vast industry*. New York, NY: Basic Books.

Stevenson, C. (1937). The emotive meaning of ethical terms. *Mind, 46*(181), 14-31.

Strawson, G. (1999). The Self and the SESMET. *Journal of Consciousness Studies 4*(5-6), 405-428.

Strawson, G. (2004). Against narrativity. *Ratio 17*(4): 428-452.

Strawson, G. (2007). Episodic ethics. In D. Hutto (Ed.), *Narrative and understanding persons*. Cambridge, England: Cambridge University Press.

Summerfield, D. and Veale, D. (2008). Proposals for massive expansion of psychological therapies would be counterproductive across society. *British Journal of Psychiatry, 192*, 326-330.

Synofzik, M., and Schlaepfer, T. E. (2008). Stimulating personality: Ethical criteria for deep brain stimulation in psychiatric patients and for enhancement purposes. *Biotechnology Journal, 3*, 1511-1520.

Szasz, T. (1963 [1989]). *Law, liberty, and psychiatry: An inquiry into the social uses of mental health practices.* Syracuse, NY: Syracuse University Press.

Szasz, T. (1970). *The manufacture of madness: A comparative study of the inquisition and the mental health movement.* New York, NY: Harper & Row.

Tang, T.Z. and DeRubeis, R.J. (1999). Sudden gains and critical sessions in cognitive–behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 67,* 894–904.

*"*The heavy burden of post-traumatic stress disorder*." The New York Times. (Jun 20 2014). Accessed Oct 28 2014 at http://www.nytimes.com/2014/06/21/opinion/the-heavy-burden-of-post-traumatic-stress-disorder.html*

The matrix revolutions. (2003). Dir. Wachowski, L. and Wachowski, A. Village Roadshow Pictures and Silver Pictures.

Theodoridou, A., Rowe, A., Penton-Voak, I. and Rogers, P. (2009). Oxytocin and social perception: Oxytocin increases perceived facial trustworthiness and attractiveness. *Hormones and Behavior, 56*(1), 128-132.

Turner, D. and Sahakian, B. (2006). Ethical questions in functional neuroimaging and cognitive enhancement. *Poesis and Praxis, 4,* 81-94.

Twain, M. (1884 [1994]). *The Adventures of Huckleberry Finn.* New York: William Morrow.

Unger, P. (1990). *Identity, consciousness, and value*. Oxford, England: Oxford University Press.

Unkelbach, C. Guastella, A. and Forgas, J. (2008). Oxytocin selectively facilitates recognition of positive sex and relationship words. *Psychological Science, 19*(11), 1092-1094.

Van Inwagen, P. (1978). Ability and responsibility. *Philosophical Review, 87,* 201-224.

Verweij, M. (1999). Medicalization as a moral problem for preventative medicine. *Bioethics, 13*(2), 89-113.

Vice, S. (2003). Literature and the Narrative Self. *Philosophy 78*(1), 93-108.

Volkow, N. and Li, T. (2004). Drug addiction: The neurobiology of behaviour gone awry. *Nature Reviews Neuroscience, 5*(12), 963-970.

Vrecko, S. (2013). Just how cognitive is "cognitive enhancement"? On the significance of emotions in university students' experiences with study drugs. *AJOB Neuroscience, 4*(1), 4-12.

Waever, O. (1994). "Resisting the temptation of post foreign policy analysis," in *European foreign policy: The EC and changing perspectives in Europe*, Carlsnaes, W. and Smith, S. (eds.), London: Sage.

Watson, G. (2004). 'Scepticism about weakness of will.' In *Agency and Answerability*. Oxford: Clarendon Press.

Weber, M. (1968). *Economy and society,* Roth, G. and Wittich, C. (eds.) Berkley, CA: University of California Press.

Wedgwood, R. (2010). "The moral evil demons." In *Disagreement* (Eds. Feldman, R. and Warfield, T.). Oxford: Clarendon Press.

Wendt, A. (1987). The agent-structure problem in international relations theory. *International Organization, 41*(3), 335-370.

Wendt, A. (1999). *Social theory of international politics*. Cambridge: Cambridge University Press.

WHO cooperative trial committee of principal investigators. (1978). A co-operative trial in the primary prevention of ischemic heart disease using clofibrate, report from the committee of principal investigators. *British Heart Journal, 40*(10), 1069–1118.

Williams, B. (1981). *Moral luck.* Cambridge: Cambridge University Press.

Winslow, J. and Insel, T. (2002). The social deficits of the oxytocin knockout mouse. *Neuropeptides, 36*(2-3), 221-229.

Witt, K., Kuhn, J., Timmermann, L., Zurowski, M., & Woopen, C. (2011). Deep brain stimulation and the search for identity. *Neuroethics*. doi:10.1007/s12152-011-9100-1.

Wolf, S. (1990). *Freedom within reason*. Oxford: Oxford University Press.

Wolpe, P. (2002). Treatment, enhancement, and the ethics of neurotherapeutics. *Brain and Cognition, 50*(3), 387-395.

Wong, D. (1984). *Moral relativity*. Berkeley, CA: University of California Press.

Wordsworth, W. (1815 [2010]). "Surprized by joy—impatient as the Wind." In *William Wordsworth,* (Ed. Gill, S.). Oxford: Oxford University Press.

Zak, P., Stanton, A. and Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS One, 2*(11), e1128.