MORAL ENHANCEMENT AND PERSONAL AUTONOMY

Lucas Venter

8703294P

A Thesis Submitted to the Faculty of Humanities,

University of the Witwatersrand, Johannesburg

in Partial Fulfilment of the Requirements for the Degree

Master of Arts

Johannesburg, March 2013

Abstract

In this thesis, I examine the extent to which moral enhancement, the biomedical alteration of an individual's disposition to act according to good or bad motives, will influence his capacity for self-governance. Following a discussion of the salient features of moral enhancement, a plausible list of conditions against which to measure the compatibility of moral enhancement with personal autonomy is expounded. The core elements of moral enhancement are weighed against these conditions in order to establish the ways in which these core elements are compatible with the conditions of personal autonomy.

I argue that moral enhancement need not lead to a diminishment of personal autonomy, provided it serves merely as a mechanism to help an agent overcome the deterministic limitations that prevent him from bringing his lower-order desires into conformity with the higher-order desires that he has arrived at through independent, thoughtful deliberation.

Declaration

I declare that this thesis is my own unaided work. It is submitted for the degree of Master of Arts in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any other degree or examination in any other university.

Lucas Venter

14 March 2013

Acknowledgments

I would like to thank my wife and daughter for their support during this extended journey, as well as my supervisor, Dr Dylan Futter, for his scholarly rigour, abundant patience and unfailing guidance.

Table of Contents

Chapter 1

Introduction

In this thesis I will try to assess whether we would make an individual less able to govern his life if we were to alter him through biomedical means so that he would be less likely to commit morally bad acts.

Much in the study of ethics, of what is right and what is good, has pursued the practical goal, stated or unstated, of improving the morality of our interactions with each other. The central preoccupation of ethics has more often, however, been to establish what, if anything, is good and right, or why that which we commonly hold to be good or right is so, rather than what, if anything, can be done to increase the prevalence of good and right in the world. Underlying this quest has been an understanding, often left implicit, that moral improvement would come in the wake of greater moral understanding.

Such philosophical enquiry as has been focused on the practical improvement of morality has often sought to achieve such improvement through distributive justice, diminishing the need for crime through more equal access to goods and to educational systems that would instil virtuous behaviour.

Social improvements in European nations did not, however, bring the necessary moral improvement to the extent hoped for, causing Bertrand Russell to write in 1924:

> If men were rational in their conduct, that is to say, if they acted in the way most likely to bring about the ends that they deliberately desire, intelligence would be enough to make the world almost a paradise. In the main, what is in the long run advantageous

> to one man is also advantageous to another. But men are actuated by passions which distort their view; feeling an impulse to injure others, they persuade themselves that it is to their interest to do so. They will not, therefore, act in the way that is in fact to their own interest unless they are actuated by generous impulses which make them indifferent to their own interest. This is why the heart is as important as the head. By the "heart" I mean, for the moment, the sum-total of kindly impulses. Where they exist, science helps them to be effective; where they are absent, science only makes men more cleverly diabolic. [..]
>
> Our unconscious is more malevolent than it pays us to be; therefore the people who do most completely what is in fact to their interest are those who, on moral grounds, do what they believe to be against their interest.
>
> For this reason, it is of the greatest importance to inquire whether any method of strengthening kindly impulses exists. I have no doubt that their strength or weakness depends upon discoverable physiological causes. (1924:59-61)

In the subsequent decades, rapid advances in our understanding of the workings of the human body have indeed begun to turn the prospect of the biomedical alteration of human beings from the realm of speculation to reality. Where such intervention seeks to improve the natural capacities of a human being, for instance by improving her physical strength or mental acuity, it is generally denoted by the term "enhancement," although opinions differ as to where therapy ends and enhancement begins, and doubt has been cast on the general validity of this distinction.

Views on the moral value of enhancement have been greatly polarised, with a spectrum of views ranging from the outright moral and political condemnation of most, if not all, forms of biomedical enhancement on the one "bioconservative" extreme, to the wholesale call for the enhancement of humans such as made by the transhumanist movement on the other. Lying between these two extremes, a growing number of nuanced positions have sought to bring more critical rigour to the debate, insisting that neither generic condemnation nor approbation of biomedical enhancement is reasonable and that each area of enhancement and the putative means by which it could be attained should be placed under the magnifying glass.

One such area worthy of potential enhancement, albeit one that has emerged only recently, is human morality, the disposition of a person to act according to good or bad motives. Despite our immense intellectual and technological progress, we are—the argument for such "moral enhancement" goes—trapped in the morality of our hunter-gather ancestors, and moral improvement by traditional means such as education cannot take us beyond certain biological limitations that have arisen as a result of natural selection. Such limitations might include the seeming inability of humans to consider the consequences of their actions for people outside their immediate circle of family and friends or for future generations. Thus we are by nature more inclined to do harm to people we do not see in front of us, whether by committing theft over the Internet or firing shots at another human through a closed door. The improvement contemplated by proponents of moral enhancement, would not, however, be moral perfection, not the "moral sainthood" which Susan Wolf (1982)—and most people—would find deeply problematic. It merely means that, *ceteris paribus*, humans morally enhanced would be less inclined to act on morally bad motives.

The various attempts that have thus far been made to outline possible means of moral enhancement are united by one common feature: such an intervention may not interfere with the ability of an agent to act up her desires or execute her decisions, but should rather seek to influence those desires so that the enhanced human herself would will her actions to be moral. The increased moral behaviour elicited by this enhancement would thus—at least for as long as it was effective—be behaviour that the subject herself desires and with which she identifies.

On the face of it, a human being morally enhanced in this manner would therefore endorse her own actions in terms of principles with which she identifies. Those principles would, however, not be entirely of her own making, but would be shaped by certain dispositions externally imposed on her by the authors of the enhancement insofar as this

intervention would be effective in terms of the goals that they set out to achieve. This would appear, on the face of things, to constitute a limitation on her capacity to act autonomously. It is therefore not surprising that not only opponents but also some proponents of moral enhancement have highlighted this conflict with personal autonomy as problematic.

It is widely held that a human can only be held morally responsible for actions that he has fully endorsed. By this account, a human can only really be called a "person" insofar as he can be said to be in control of his own life and the decisions that he makes. According to this view, for morality to be truly "moral," it has to come from within. Thus, if it were to be found that a morally enhanced human is not autonomous in respect of her moral decision-making, one would not, according to such views as make personal autonomy a necessary condition for moral agency, be able to consider a human thus enhanced to be a moral agent any longer, in which case the moral enhancement project would be a self-defeating endeavour.

In this paper, I will focus on the implications that moral enhancement would have for the capacity of an agent to govern himself and take ownership of his life. In order to measure the potential influence of moral enhancement on personal autonomy, it is necessary to have a clear understanding of these two notions, an endeavour which is complicated by the fact that moral enhancement is still a very young, mostly theoretical concept and personal autonomy is a very contentious notion in respect of which many conceptions (some diametrically opposed) exist.

I therefore commence the paper with an attempt to reach, at most, a minimum conception of moral enhancement on which to base my subsequent argument. This is followed by a brief discussion of the empirical assumptions on which its feasibility as an actual form of biotechnology is based.

In Chapter Three I discuss personal autonomy and attempt to arrive, despite the contentious nature of this concept, at a plausible list of conditions for an individual to be deemed autonomous. These conditions I then employ in Chapter Four to ascertain the extent to which certain core elements of moral enhancement would influence the autonomy of an individual thus enhanced.

This comparison leads me to conclude that moral enhancement need not lead to a diminishment of personal autonomy, provided it serves merely as a mechanism to help an agent overcome the deterministic limitations that prevent him from bringing his lower-order desires into conformity with the higher-order desires that he has arrived at through independent, thoughtful deliberation.

Chapter 2

Moral Enhancement

2.1 Introduction

The purpose of this chapter is to present a more detailed overview of moral enhancement, what it is or purports to be, what it seeks to achieve and how it may do so. As there is already some disagreement about the notion, it will be necessary in the first half of this chapter to arrive at a "minimum conception" of moral enhancement. The purpose of this minimum conception will be to guide my subsequent argument and delineate its conceptual limits.

This conception will have to be of the "minimum" variety, as there are still too many empirical lacunae in our understanding of the subject and too much philosophical disagreement, especially in metaethics and philosophy of the mind, to make a more comprehensive conception of moral enhancement possible.

The minimum conception of moral enhancement which I will attempt to formulate, will consequently be limited to those features of moral enhancement that are required for a meaningful discussion of the topic of this paper, namely the compatibility of moral enhancement with personal autonomy, while also side-stepping, insofar as this is possible, areas of philosophical disagreement that could impede the formulation of an exhaustive definition.

I will then, in the second half of this chapter, present an overview of the general claims on which the theoretical notion of moral enhancement are based and the current state of empirical evidence for these claims. This overview will, as a matter of necessity, require me to touch in a cursory fashion on the neural correlates of morality, given the very real risk that any alterations to this domain will, as a result of the complexity of the brain, result in consequences that extend beyond the focus of the intervention, including consequences that may have some effect on an individual's autonomy or capacity for autonomy.

2.2 A minimum conception of moral enhancement

Defining the term "moral enhancement" is not without its complexities, which largely stem from the different meanings that can be imputed to the two words that make up the term: "moral" and "enhancement". I will therefore begin this discussion with an analysis of these two concepts seen in separation, and then look at the term as a whole.

2.2.1 "Moral"

When I refer to a specific kind of enhancement as a "moral" enhancement, this can mean one of two things. I can either be referring to the moral value of this particular enhancement or to a specific feature or features of the agent that this kind of enhancement will seek to alter. As I have indicated in Chapter One, the "moral enhancement" which forms the subject matter of this thesis is a specific kind of biotechnical enhancement that targets the morality of an agent. At the same time, my subject is concerned with the possibility that enhancements of this kind could, by somehow endangering the

agent's ability to govern himself, be bad for the agent. This means that "moral enhancement" in both these meanings is relevant to my enterprise.

In the next two sections I therefore discuss the "moral" in "moral enhancement" in both these senses, as referring to the moral value of the enhancement on the one hand, and the domain earmarked for alteration on the other.

The moral value of the enhancement

"Moral" enhancement as referring to the moral value of the enhancement can be subdivided further, either referring to (i) the morality of the means by which the intervention is to be attained, (ii) the morality of the goal towards which the enhancement strives or (iii) the morality of the actual consequences of the enhancement.

Concerning the morality of the means of enhancement, we may say, for instance, that an enhancement is not a "moral" enhancement if it is brought about in a subject through force or deception or if it debases or causes suffering to the agent undergoing the intervention.

Even if the means by which the enhancement is brought about is moral, the goal of the enhancement may be deemed immoral. An enhancement may be moral in respect of the goal it pursues if that goal is itself a moral one; thus an enhancement that is aimed at reducing the prevalence of a debilitating disease in the world may be considered a "moral" enhancement.

Lastly, an enhancement may be deemed moral if the consequences of the enhancement are deemed to be good. An enhancement may be achieved through moral means and pursue a moral goal, but if it has bad consequences it will not be considered moral. Thus, if a cognitive enhancement makes the subject more intelligent, thereby increasing

his prospects for a career with affluence, and if such enhancement is only affordable to a certain section of the population, this would lead to a bad consequence – a decrease in equality in the world. Such an enhancement would not be a "moral" enhancement. More pertinent to the subject matter of this paper, if it is good for an individual to be in control of his life and an enhancement lessens this control (the subject of Chapter Three), such an enhancement cannot be deemed moral.

Enhancement of the subject's morality

In the second sense of "moral enhancement," "moral" refers to an intervention that is aimed at altering the morality of the human subject, for instance by making an agent less inclined to commit morally bad acts such as a deed of deception or violence.

We can distinguish at least two different kinds of claims that could be made about the nature of the alteration brought about in the agent by such a kind of enhancement.

The first claim would focus on the normative value of the *actions* brought about by the alteration, e.g.

C.1 The enhancement caused the agent to behave badly less often.

The second claim would largely refer to the value of the *character* of the agent post-enhancement, e.g.

C.2 The enhancement caused the agent to become a better person.

Several critics of moral enhancement, such as Fabrice Jotterand (2011) and John Harris (2011), have cast doubt on the second claim, as certain conceptions of morality may set requirements for moral character that would automatically disqualify most, or all, forms of moral enhancement, for example the requirement that morality stem from

moral virtue that has been acquired "the hard way" as a result of a certain kind of experience in the course of a lifetime.

This is an important question in the debate around moral enhancement, and it is not unrelated to the central question of this paper, for if moral enhancement changes the propensity of an individual to engage in immoral acts without changing the individual's character accordingly, would this not lead to an increase in the divide between who the person is (or considers himself to be) and what he actually does?

In order to arrive at our minimum conception of moral enhancement, however, it is not necessary to settle the question at this stage of our discussion. If a form of moral enhancement leads to the second kind of alteration, namely an alteration of the character of the agent, thereby causing him to become a better person, this should also entail the first, action-oriented, claim about moral enhancement, namely that the agent thus enhanced will behave badly less often, as it seems reasonable to expect that the better the moral character of an individual, the less will be the likelihood that his actions will be morally bad. A certain value-laden expectation regarding the actions of the post-enhanced subject is thus applicable to all forms of enhancement that seek to alter the moral character of the subject.

Whether the enhanced individual's character concords with his actions post intervention relates rather to the moral value of the enhancement itself as defined in the first sense of "moral" enhancement, and to incorporate such a requirement into our minimum definition of moral enhancement at this juncture may unnecessarily exclude certain kinds of moral enhancement from our scope. Moreover, by requiring change to moral character as a necessary goal of moral enhancement at this stage will embed a

certain presupposition about the possible effects of moral enhancement in my argument and therefore contains within it the risk of begging the question.

I will therefore return to this question, but for the time consider it prudent to keep the minimum conception of moral enhancement free of this problem and to state merely that I will treat the "moral" in "moral enhancement" as referring to the more factually verifiable morality of the actions of the enhanced subject, without making any assumptions about the possible change in moral character that it may or may not bring about.[1]

Even if our minimum conception focuses on moral action rather than moral character, the question will still arise as to which kinds of actions are good and which ones are bad. Would an intervention in a highly patriarchal state that caused women to be more submissive towards men be a "moral enhancement"? There are still cultures in existence that would no doubt consider such an enhancement to be highly moral, and it would be very easy for a definition of the term "morally bad actions" to get bogged down in a debate about moral relativism. The problem of moral disagreement has received extensive coverage in the debate about moral enhancement.[2]

We can bracket this thorny issue safely for the following two reasons: firstly, enough general agreement exists about the immoral nature of certain types of acts (such as the torture of an innocent child) to allow for a hypothetical intervention to limit itself to actions in respect of whose morality there could be broad agreement. A vast proportion of the suffering in the world is due to actions in respect of whose moral value there

---

[1]For related attempts, albeit not as starkly contrasting moral action from moral character, to distinguish predominantly factual claims about moral enhancement from purely normative ones, see Pacholczyk (2011) and Bruni (2011).

[2] See Douglas (2011), Persson and Savulescu (2011), Harris (2012), Shook (2012), DeGrazia (2013).

is little room for doubt. I will therefore not attempt to discover what the actions of a morally enhanced person would be, were she to find herself in a situation where she could save the lives of many by pushing a fat man in the way of a runaway trolley. If moral enhancement can prevent the individual who tampered with the trolley's brakes to begin with from doing so, that will be a more than adequate result.

Secondly, even based on the limited literature on the subject, it is by no means clear if generic varieties of moral enhancement that can be applied across the board to all individuals are a feasible proposition; moral enhancements likely to appear in the foreseeable future could well have a highly localised and narrow field of application and may even have to be of the "designer" variety, aimed at specific areas earmarked for enhancement in a given individual (Shook 2012).

2.2.2 "Enhancement"

The term "enhancement" can also be interpreted in different ways. In this section I will look at three of these possible meanings: whether enhancement should include the notion of therapy; whether enhancement should be a boon to the agent; and whether enhancement should include traditional, non bio-medical means of alteration.

Enhancement vs. Therapy

According to Eric Juengst, the term "enhancement" is "usually used in bioethics to characterise interventions designed to improve human form or functioning beyond what is necessary to sustain or restore good health" (1998:29). Adherents of such a definition of enhancement would distinguish it from "therapy," which they consider to be the bringing of a trait that is below the human "norm" in line with that norm. Thus, if anabolic steroids are administered to a child with growth failure, such an intervention

would be deemed therapeutic, as it would merely aim to bring the growth of the child in line with that of his peers. If steroids are administered to a healthy athlete so that they increase his strength beyond what could be obtained through even the most rigorous exercise, such intervention would, according to proponents of this kind of distinction, be considered "enhancement".

This distinction has been central to certain arguments concerning the moral admissibility of enhancement, such as the contention by Leon Kass (2002) that normal human functioning of an individual is a result of the evolutionary history of the human species that we should alter only at our peril. Moral enhancement can likewise aim to increase a human being's capacity for moral action to a level within current norms encountered in the species or beyond. Nicholas Agar refers to this norm as "normal moral competence":

> What we might call normal moral competence emerges from the interplay among normal moral motivation, normal moral insight and normal moral behavioural capacities. People with normal moral competence are often, but not always, motivated to do what is right. They have good, but not perfect insight into what is right. They often, but not always, conform to appropriate norms. Morally normal people are not Josef Stalins, but they are not Mahatma Gandhis either. (Agar 2013:1)

Biomedical increase of a psychopath's "moral competence" to bring this capacity to within the boundaries of this norm could ostensibly be termed "moral therapy". Likewise, Agar holds, it may be possible to increase a human's moral competence to beyond the current norm. "Beyond the current norm" could mean altering a person whose moral competence is within the boundaries of normalcy to the higher-end scale of extremely virtuous humans currently already in existence, or even beyond. It is this latter kind of radical moral enhancement that Persson and Savulescu contemplate when they argue that the human species is inadequately endowed for survival beyond the twenty-first century, as a result of a norm of moral competence that has not evolved beyond the

realities of the small communities in which our species has lived for the greater part of the past 100,000 years (Persson and Savulescu 2012).

From the perspective of our argument, both moral "therapy" and moral "enhancement" could potentially threaten the autonomy of a subject, and this factor, together with the difficulties in deciding where the current norm starts and where it ends, makes it undesirable for our minimum conception of moral enhancement to exclude the one or the other from our argument. I will thus include potential moral "therapy" under the umbrella of enhancement in this paper.

Augmentation and benefit

Apart from the enhancement–therapy distinction, "enhancement" can also be taken to mean merely the augmentation of a characteristic or the augmentation of a characteristic such that it is of benefit to the subject. This is a distinction that is frequently drawn in discussions of other kinds of enhancement, and many of these clearly bestow some benefit to the subject – whether greater athletic strength, increased physical attractiveness or a better memory. Moral enhancement may differ in this respect, as it is quite possible that such enhancement may often not be prudentially advantageous to the subject thus enhanced. In a world where humans frequently further their ends through immoral actions, it is not unlikely that a morally enhanced person may suffer more harm at the hands of the morally unenhanced than would otherwise have been the case.

This distinction is likewise not of relevance to our minimum conception, as enhancement of both kinds could potentially limit the autonomy of the subject. Limiting our understanding of "enhancement" to mere augmentation as opposed to augmenta-

tion+benefit also holds the risk that certain important kinds of moral enhancement may end up excluded from the scope of this thesis.

Enhancement and biomedical enhancement

Lastly, some of the authors on moral enhancement have switched to the term "moral bioenhancement" to distinguish it from traditional moral improvement, on the basis of the thesis that many traditional forms of alteration of the self are indistinguishable from biotechnical enhancement and that the dichotomy is therefore a false one. That is a matter of convention, and I will be using the term "moral enhancement" exclusively in the sense of biotechnical moral enhancement, reserving the term "moral improvement" for traditional means of reducing immoral action in an individual.

2.2.3 "Moral+Enhancement"

Having examined the terms "moral" and "enhancement" separately, I can now turn to the minimum conception of "moral enhancement" that I will be using in this paper.

According to Thomas Douglas, who set off the current "moral enhancement" debate with the publication of his article by the same name in 2008, "[a] person morally enhances herself if she alters herself in a way that may reasonably be expected to result in her having morally better future motives, taken in sum, than she would otherwise have had" (2008:228).

This definition is notable in several respects: firstly, it does not demand that the intervention will actually be successful; it is enough for an increase in morally good motives to be the goal of the enhancement in order for it to be termed "moral enhancement". Secondly, it limits itself to the morality of the actions of the subject, not the morality of the character of the subject, thereby sidestepping the separate issue (an

important one) of whether the subject of the enhancement would, herself, become a more moral individual as a result of the enhancement.

Lastly, the definition does not limit the means by which the enhancement is to be attained, provided that the intention behind the enhancement specifically concerns moral improvement of the motives of an individual. If by "motives" we understand such first-level pro-attitudes that actually move the subject to act in a specific way, these can presumably be influenced in different ways, whether through targeting the attitudes themselves or higher-level forms of conscious deliberation. Understood in this way, I have no objection against the use of the concept of "morally better motives" in a definition of moral enhancement, even if the term "motives" may prove problematic as a result of a lack of agreement in the philosophy of action as to how we are move to act the way we do.

I will therefore define the minimum conception of moral enhancement, for the purposes of this thesis, as *any biotechnical intervention in respect of a human being aimed at decreasing the prevalence of morally bad actions performed by that individual.*

2.3 The claims of moral enhancement

Having established our minimal definition of moral enhancement, I am now in a position to discuss certain key claims that proponents of moral enhancement make and on which they base their arguments for the feasibility of such enhancement as well as the arguments in favour of the need for moral enhancement. I will then conclude the chapter with a discussion of the most controversial of these claims:

C.1 Human beings have varying moral dispositions, i.e. tendencies to commit actions or omissions that are morally bad and can harm other sentient beings.

C.2 It would be good to alter these dispositions in order to reduce the number of morally bad actions committed by human beings and, perhaps, also to increase the number of morally good actions by humans, provided that the means by which this change is attained does not lead to a greater wrongdoing or other undesirable consequences.

C.3 Society has traditionally attempted to reduce the disposition of individuals to engage in immoral actions by means of moral education, whereby individuals will themselves become less inclined to commit immoral actions.

C.4 Some of the biological causal factors that determine the moral disposition of an individual seem to be so thoroughly "wired" into our biology that the effectiveness of both traditional means of moral education and careful, conscious, rational deliberation in overriding these dispositions is significantly limited.

C.5 If certain behavioural characteristics (including moral dispositions) of individuals and the entire species are influenced by facts about our biology, then the alteration of certain biological features of humans will also be capable of altering the characteristics.

By far the most contentious of these premises is C4, as social sciences have long adhered to the empiricist dogma that the mind of the human being is a blank slate at birth and that traditional means of education from an early age can mould an individual, regardless of his background or natural endowments, into an upstanding citizen. Doubts are also expressed regarding C5, given the complexity of moral psychology on a neurological level.[3]

[3] See for instance Zarpentine (2012).

2.3.1 Nature vs. nurture

The fourth claim, which pertains to the limit of traditional moral improvement as a result of biological predispositions to immoral activity, may be viewed as contentious by some who believe the new-born mind to be almost entirely a tabula rasa or "blank slate" whose features will be almost wholly determined by social factors such as environment and upbringing and that all—or most—behaviour is learned behaviour.

Such "all-or-nothing" views, which were predominant for a substantial part of the twentieth century, are now beginning to subside. The current state of this debate sees an increasing number of contemporary students of human behaviour viewing psychological traits as being caused by a complex combination of both environmental and genetically inherited factors.

There is a striking body of evidence to demonstrate that certain antisocial behavioural traits directly associated with immoral actions, such as a propensity to violence, are to an extent hereditary. A recent study that has sought to collate the findings of more than 100 quantitative genetic studies of antisocial behaviour has found that genes influence approximately 50% of population variation in antisocial behaviour (Moffitt 2005).

Thus, studies performed on monozygotic (identical) and dizygotic (non-identical) twins have demonstrated a stronger than average correlation in the ability to empathise on the part of the monozygotic twins, which is indicative of at least some genetic basis for aspects of empathic behaviour (Davis et al. 1994). There is also a significantly greater similarity in the way in which monozygotic twins follow a "tit-for-tat" strategy, than seen among dizygotic twins (Wallace et al. 2007). This is indicative—among other things—of at least a partial genetic basis for a propensity to fairness.

In a study published in 2008, 409 pairs of twins were made to believe that their mothers were in pain. The reactions of the twins to this display of suffering were measured as an indication of empathy at various intervals between the ages of 14 and 36 months. It was found that environmental effects decreased and genetic effects increased over time (Knafo et al. 2008).

Simon Baron-Cohen cites several genes that seem to play a role in the shaping of moral behaviour:

- The MAOA (monoamine oxidase-A) gene seems to play a role in determining the extent to which serotonin (a neurotransmitter which in excess may lead to increased aggression in an individual) is cleared between the synapses in the brain.
- The serotonin transporter gene SLC6A4 and AVPR1A (the arginine vasopressin receptor 1A gene) both seem to influence the ability of an individual to recognise specific emotions in other persons, an important component.
- A study of a group with a low "empathy quotient" (EQ)5, showed similar prevalence of CYPB11B1, a gene that influences the production of oxytocin, a hormone that has been found to influence the level of trust and empathy an individual can experience, as well as genes WFSI linked to social-emotional behaviour, and NTRK1 and GARBR3 which both influence neural growth (Baron-Cohen 2011:131-138)
- Lastly, Baron-Cohen lists several genes that seem to show a "significant association" with autistic traits, some of which are also concomitant with a-social behaviour.

The interrelationship between genetic and environmental factors that influence the moral dispositions of an individual is a complex one. In the case of the MAOA polymorphism, studies have indicated that the presence of the genotype could only be statistically linked to antisocial behaviour if an individual with that genotype had been exposed to abusive or adverse environmental conditions (Moffitt 2005:82-85). Thus a person who has the MAOA genotype and the misfortune of growing up in an abusive environment has a high probability of antisocial behaviour. A person without the genotype would, conversely, have a better chance of overcoming adverse environmental conditions.

These genetic factors do not only influence moral behaviour, but may also limit the extent to which environmental factors can shape and alter the behaviour of an individual – that is to say, in those cases where environmental factors play a substantial role in shaping the behaviour of an individual, that role may itself be limited by the genetic makeup of the individual. Such limiting of environmental influence may, depending on the context, play both a positive and negative role in the formation of morality, as it may limit the influence of both positive influences such as moral education and negative influences such as sustained abuse.

2.3.2 Knowing the good and doing the good

At this stage it may be countered that a limiting influence of biological factors on the ability of an individual to behave morally can be counteracted by a moral education. Prominent bioethicist John Harris sums up this view when he counters the moral enhancement project by stating that cognitive enhancement, i.e. the biomed-

ical enhancement of an individual's ability to store and process information, should be sufficient to "morally enhance" individuals:

> The most obvious countermeasure to false beliefs and prejudices is a combination of rationality and education, possibly assisted by various other forms of cognitive enhancement, in addition to courses or sources of education and logic. (Harris 2011:105)

It would appear as if the ability of human individuals to act in accordance with the fruits of their moral deliberation is somewhat limited, however, by the architecture of the brain. A detailed discussion of the functioning of the brain is entirely outside the scope of this paper, but a cursory overview of what neuroscience has established—or strongly suspects—about the neural mechanisms involved in morally relevant cognitive activities will be necessary for my argument later on.[4]

Even prior to the advent of modern neuroscience, introspection and our intuitions pointed towards at least two different processes at work inside the human mind, reason and feeling. The interplay between the cognitive and affective spheres of the human psyche has been the subject of much speculation in philosophical thought, with Aristotle maintaining that a life lived in accordance with reason is necessary for a human to flourish (*Nicomachean Ethics*, chapter 1) , while Hume famously stated that "reason is, and ought to be, the slave of the passions" (*A Treatise of Human Nature*, B2.3.3).

The notion of a conflict between reason and feeling is still to an extent supported by certain neuroscientists, although there are increasing indications that the picture is a bit more complicated. Modern neuroscience has largely confirmed that the brain consists of numerous computational centres, seemingly operating with significant independence from

---

[4] I will not be able, within the constraints of this paper, to present an elaborate argument for the nature of the relationship between the mind and the brain. Suffice it to say that I hold mental properties to depend on physical ones, regardless of whether the mental properties can be reduced to the physical ones.

one another. A reasonable amount of light has been shed on the localisation of these computational centres and some of the broad functions that they serve. This fragmentation seems to be mirrored by the mind, and our mental events are likewise not the result of a single "self" within the mind.

The brain only has a limited amount of mental resources it can utilise for the vast amount of computational activity that has to occur within it at any given moment[5] (Baumeister et al. 1998). Consciousness, conscious deliberation and especially conscious acts of volition are particularly slow, "costly" and computationally inefficient and the brain therefore relegates as many activities as possible to computational centres that are preconscious (Kahneman 2002). An overwhelming amount of evidence has been amassed to show that perception "takes a short cut" directly to action, rather than via the far more computationally inefficient route of conscious mediation (Bargh and Chartrand 1999). This side-lining of conscious rational thought is not only present in semi-instinctive actions such as the tying of one's shoelaces, but also predominates in acts of moral judgment.

Moll et al (2008) cite the results of numerous recent studies that seem to indicate that activation of brain regions in purely rational choices cannot lead to action as they lack motivational power. They summarise the theory as follows:

> [A]ll morally relevant experiences are considered to be essentially cognitive-emotional association complexes. Instead of competing with each other, cognition and emotion are continuously integrated during moral decision making. (Moll et al. 2008:167-168)

Thus it would seem that reasoning on its own may be largely bereft of motivational force and that some form of affective stamp of approval is neurologically required before reasoning can lead to some kind of action. Internal conflicts over choices would therefore

---

[5]It has been calculated that the brain receives approximately 11,000,000 pieces of information per second, but that we can consciously process at most 40 pieces of information per second. (Wilson 2004:24)

not be between reason and passions, but between different of cognitive and emotional "pairs".

The experiments performed by Jonathan Haidt go even further. According to Haidt, moral reasoning is seldom if ever seen to cause moral judgment and moral reasoning is more often than not a *post-hoc* construction that comes after the moral judgment has already been made. As Haidt puts it: "Conscious reasoning functions like a press secretary who automatically justifies any position taken by the president". (2012:91)

These findings are by no means the last word to be said in respect of the cognitive underpinnings of moral thought and moral action, but they seem not to square with internalist views that posit motivation as necessary for moral judgment. The combination of "rationality and education" that Harris sees as the key to moral improvement may therefore be of limited use in improving human morality without also targeting the preconscious processes that turn moral judgment into moral action.[6]

Most proponents of moral enhancement therefore believe that for such enhancement to be successful, it would have to target both our motivations and ability to reflect on what is right. David DeGrazia states:

> Motivational improvement and improved insight are conducive to behavioural improvement. Other things being equal, either type of improvement will tend to bring about better behaviour. If someone's moral motivation remains constant as she gains moral insight, she will become more likely to do what is right (assuming she has any motivation to do so) because she's more likely to know what is right and therefore what to do. Conversely, if someone's moral insight remains constant as his moral motivation improves, he will become more likely to do what is right (assuming he has any moral insight at all) because he will be more motivated to do it. A fortiori,

[6] For the limited extent to which improvement in moral reasoning seems to improve moral action, we may also refer, only partially tongue-in-cheek, to the empirical research of Eric Schwitzgebel, who found no difference between the behaviour of moral philosophers and other individuals. As Schwitzgebel states: "It remains to be shown that even a lifetime's worth of philosophical moral reflection has any influence upon one's real-world moral behavior." (Schwitzgebel and Rust 2011:67)

> the conjunction of motivational improvement and improved insight will be conducive to behavioural improvement. (DeGrazia 2013:2)

2.3.3 The means of moral enhancement

To conclude this chapter, I will briefly examine the claim made by proponents of moral enhancement that it is possible to alter the biological traits of individuals responsible for immoral behaviour, as well as some common objections to these proposed interventions.

In the first theoretical works on moral enhancement, two basic approaches to the alteration of dispositions have been proposed: the amplification of dispositions that will bolster moral motivation and the reduction of mental processes that impede moral motivation.

Persson and Savulescu (2008) postulate that moral enhancement should strive to increase dispositions such as altruism or a sense of justice that are conducive to moral behaviour, on the basis of the hypothesis that human beings have, by virtue of their biological make-up, certain "natural tendencies" or "core dispositions" to respond in specific ways to specific types of environmental stimuli (2008:168). Thus there is a substantial body of research into certain forms of innate psychological dispositions that seem to lie at the roots of what is generally known as "racism" and that "[people] encode the race of each individual they encounter, and do so via computational processes that appear to be both automatic and mandatory" (*ibid.*).

While such a type of disposition leads to immoral behaviour, there are certain other core dispositions—according to Persson and Savulescu—which are conducive to moral behaviour, and whose enhancement can be taken to amount to a form of moral enhancement. As noted above, the authors highlight two specific dispositions that would seem to lie at the root of moral behaviour – altruism and a sense of justice or fairness. By altruism, Persson and Savulescu understand the ability to "sympathize with other beings, to want their lives to go well rather than badly for their own sakes," (*ibid.*) and consider it to be one of the most generally accepted components of morality.

A sense of justice or "fairness" is the second identified category of moral behaviour that the authors consider to be central to human morality. This sense of justice seems to be composed of a complex set of different core dispositions, collectively known as "tit-for-tat," i.e. dispositions that are conducive to reciprocity. Thus, for instance, humans may reciprocate the granting of a favour with gratitude and an offence with anger.

Persson and Savulescu hold that moral enhancement of a subject would require a careful balance of altruism and fairness:

> More altruism is likely to initiate more tit-for-tat exchanges, though too much altruism may be an obstacle by making us turn the other cheek when tit-for-tat requires retaliation. Too little gratitude may provoke anger and aggression in benefactors rather than further favours; too much anger in response to aggressors may spark off an escalation of violence rather than simply deterrence of future violence, and too little anger might not be a sufficient deterrent; the same is true of too little and too much forgiveness. So, altruism, and tit-for-tat emotions need to be properly attuned to be maximally useful. (*ibid.*)

Thomas Douglas has proposed that moral enhancement be achieved by means of "causing ourselves to have morally better motives" (2008:229). Douglas understands motives to be "the psychological—mental or neural—states that will, given the absence of opposing motives, cause a person to act"(*ibid.*).

Douglas expresses the opinion that there are several potential interventions that could uncontroversially be labelled as "moral enhancement" for a large number of persons and in different situations, of which he outlines one:

> [T]here are some emotions—henceforth, the counter-moral emotions—whose attenuation would sometimes count as a moral enhancement regardless of which plausible moral and psychological theories one accepted. I have in mind those emotions which may interfere with all of the putative good motives (moral emotions, reasoning processes, and combinations thereof) and/or which are themselves uncontroversially bad motives. Attenuating such emotions would plausibly leave a person with better future motives, taken in sum. (2008:231)

Douglas cites two possible species of "counter-moral emotions"[7] whose diminution may lead to an increase in morally good motives, namely an aversion to persons of particular race groups and the impulse to commit acts of aggression. These dispositions may undermine the ability of an individual to engage in moral behaviour.

Despite some tentative empirical studies in this regard[8], the extent to which an intervention could be applied with such pinpoint precision is still very much open to doubt. If the intervention is not sufficiently precise, the danger exists that it may throw the baby out with the bathwater.

Thus, a frequent, and plausible, objection against alteration of dispositions, whether through their strengthening or attenuation, is that all of these dispositions may, in certain circumstances, be required for moral behaviour. It may well be that the aversion to individuals of other racial groups forms part of a more complex neural basis responsible for the identification of, and the emotional stance towards, individuals as members of groups: feelings of racism and feelings of kinship may be opposite sides of the same neurological coin. It may not be possible to single out racial aversion for attenuation without weakening an individual's attachment to his family.

Aggression (including quick, impulsive aggression) may also be entirely warranted in certain situations to avoid a greater wrong from occurring (Chan and Harris 2011).

This objection may be countered in two ways. Firstly, it is highly likely that moral enhancement treatments need not be of the "one-size-fits-all" variety. Shook (2012) predicts that some moral enhancers may take the form of highly specialised, even individu-

---

[7] Douglas uses the term "emotion" in a neurobiological, rather than the traditional philosophical, sense of the word, as referring to a disposition that is accompanied by certain conscious psychological sensations and is frequently endowed with motivational force.

[8] For an empirical study on possible inhibiting effects by the beta-blocker Propranolol (known in South Africa under the brand name "Inderal") on "implicit racial bias," see Terbeck et al. (2012). Deep brain stimulation as a means to countering aggressive impulses is discussed by Franzini et al. (2005)

alised forms of intervention designed according to the preferences and needs of the subject. Each form of moral enhancement would have to be assessed carefully in terms of its effects and side-effects. It may well be that some complex forms of immoral behaviour will not be modifiable without serious side-effects and therefore have to be rejected. Moreover, in order not to affect other psychological traits that may be dependent on the same neural mechanisms, interventions may have to be limited to more modest objectives.

Pacholczyk (2011) maintains that some of the proponents of moral enhancement suffer from over-heightened expectations. One should perhaps approach moral enhancement not as something that can solve all of society's problems overnight by turning reprobates into saints, but rather with the same, more modest, expectations that we have of, say, anti-depressants or pharmaceuticals targeting ADHD, which are occasionally ineffective and often have subtle, rather than earth-shattering, effects.

Another important objection to the modulation of counter-moral emotions or the strengthening of motivations such as fairness or empathy as proposed by Persson and Savulescu concerns the seeming disregard in which these methods seem to hold human rationality and human free will.

It is this type of non-rational moral enhancement that has been criticized explicitly by some observers, most prominently John Harris (2011, 2012) for decreasing the freedom of the enhanced subject. Harris is not opposed to the general notion of moral enhancement, provided it is based on a type of cognitive enhancement that will lead humans thus enhanced to reflect more rationally about why certain types of action are bad. But once a cognitively enhanced human has, for instance, come to the realisation that stereotyping another human on the basis of his race is actually an irrational act with no basis in any objective features of representatives from that race, he should still be allowed the freedom to act in accordance with his reasoning or not.

Persson and Savulescu (2011:11) counter Harris with an argument based on free will and determinism. They hold that our actions are either fully determined by biological fea-

tures or they are not. If our actions are fully determined, we can only enjoy freedom inasmuch as it is compatible with this determinism, and "judiciously applied" moral enhancement will only mean that we are now determined to act in accordance with our conception of the good, i.e. to behave in exactly the same way as a person who is morally good through natural moral endowment. If, on the other hand, our actions are not fully determined by our biology, then moral enhancement of the kind that, for instance, increases our disposition for empathy, would not curtail our freedom, as it would be incapable of influencing these free-will actions. Thus there would, according to Persson and Savulescu, be no difference between the freedom enjoyed by people who are "naturally moral" and those whose dispositions have been altered in a more moral way.

Douglas, who has proposed a form of moral bioenhancement that would directly modulate emotional biases impeding moral reasoning, agrees with Harris that moral enhancement may, in certain forms, reduce the freedom of humans to have immoral motives. (2011:22)

This freedom to be immoral may have significant value, by virtue of its being a consequence of the fact that we have any freedom at all. Douglas is even willing to grant that this value may in fact redeem the substantial disvalue brought about by the freedom to act immorally, the disvalue of immoral acts committed by so many people. However, he believes that the emotional biases that afflict prudential and moral reasoning act as an even greater "brute constraint" on freedom. Douglas' proposed type of moral enhancement, seeking as it does to modulate these biases, may therefore increase our freedom to act morally without necessarily lessening our freedom to act immorally.

2.4 Conclusion

In terms of our minimal conception of moral enhancement, a great many forms of potential interventions could fall under this category. A single answer to the question of whether

moral enhancement could negatively affect personal autonomy is therefore not possible, and we will have to assess the effects of specific kinds of intervention on personal autonomy. Before we can do that, however, we need to gain additional conceptual clarity in respect of personal autonomy. That is the subject of our next chapter.

Chapter 3

Personal Autonomy

3.1 Introduction

As I stated in Chapter One, personal autonomy is considered to be of distinct value to an agent and a *prima facie* right that no other party may infringe upon without a very weighty reason to do so. In order to establish whether moral enhancement would constitute a threat to personal autonomy, it is necessary to gain some understanding of what personal autonomy is and what the conditions for its presence in an individual are.

"Autonomy" is originally a political term, used to refer to the right enjoyed by nations to legislate for themselves. When applied to agents, an autonomous person is somebody who is able to determine or "govern" his life, i.e. able to conduct it in accordance with values, preferences and desires that she has authored or at the very least endorsed.[1]

Various accounts have been given of ways by which we may reliably establish whether the actions of an individual are indeed independently approved of by the agent, but these have been bedevilled by the pervasiveness of influences, subtle and not-so-subtle, that could subvert the agent's independence as well as the immense difficulties in finding an authoritative "core self" to which we could anchor a plausible account of an autonomous individual. The various debates in this regard have therefore led to a state of affairs where

[1] This kind of autonomy, often referred to as "personal autonomy" is to be distinguished from "moral autonomy," especially of the Kantian variety, which holds an individual to be acting autonomously insofar as he is "not guided just by his own conception of happiness, but by a universalized concern for the ends of all rational persons" (Waldron 2005:307).

authors incessantly go back to "tweak" their accounts, heaping condition upon condition. The challenge of this chapter is to arrive at a plausible list of conditions that I will use in subsequent chapters as a touchstone for the evaluation of the influence of moral enhancement on personal autonomy. In order to do this, I will commence with an overview of the different accounts of personal autonomy, picking out the relevant factors that each of these accounts can contribute to such a list. I conclude the chapter with a summary of these conditions.

3.2 Accounts of personal autonomy

Autonomy as self-government is at its heart a procedural notion, as it merely maintains that the internal process by which individuals endorse or determine what kind of life they wish to live should be the agent's own, without placing any normative strictures on what kind of life this must be. Critics of the notion of personal autonomy have maintained that such a purely procedural notion originates from a particular Western (and male) species of individualism that ignores the essential relationship between the individual and other people and should include some substantive normative content as well, such as the ability to discern right from wrong.

Substantive accounts of autonomy are, in turn, criticised for trying to graft onto it more baggage than it can reasonably bear and turn it into a single concept that is not only a necessary but also a sufficient condition for human flourishing. As Neil Levy points out, personal autonomy defined as "self-government" *is* essentially a structural notion:

> If autonomy is self-government, then (absent further argument) it seems that an autonomous individual will not necessarily live a flourishing life. She need not be happy, or have largely true moral or nonmoral beliefs. Nor (more contentiously) need she enjoy liberty, at least not to its fullest extent. The autonomy debate is not the free will debate, nor is it the moral responsibility debate (though basic autonomy is closely connected to these notions, it is not identical to them). Self-government is a procedural notion, and it is at least *prima facie* a near synonym for autonomy in its basic sense." (Levy 2006:430)

The kind of personal autonomy I will be considering in this paper is not of the strongly substantive kind. My goal is to establish if a morally enhanced individual can still be considered capable of being in charge of his life. If there are additional content-filled conditions for personal autonomy, such conditions may very well encompass aspects of morality that a specific kind of moral enhancement would strive to improve, such as the capacity to distinguish right from wrong, but an examination of the compatibility of such a conception of personal autonomy with a particular kind of moral enhancement may very well be accused of question begging. For any conception of autonomy where autonomy is a certain kind of procedure+x and x is some additional normative value, the normative values imposed by a given species of moral enhancement would either bolster x, be neutral in respect of x or oppose it. The influence of particular kinds of moral enhancement on such normative values would have to be considered on a one-by-one basis. If any such substantive conditions are endangered by moral enhancement, I leave it to others to examine the extent to which they could be affected by a postulated species of moral enhancement.

3.2.1 Capacity for autonomy

Most authors on the subject of personal autonomy have noted that total autonomy is impossible to achieve, as our desires can never quite fully be of our own making.[2] An individual's capacity for autonomy is determined by the extent to which he is endowed with certain abilities or traits that allow him to be in charge of his life. What these may be will be discussed below, but for the time being I will only state that it is necessary to distinguish the extent to which an individual is actually autonomous from his capacity to be autonomous. It seems plausible to hold that an individual may have the capacity to be autonomous without actually exercising that capacity to its fullest.

[2] See Feinberg (1986), Dworkin (1988).

Even if we believe the capacity for autonomy to be insufficient for an individual to be deemed autonomous, a decrease only in this capacity, all other things remaining equal, will be a loss to an individual, given that those constraining factors which prevented him from previously exercising this capacity may disappear but leave him without the ability to make use of this capacity. Thus, a morally enhanced individual may not have exercised his capacity for personal autonomy prior to enhancement, but suddenly realise the need to do so after enhancement and yet find himself unable to do so as a result of the enhancement. In this kind of scenario, there will be no decrease subsequent to the moral enhancement in the extent to which the individual actually behaves autonomously, but he will still partially or fully have lost his capacity for personal autonomy.

3.2.2 Hierarchical accounts of personal autonomy

Most conceptions of personal autonomy have taken as their basis the "hierarchical" account of personal autonomy as expounded in the work of Harry Frankfurt (1971) and Gerald Dworkin (1988). In an influential account of freedom of the will that has also proven preeminent for accounts of personal autonomy, Frankfurt states that persons, unlike "non-persons" such as animals, do not only have desires but also desires about desires as a result of their ability to reflect about the things that we desire (1971:70).

There is thus, according to such accounts of autonomy, a hierarchy of lower- or first-order desires, values and preferences and a higher-order set of "desires-about-desires," beliefs and principles that have as their object the lower-order desires, values and preferences. In the words of John Christman, "[L]ower-order desires (LODs) have as their object actual actions of the agent: a desire to do X or Y; higher-order desires (HODs), have as their object other, lower-order desires." (1988:112)

For both Frankfurt and Dworkin, these higher-order desires, beliefs and principles can, if they are based on a process of critical self-reflection, bestow the status of autonomy on the lower order, through an act of endorsement that both Frankfurt and Dworkin refer to

as "identification". Dworkin (1988:25) refers to such cohesion between higher and lower levels on the basis of identification as "authenticity," and maintains that for such identification to be authentic, it should be the result of critical reflection (1989:61).[3]

Two key objections have been raised in respect of hierarchical conceptions of autonomy, however. Firstly, if identification on a higher level is required to make the attitudes of a lower level autonomous, would not an even higher level of identification be required to render the second-tier attitudes autonomous, and so on *ad infinitum*? Secondly, the danger exists that a focus on predominantly internal consistency will make such an account blind to threats of external manipulation.

We turn therefore to procedural accounts of autonomy that attempt to address these two problems.

3.2.3 Coherentist accounts of personal autonomy

The hierarchical schemas of Frankfurt and Dworkin have been deemed problematic by some philosophers as a result of the ontological primacy that Frankfurt and Dworkin bestow on the higher-order principles, for if each level of desires needs to be endorsed by a yet higher level of desires, we will go into eternal regress. If we state, on the other hand, that no endorsement at an even higher level is required for the higher-order attitudes, we are faced with the situation that an agent's autonomy originates outside the hierarchy of desires and that our higher-level desires are thus themselves not autonomous (the *ab initio* problem).

What is required is some lodestar other than the purely structural notion of conformance between the lower and higher orders against which to measure the authenticity of an agent's values, desires and preference. The simple solution would be to appeal to the

[3] Dworkin (1988) has subsequently distanced himself from the requirement of authenticity, maintaining that an agent should rather be able to alter his preferences and make them effective in driving his actions. The extent to which people can do that is, however, severely constrained.

autos in autonomy and to say that the higher-order attitudes need to correspond to the agent's "core self". Phrasing the matter this way would merely defer the problem, as we would merely be replacing an infinite regress of hierarchies. Moreover, the absence of a unified mind empirically militates against such a single "centre".

Coherentist accounts seek to root the authenticity of our desires, preferences and values not merely in the act of identification between a lower- and higher-order entity, but also in the belief that the desires, preferences and values must also be coherent, i.e. not contradict one another too starkly.

Two kinds of coherentist accounts of personal autonomy predominate. Synchronic coherence requires that the various desires, preferences and values form part of a reasonably integrated core at a given point in time, whereas diachronic coherence requires that they cohere across time. Laura Ekstrom (1993) posits a condition of synchronic coherence when she states that our preferences must be compatible with the full picture of a person's character, i.e. the entire inter-connected system of acceptances and preferences of an individual.

The problem with such a largely synchronic account is that it is not sufficient for higher-order desires to cohere among each other in order for us to deem them autonomous. It is quite possible for such an integrated set of desires to be the result of manipulation. Even if these desires cohere, can we still refer to the owner of this integrated set of desires as autonomous?

3.2.4 Historic accounts

One way of addressing this problem is to add a diachronic element to the picture, by insisting that our desires, beliefs and principles should not only cohere with one another at a given time, but also across time. As Michael Bratman states:

> An agent acts at a particular time. But adult human agents are not simply time-slice agents. Adult human agents persist over time, and their practical thinking concerns

> itself with and plays central roles in the organization and coordination of their activities over time. In this sense their agency is temporally extended. (Bratman 2007:98)

For Bratman, the long-range plans that individuals make in the course of their lives form a continuum over time that can give at least some semblance of a consistent self against which to measure the coherence of our higher-order desires. Such plans can, and do, change over time, but for the minimally autonomous agent, at least, there should be some continuous thread at any given instance of time that confers authenticity on the higher-level desires of an individual, without the need for higher levels of identification.

The extent to which long-range plans shape all our daily actions can be subjected to some doubt, however. Several other authors, such as John Christman (1991) and Alfred Mele (2001) have combined a diachronic approach with Frankfurt and Dworkin's requirement of critical reflection, requiring that the causal history of such reflection be of a kind that would exclude factors that would limit the agent's ability to engage in such reflection.

Exactly how such reflection should take place has likewise been the subject of much deliberation. Fischer and Ravizza (1998), who also view the actual acquisition process of a character trait as crucial to its status as autonomous, maintain that there must be some some reason-responsive mechanism ("guidance control") which is causally responsible for the preferences of an agent. An agent can only really be self-governing, this view holds, if he fully understands the reasons that ground his preferences. If the reasons are opaque to him, then he is not really able to account for his higher-order desires and cannot accept full ownership of them.

The requirement that the autonomous individual be responsive to reasons also raises the issue of such an individual's ability to reason. There is thus a prevalent view that the higher-order desires themselves must be based on a process of reasoning by the autonomous agent. If an individual, for instance, were to hold simultaneously different preferences which clearly contradict one another, could such an individual be considered autonomous? At the very least there should be some level of minimal rationality (Christman 1991:15) so

that no manifest inconsistencies would exist between the different preferences held by an agent:

> What this requirement for consistency entails, however, is that the autonomous agent does not act on the basis of mistaken inferences or violation of logical laws. If I believe that 'p' and I believe that 'if p then q' but I desire something X which is based on the belief that 'not-q' then the desire for X is not autonomous. (*ibid.*)

Both Christman and Berofsky (1995) believe that the requirement that a higher-order preference be consciously adopted is too strict a requirement, given the various hereditary and social origins that may underlie our preferences, as well as our inability to be fully mentally engaged at all times. As long as we are able to revisit such a preference, regardless of its origin, reflect on it in a rational way and then either accept or reject it, we should be at least minimally autonomous in respect of that desire. Berofsky writes:

> What matters to my status as an autonomous agent is the current disposition of the items, my present ability to understand, use, and, if necessary reject or modify the rule or principle. We may have acquired the rule by rote, but so long as it is sustained in virtue of its rational ground and in such a way that, were there good reason to reconsider its acceptability, that earlier learning experience would not inhibit a rational re-evaluation, our autonomy is not diminished. (Berofsky 1995:124)

There is, however, reason to doubt the extent to which most people are able to engage in careful reflection (even after the fact) of the reasons for their desires, beliefs and principles, and we may wonder if much (or most) of such reflection is not more akin to rationalisation than critical identification. It is not inconceivable—especially when one examines the divided view of the mind that some authors have extrapolated from the findings of neuroscience—that our higher-order attitudes are to an extent shaped by the lower orders.

Daniel Dennett (1992) famously draws an analogy between the self and the centre of gravity of an object, which is an abstraction and yet one that is very useful to characterise the behaviour of an object in space. The self proposed by Dennett is a coherent narrative self that is fashioned by individuals to impose an order on their existence – a chronological

order that explains their lives in terms of certain leitmotifs, causes and effects and imposes coherence upon something that may not be so coherent at all. Such a fictional narrative may itself provide a kind of core "self" against which to measure the authenticity of our attitudes and preferences.[4]

3.2.5 Externalist considerations

As noted earlier, there is a very real sense in which we cannot consider an individual who holds certain attitudes as a result of manipulation or deception as autonomous. One of the paradigmatic examples of autonomy deprivation in the literature is, after all, Iago's manipulation of Othello into murdering Desdemona. Purely hierarchical accounts of autonomy would not deem Othello less autonomous in his actions towards Desdemona, and yet if Iago had not committed his acts of deception, Othello would not have embarked on his murderous course of action.

The problem that all conceptions have to counter is this: all humans are subjected, from the moment of conception, to a barrage of external influences that we can only counter to a limited extent. In order for the very notion of autonomy to have any value, such a conception has to outline a plausible means for a person to maintain at least a measure of independence in the light of all these deterministic influences.

It is for this reason that these conceptions all focus predominantly on internal capacities that the agent must hold at least minimally in order to be considered autonomous, regardless of the provenance of external influences. If we view autonomy as a right, we cannot sustain a purely internalist conception of autonomy, but need to specify why certain external influences are unacceptable and others not. Some of these influences, such as traditional childhood education by caring parents, have great intuitive appeal, however, while others such as the swaying of opinions by means of manipulation clearly do not.

---

[4] For a recent consideration of authenticity based on such a narrative self, see Davenport (2012).

Some means of distinguishing "positive" influences from "negative" ones are therefore required. As Dworkin puts it:

> The problem of analyzing procedural independence is the task of characterizing those influences which in some way prevent the individual's decisions from being his own. [..] With respect to autonomy, conceived of as authenticity under conditions of procedural independence, the paradigms of interference are manipulation and deception, and the analytic task is to distinguish these ways of influencing people's higher order judgments from those (education, requirements of logical thinking, provision of role-models) which do not negate procedural independence. (1976:25-26)

Most accounts of personal autonomy have therefore stipulated an additional condition for autonomy, namely that the identification of the higher-order desire with the lower-order desire must not have been unduly influenced by factors external to the agent; in other words, the act of endorsement must, itself, conform to the condition of what Dworkin calls "procedural independence".

Dworkin (1976:26-27) provides some guidelines regarding acceptable methods of influencing an individual. Unacceptable methods are, for Dworkin, methods that either adversely affect the individual's capacity for critical reflection, or thwart such reflection through the deceptive provision of false values, reasons and the like. If the formation of a higher-order identification is based on critical reception, it should after all involve a scheme of evaluation that grounds its identification (or the withdrawal of such identification) on some value or reason. Even a highly developed capacity for critical reflection may after all be useless in the face of manipulation or deception.

Not only certain cognitive functions (such as the capacity for rational deliberation) seem to be necessary for personal autonomy, but that affective responses of a specific kind may be required as well. An individual must, it would seem, have a sense of self-worth in order to "be his own person" and to take a stance on his preferences. If the altering influence were to demean that sense of self-worth, an individual who in all cognitive respects seems able to arrive at his higher-order preferences in an independent manner, may consider

himself to be incapable of doing so and therefore abstain from altering these preferences. Thus, for Dworkin, methods of influence that undermine the "self-respect and dignity of those who are being influenced" (*passim*) are unacceptable because these undermine the faith that an agent must have in his own ability to engage in critical reflection.

Methods of influence that cause irreversible or long-during changes during which the agent will be "stuck" with the changes, as well as changes that take place with such rapidity that the agent does not have the opportunity to engage in critical reflection need to be approached with great caution.

Methods which "affect in fundamental ways the personal identity of individuals" are also undesirable. For Dworkin, such influences would cause discontinuities in a person over a short space of time, as we need to "maintain a coherent and unified conception of our own identity. Notions of personal responsibility, long-range plans, a connection with our past, all depend on our being sufficiently similar at various time-stages to be thought of as the same person" (*passim*).

Lastly, Dworkin gives preference to methods which "work through the cognitive and affective structure of the agent, which require the active participation of the agent in producing the change, to those which short-circuit the desires and beliefs of the agent and make him a passive recipient of the changes" (*passim*). This last stipulation by Dworkin seems plausible when taken to refer to the alteration of the higher-order attitudes of the subject, but problematic when applied to the lower-order attitudes of the mind, as numerous traditional methods of influencing do so in an uncontroversial manner, given the limitations of affecting change through the active participation of the subject. Nor does it take into account the possibility that the agent may, himself, wish to short-circuit his desires and beliefs if he wishes on the basis of his higher-order attitudes to alter them but is somehow limited in his abilities to do so. We will return to this issue in the following chapter, but shall for the time being stipulate only that methods which bypass the cognitive and affective structures of the mind will only uncontroversially decrease the autonomy

of the subject if they alter the higher-order desires, beliefs or principles of the subject in an irrevocable manner and without the agent being aware of these changes.

Some of the "negative influences" listed by Dworkin may, in the long run, have a positive effect on the agent's ability henceforth to engage in independent deliberation about what constitutes his values; others will undermine that ability. Where an ordinary education from parents or teachers instils in an individual the ability to reason and to reflect on the basis of such reasoning on what it is he values in life, such an influence may be deemed positive, and even if that education includes a healthy dose of "prepackaged" values, the reasoning skills necessary for an evaluation or re-evalution of our desires acquired in the course of the education will help that individual to revise his stance on those values at a later stage.

What is important, according to Dworkin, is the requirement that these influences do not bypass our consciousness, are not irreversible and do not affect our subsequent ability to reflect in a consciously rational way about our preferences and to alter them accordingly.

Christman provides a useful rule-of-thumb when he states:

> [A]ny factor affecting some agent's acts of reflection and identification is 'illegitimate' if the agent would be moved to revise the desire so affected, were she aware of that factor's presence and influence. (Christman 1989:61)

Dworkin's stipulations concerning procedural independence bring us closer to a view of personal autonomy that would disallow preferences shaped by manipulation. The procedural independence he states as a condition for personal autonomy does not seem to pertain only to the act of endorsement either, but also to the aetiology of both the lower-order and higher-order desires. This departure from the strictly structural approach of the standard hierarchical conception of personal autonomy underlines the importance of viewing the historical context within which the specific attitude of an agent has arisen.

3.3 The conditions of personal autonomy

As can be seen from the overview, various different sets of necessary and sufficient conditions for personal autonomy have been proposed by theorists. To conclude this chapter, I will outline the various conditions that I will use to assess the influence of moral enhancement on personal autonomy. I will try to be as comprehensive as possible in my acceptance of these conditions in order to make my requirements for personal autonomy as strict as possible.

Christman (2011) provides a useful schema for the categorisation of conditions necessary for autonomy to obtain in an agent:

> To govern oneself one must be in a position to act competently and from desires (values, conditions, etc.) that are in some sense one's own. This picks out the two families of conditions often proffered in conceptions of autonomy: competency conditions and authenticity conditions. Competency includes various capacities for rational thought, self-control, and freedom from debilitating pathologies, systematic self-deception, and so on. [...]
>
> Authenticity conditions often include the capacity to reflect upon and endorse (or identify with) one's desires, values, and so on. (*ibid.*)

Based on this overview, in order for a species of moral enhancement to be considered non-threatening for the personal autonomy of the subject of the intervention, it may not:

i. *Competency conditions*

   (a) Diminish an agent's cognitive abilities or psychological traits such as the ability to engage in critical evaluation thereby weakening the agent's ability to identify with or endorse his first-order desires, values or preferences.

   (b) Diminish an agent's ability to alter his higher-order desires, beliefs or principles.

ii. *Authenticity conditions*

(a) Cause a real or potential decrease in the coherence of the individual's higher and lower attitudes.

(b) Irrevocably alter his higher-order desires, beliefs or principles.

(c) Occur without the awareness of the subject, e.g. by means of deception.

(d) Cause a sudden and unwanted rift in the narrative identity of the individual's conception of his own self.

Despite the various differences that exist over the conditions of personal autonomy, the criteria outlined above should enjoy a substantial amount of support among a wide spectrum of authors. Not all these conditions are required by all views, and there is also disagreement as to which of these conditions are necessary for personal autonomy to obtain in a subject, and which are sufficient. As outlined in the introduction, I intend treating all of these conditions as necessary, in order to establish if any kind of moral enhancement could exist that would satisfy even the strictest conception of personal autonomy.

Chapter 4

The compatibility of moral enhancement and personal autonomy

4.1 Introduction

As I attempted to demonstrate in Chapter Two, moral enhancement could potentially develop in all manner of directions. For the sake of establishing what kinds of moral enhancement may have a negative influence on the personal autonomy of an individual, I will now, on the basis of the minimum conception that I developed in Chapter Two, categorise the different core elements of moral enhancement that are relevant to the personal autonomy of a human being. I will then, on the basis of the principles developed at the end of the previous chapter, proceed to discuss the ways in which these core elements are compatible with personal autonomy. This will provide me with a set of rules against which to measure the compatibility of any form of moral enhancement with personal autonomy. If any form of moral enhancement manages to conform to all these criteria, we can state that at least some form (or forms) of moral enhancement is compatible with personal autonomy.

4.2 The core elements of moral enhancement

In the section that follows, I will attempt to identify the core elements of moral enhancement. In order to simplify the discussion, I will seek to categorise these elements in terms of those aspects of the mind that are relevant to the conditions of personal autonomy outlined at the end of Chapter Three: the higher- and lower-order attitudes of the agent and the mental abilities necessary for the alteration of these attitudes and for critical self-reflection.

As we are discussing moral enhancement from the perspective of the alteration of actions, it is necessary to discuss some of the complexities that may arise out of disagreement over the causation of actions in the human mind.

4.2.1 Alteration of actions

In accordance with the minimum conception of moral enhancement that I offered in Chapter Two, moral enhancement is any biotechnical intervention in respect of a human being aimed at decreasing the prevalence of morally bad actions in that individual, i.e. by causing the ratio of morally bad actions compared to good actions to decrease. This goal can be achieved by causing more good actions to occur and/or by causing fewer bad actions to occur.[1]

An action can either be voluntary or involuntary. A muscle spasm that leads to the twitch of a limb is an involuntary action caused by a non-mental event. Voluntary actions are actions that are somehow willed by the agent to happen. There is substantial disagreement in philosophy concerning the aetiology of such voluntary action. At the root of this disagreement lies the fact that any given action may be preceded by various different mental events, many of which seem to be linked to one another in complex, mutually influential, relations. Thus a value or emotion may influence a desire, but a desire or value may also influence an emotion. Our lack of certainty about whether any of these mental events or states can, if willed, be sufficient to direct the actions of an agent in a specific direction need not frustrate this discussion, however. If these mental states influence each other,

[1] This is an area where deontological and consequentialist views of moral enhancement may disagree. It seems to me that if this is an either/or choice, it would be more important for moral enhancement to cause a decrease in bad actions than an increase in good actions. If a murderer after undergoing moral enhancement continues to murder, but instead of watching television in the intervals between his murders now participates in charitable life-saving activities, this would be a less desirable state of affairs than if the murderer stopped murdering, but spent most of his time watching television.

this would entail that the actions of an agent can also be influenced by targeting any of these mental states. If one were to hold that only a desire with the proper volitional force can precipitate action, but agree that the content of a desire can be altered by a specific kind of emotion, this would allow us to conclude that we could influence the action of the agent, albeit indirectly, by influencing the emotion. If it were to be true that an emotion may, with the necessary volitional force, directly lead to a specific action, this may be phenomenologically indistinguishable from a situation in which the altered emotion first had to lead to a particular kind of desire[2].

In the discussion that follows, I will speak about lower-order desires as the precursors to action, with the understanding that these desires may themselves be caused or influenced by other lower-order states such as values or emotions. If I therefore state that a lower-order desire can be altered, for instance, by decreasing the prevalence of anger, this should be understood to mean that the alteration of the prevalence of a lower-order emotion of anger will, in turn, alter some lower-order desire of the subject. I will also assume that these mental states do not only influence the content of desires or lead to the formation of new desires[3], but may also—importantly—increase the volitional strength of a given desire. Thus they may influence in a decisive manner which of a number of competing lower-order desires is eventually effective in moving the agent to act in a specific way. Lower-order desires may influence each other in like manner, both with respect to content and strength.

In a similar fashion, the higher-order mental states that have as their object the lower-order mental states may influence the lower-order states: a higher-order belief that a spe-

---

[2] It may increase the effectiveness of an intervention that targeted the emotion directly, however. We will discuss the relevance of the effectiveness of an intervention below.

[3] I am not altogether sure whether a desire can actually be altered, and whether such an alteration is not, in fact, the replacement of the one desire with another one, but for our purposes this distinction seems immaterial.

cific lower-order value is a good value to have, may influence the lower-order value to give rise to a new desire, alter an existing desire, or strengthen or weaken the volitional strength to render the desire effective or ineffective.

In Chapter Two, we defined moral enhancement as *any biotechnical intervention in respect of a human being aimed at decreasing the prevalence of morally bad actions performed by that individual.* If an action can only be caused by a desire endowed with the necessary volition force, we may therefore restate our minimum conception of moral enhancement as *any biotechnical intervention in respect of a human being aimed at decreasing the prevalence of effective morally bad desires in that individual.*

The above discussion allows us to outline the following basic means through which moral enhancement may alter the prevalence of morally good actions in an individual:

1. It may alter the actions directly without influencing the lower-order desires.
2. It may alter the lower-order desires.
3. It may alter the higher-order desires.
4. It may alter the force of the will so as to favour a particular lower-order desire to render it effective or ineffective.

I will now discuss each of these means of influencing the actions of an agent in turn.

4.2.2 Direct determination of actions

It is hypothetically plausible to determine the actions of an agent without the need for a causal lower-order desire. Thus one could conceive of a biotechnical mechanism, implanted in the body of an individual, that is able to monitor the surroundings and mental activity of an individual without forming a part of the mind of an agent. This mechanism would, based on the situation and mental activity of its host-subject, be programmed to determine when an agent is about to act in a morally bad way and subvert that action, for instance

by mechanistically influencing the motions or speech of the subject. If the device were to establish that the subject is about to harm an individual, for instance by pointing a gun in the direction of that individual and pulling the trigger, the device would ensure that the muscle actions in the arm that would lead to the raising of the gun to the necessary level would not take place and that the muscles in the trigger finger cannot contract. The actions (or lack of actions) of an individual with such an implantation would, insofar as they were determined by the implantation, be wholly involuntary. Such actions or inactions would be entirely heteronomous, as they would be determined by the device and not by the individual in whom the device has been implanted. The actions performed by the subject are not actually performed *by* the subject, but *by means of* the subject, who is no more than a tool in the hands of the device; a means and not an end.

Now it might be the case that the subject himself desired the device to be implanted in him. Take the example of Joe, a mild-mannered accountant who is prone to occasional fits of murderous rage that are triggered by particular, sometimes seemingly trivial, situations. When Joe is in the throes of such an episode, he spouts obscenities and will reach for the first object he can use as a weapon to inflict harm on whomever is around him, regardless of whether such a bystander even caused the onset of his state. While in such a state, Joe is unable to account for his actions, and only the passage of time allows Joe to regain his self-control. Afterwards, Joe inevitably feels immense remorse over his actions, and desires not to have the desires to commit the violent actions that he is capable of when in the hold of such a state of fury, but lacks the strength of will to make effective this higher-order desire not to be prone to fits of rage. Therapy and medication are powerless and he decides to have such a controlling device installed in him. The device is attuned to Joe's state of mind and will remain passive until it detects the onset of such an episode, at which point it will immediately kick into action and disable all signals that are sent to the muscles of Joe's speech apparatus, arms and legs. Once the fit has passed, the device goes into passive mode again. To an external observer, Joe, when slumped in a lifeless heap while under the

control of the device, appears to be in a vegetative state.[4] Inside his mind, however, Joe is still a seething inferno of anger, desperately desiring to hurt everybody in his sight, but totally impotent to do anything.

The example of Joe is not unlike the example of the hero Odysseus in Homer's *Odyssey* who orders his men to tie him to a mast, stuff beeswax in his ears and refuse to obey his commands when they encounter the Sirens. While the Sirens are singing, Odysseus tries to wrench free from his bondage, but his men do not heed the commands to untie him. The question of whether Odysseus, while powerless, is being autonomous, is answered as follows by Gerald Dworkin:

> Although [Odysseus'] behavior at the time he hears the sirens may not be voluntary—he struggles against his bonds and orders his men to free him—there is another dimension of his conduct that must be understood. He has a preference about his preferences, a desire not to have or to act upon various desires. He views the desire to move his ship closer to the sirens as something that is no part of him, but alien to him. In limiting his liberty, in accordance with his wishes, we promote, not hinder, his efforts to define the contours of his life. (Dworkin 1988:106)

Through the use of the device, Joe is ensuring that his actions conform to the higher-order desire he has of not harming individuals. This hypothetical device, as described, does not lead to the formation of particular desires in Joe but circumvents the mind entirely, directly causing actions (or inactions). It is also alien to his mind, without the ability to integrate into it and influence his desires, emotions or feelings[5]. The objection could thus be raised that Joe cannot possibly be construed as autonomous when he is lying incapacitated on the floor, desiring to act and yet finding something preventing him from acting.

This example shows the limitations of viewing autonomy outside of a diachronic, narrative context. If Joe is a loving family man for the greater part of his conscious life, such

---

[4]The use of such a device as a means to combating road rage is not recommended.

[5]Conceptions of the mind that allow for the possibility of an "extended mind," may consider the device, if endorsed by Joe, to be such an extension of his mind.

fits of violence stand out as anomalies in the story of his life. Joe would obviously prefer not to have the device implanted in his body, and he would obviously prefer, instead of the humiliating scenes where he is lying helplessly on the floor, for no such incidents to happen at all. Yet the lottery of nature has endowed him with occasional lower-order desires that he cannot overcome, despite having higher-order desires for these lower-order desires not to be effective. It may be that for the duration of such a fit of rage, his higher-order desires cohered with the lower-order desires to act violently or he may even have no conscious higher-order desires during such an episode. Any higher-order desires during an episode of blind rage seem to be inauthentic against the bigger backdrop of his life, however, and Joe with the device seems to be more autonomous diachronically, as he is merely forsaking one kind of heteronomous behaviour, unwilled by him, for another kind—that he wills—when he allows the device to deprive him on occasion of the ability to will his higher-order or lower-order desires.

The objection of John Harris, previously alluded to, that moral enhancement would deprive us of the valuable "freedom to fall" seems not to take into account the fact that Joe would, without the technological implant, have no freedom to chose whether to stop acting violently. He does have the freedom to chose whether to have the device installed or not, and when he makes that choice, he is effectively willing his actions to be of the kind that the device determines.

If Joe were a different kind of person, for instance one who secretly despised his spineless existence as an obsequious accountant and relished the anger attacks for the fear they caused in those around him, there would be greater coherence between the higher-order desires of the ordinary Joe and the murderous impulses he would occasionally experience. Such bouts of anger might seem to others to be anomalies against the general backdrop of his life, but these episodes would form a part of the texture of his own narrative. If others forced this second Joe to have the device implanted in him, the actions or inactions caused

by such a device would not be actions that he desired to have and the device would decrease his autonomy.

4.2.3 Alteration of lower-order desires

After this extreme example, we can now turn to forms of moral enhancement that alter the lower-order desires. Such forms of moral enhancement would, unlike the kind reviewed in the preceding section, not leave the mind as a bewildered onlooker, but operate through, and as a part of, the mind.

There may be many means of altering the lower-order desires: either directly, by causing a specific kind of desire to arise, or indirectly, by causing or altering another mental state—an emotion, for instance—that in turn leads to the formation of a lower-order desire.

Such a means of moral enhancement would attempt to assist the accountant described in our previous example so that he would no longer experience any episodes of murderous rage. It should be clear that action that springs from the lower-order desires of the subject's mind is greatly preferable to action initiated by something foreign to it. Whereas Joe with the implant would be totally autonomous in respect of the actions initiated by the device, there would be no such jarring episodes if the actions originated from his mind. This does not mean that such actions will never be heteronomous, however.

If the lower-order desires caused by the alteration do not cohere with his other lower-order pro-attitudes or with his higher-order attitudes, they will fail the test of authenticity. Let us take a different example, that of a racist called Andrew. Andrew grows up in a family where he is taught from a young age that people of other races are inferior to that of his own. It may be that this education simply serves to reinforce certain atavistic feelings of fear and loathing that he feels with regard to other races and that the racist education merely grounds these feelings in corresponding higher-order beliefs, such as the belief that members of other races are mentally inferior or have a greater innate disposition towards violence than members of his own race. It may also be the case that Andrew's

racist upbringing actually causes the lower-order feelings of revulsion and suspicion that he experiences with regard to other races. If the moral enhancement undergone by Andrew only addresses some of the lower-order attitudes that lead to racism, these attitudes will not cohere amongst themselves, thereby leaving him conflicted and in perennial doubt about his "true" desires.

Likewise, if Andrew's higher-order desires, beliefs or principles are still endowed with racist content, the presence of lower-order attitudes that contradict these higher-order desires will be alien to him. Andrew may, as a result of the moral enhancement, find himself falling in love with a member of another race and yet experience a sense of self-loathing, given that he still continues to believe in the inferiority of people who do not belong to his race. He may attempt to will himself not to have these desires, but be unable to do so. This lack of coherence between his higher-order beliefs and desires and his lower-order desires will make Andrew heteronomous in respect of these higher- and lower-order desires.

Such a state of affairs can be contrasted with another. Suppose Andrew has, prior to moral enhancement, begun to question his belief in the inferiority of other races as a result of critical reflection informed by day-to-day observation of members of these races. Andrew begins to realise that his upbringing was faulty and he no longer wishes to harbour racist feelings, but is unable to do so. He decides to undergo a kind of moral enhancement that will address the racist lower-order emotions. His higher-order desires now cohere with his lower-order desires and he is now more autonomous, as the enhancement has allowed him to have the lower-order desires that he desires to have.

As in the case of Joe, the moral enhancement has merely served to allow him to have greater influence over his actions, and is therefore a means to greater self-determination. Had he undergone the enhancement at the behest of another party, such as a paternalistic state, this may not be the case and he might find his beliefs at odds with his actions and be less autonomous for it. Such heteronomy may be short-lived, however, given the possibility

that the higher-order desires may come to endorse his imposed lower-order desires. I will discuss this scenario under the next heading.

4.2.4 Alteration of higher-order desires

The next category of moral enhancement that we will examine from the perspective of higher- and lower-order desires is the alteration of higher-order desires. Such a species of moral enhancement that attempts to increase the prevalence of moral action would target the higher-order desires in the hope that they would be effective in causing corresponding effective lower-order desires for such action to take place. As in the case of the lower-order attitudes, the higher-order attitudes could hypothetically be altered directly or through the alteration of other attitudes of the subject that would influence the higher-order attitudes.

Direct alteration of the higher-order attitudes can only plausibly mean the artificial implantation of such attitudes, for example the placement of a desire to desire some action, or a belief or principle in the mind of a subject. It is difficult to call this by any name other than brain-washing, as a higher-order attitude that is implanted in the mind of a subject without the subject having arrived at that attitude through his own mental processes, whether consciously or unconsciously, cannot rightly be thought of as being his own, even if it were somehow to cohere with his other lower- or higher-attitudes. Direct alteration of higher-order attitudes therefore fails the test of autonomy on procedural grounds, even if the attitudes themselves were to have a semblance of authenticity.

The question remains whether it is permissible on grounds of autonomy to alter the higher-order attitudes indirectly, by influencing them through the alteration of some other mental state or process capable of altering the higher-order desires. The most apparent means by which this could be done, would be through the alteration of the lower-order desires or through the alteration of such aspects of the cognitive apparatus of the subject by which he arrives at his higher-order desires. We will first turn to the alteration of higher-order desires by means of the alteration of lower-order desires.

If an altered lower-order attitude does not have a corresponding higher-order attitude, the agent can either endorse it or live in a state of conflict. Such an endorsement can either be as a result of critical reflection about the new lower-order attitude, or a post-hoc justification of that attitude, as it seems to be a common feature of human psychology that individuals will try to integrate even the most dissonant lower-order desires within the greater complex of higher-and lower-order attitudes. Thus it seems highly likely that alterations in lower-order attitudes that have not arisen as a result of higher-order volitions will frequently lead to an adaptation of the higher-order attitudes.

We can imagine a scenario in which racist Andrew does not voluntarily undergo moral enhancement to address his racist beliefs, but is coerced to do so by a paternalistic state. Andrew suddenly finds himself "looking with new eyes" at people of other races. Initially his higher-order beliefs about the inferiority of other races reject what his feelings tell him, but with the passage of some time he starts at first to question those beliefs, then to revise them, until he fully endorses them. He feels ashamed of his old, pre-enhancement self, and thankful that he was coerced into undergoing the enhancement. The moral enhancement has thus indirectly led to an alteration of his higher-order desires.

This example raises the problem of which higher-order desires to consider authoritative when it comes to the endorsement of a post-enhancement status quo: those held prior to enhancement or those held subsequent to it. In the case of Andrew's forced moral enhancement, his post-enhanced autonomous endorsement of the moral enhancement seems to make amends for the lack of such an endorsement prior to the enhancement. This is the approach of David deGrazia, who believes that only the retrospective higher-order desires need to be taken into account when establishing whether the enhancement is authentically the subject's or not (2005:112).

This scenario highlights one of the central problems that all conceptions of personal autonomy battle to deal with. Our lower-order attitudes are frequently, perhaps more often than not, caused by factors over which we have little to no control. Andrew's original racist

attitudes may have arisen in him at a very young age, through socialisation, a particular psychological endowment, or as a result of a traumatic incident that happened to involve a member of another race. Likewise, it may very well be that he experienced a fateful "Damascene conversion" at an advanced age, when a sudden, unexpected and entirely undeserved act of kindness by a member of another racial group set in motion an alteration of his racist attitudes. What, then, is the salient difference between a "natural conversion," precipitated by chance, and one that has been engineered in a scientist's laboratory and imposed on a subject, equally powerless to oppose it? And how is one to distiguish the moral education received by a child at a government school when he is at his most malleable from such an imposition upon an agent in adulthood?

The unease that we feel at the notion of an external party imposing its will on our mind seems to have several causes. Firstly, we see a difference between things that happen to us by chance and things that happen because they are intentionally willed by other people. If a mountaineer were to be crushed by a boulder that had come loose, we would hold that this is a possible risk that the mountaineer should have foreseen. If another individual had intentionally caused the boulder to fall on the mountaineer, we will not say that this action is permissible merely because the boulder could have fallen of its own accord by chance. We may object to this analogy, however, on the grounds that intentionally causing a harm is not the same thing as causing a benefit – we do not object when the state coerces us to observe laws if these laws are for our benefit, and moral enhancement could similarly be for our own benefit. Moreover, we already allow the state to influence the attitudes of children when they are of an age when they cannot reasonably be expected to have the cognitive capacity to shape those attitudes themselves, and moral enhancement imposed on an adult would merely serve to overcome a similar impairment of an adult set in his views. Lastly, many observers would argue that allowing another party this kind of power may lead to a slippery slope where governments (consisting of other individuals equally in need of moral enhancement) will increasingly interfere in the mental life of their citizens,

initially perhaps with good reasons, but eventually with greater and greater totalitarian intentions, until we find ourselves living in a brave, new world. Such fears may not be unwarranted, but "slippery slope" arguments are based on epistemic uncertainty and fears of an unknown future and therefore of limited value in reasoned discourse.

Most importantly, none of the above considerations seem to bear any relevance to the autonomy of the subject post-enhancement. It seems as if the biggest risk that is posed by an imposed alteration of lower-order attitudes to the autonomy of the subject is the risk that it may cause a rift in the narrative identity of the subject pre- and post-enhancement, especially if that imposition is of a forced nature and significantly upsets the continuity of the agent's long-term plans, attitudes and preferences pre- and post-enhancement. It is only a risk, however, and could be mitigated if the change brought about by the enhancements is subtle or if the imposition is of a gentler, conditional kind, such as if the state were to require that a subject who wishes to undergo cognitive or physical enhancement undergo moral enhancement first.[6] In such a scenario, a subject would be able to incorporate the imposed moral enhancement into his narrative identity, even if he did not have a higher-order desire of his own to undergo the moral enhancement.

We can therefore state that if a moral enhancement alters only the lower-order desires of the subject, if the subject in a procedurally independent manner endorses the new lower-order desires post-enhancement and if these desires and the manner in which they arose can be integrated by the subject into the diachronic narrative identity that straddles his life pre- and post-enhancement, then such an enhancement cannot be deemed to have caused a decrease in the autonomy of the subject.[7] It may well be that there are other

---

[6]It would be impermissible, from the perspective of personal autonomy, if the state were to achieve such enhancement through deception, for instance by surreptitiously lacing the drinking water with an imperceptible pharmaceutical agent.

[7]If a third party were to impose a moral enhancement on an agent, it would have to ensure prior to enhancement that all these conditions could be met, otherwise there will be a real risk that the subject will not be able to assimilate the changes with his lower- and higher-order desires. Unless

reasons why it should be morally impermissible for a party other than the subject himself to cause these lower-order desires in the subject, but such reasons would have to be based on considerations other than the personal autonomy of the subject.

A last way in which the higher-order desires of the subject may be indirectly altered, is through the alteration of the agent's cognitive abilities or psychological traits that play a role in the formation of the agent's higher-order desires. John Shook (2012) highlights such methods of moral enhancement in an overview of the subject. According to Shook (*op. cit.* 5–7) moral enhancement could reach its goals by enhancing an individual's thoughtfulness about "doing the right thing" or enhancing a person's ability to make moral judgments. These are both examples of cognitive enhancement, and would alter the higher-order attitudes of the subject concernibg the kind of lower-order attitudes that the subject would like to have.

It is not entirely apparent how successful such enhancement would be in targeting those aspects of cognition that would lead to an increase in correct moral judgments or thoughtfulness. There are enough examples of immoral behaviour by intelligent individuals (Gottlob Frege's brilliance did not lead him to deduce that racism is morally wrong) to doubt whether merely increasing the computational efficiency of the mind would lead to an improvement in, say, moral judgment. Regardless of these worries, the biggest concern about cognitive enhancement taken on its own as a means for moral enhancement lies in doubts about the efficacy of such moral higher-order judgments to alter the lower-order desires and give them volitional force, especially in the light of the doubts, already discussed, that have been raised by empirical research, such as is to be found in the work of Jonathan Haidt, concerning the ability of moral judgments to cause action.

---

this risk is a certainty, which it does not seem to me, moral enhancement imposed by a third party that does not decrease the personal autonomy of a subject is, under certain conditions, possible, which is all that my present argument requires.

The risk of such enhancement for personal autonomy lies in the fact that if the higher-order desires of the subject are incapable of altering the lower-order desires, this will decrease the cohesion between the two levels, and diminish the autonomy of the individual. If a person who frequently acts on racist motives and has incorporated these motives into a world-view undergoes cognitive enhancement that makes him realise the wrongness of his motives, but is unable to change his actions, this would lead to a state of self-abhorrence and dissonance between his higher- and lower-order attitudes. As this individual now absolutely understands that such actions can never be right, it would be impossible for him to justify or integrate them. Purely cognitive moral enhancement without some means of ensuring the effectiveness of the desires thus brought about thus faces a real risk that it may decrease, not increase, the personal autonomy of the subject.

Alteration of the higher-order desires through cognitive enhancement without a means of ensuring that the higher-order desires can effectively will the lower-order desires to be of the kind that they desire, may therefore be largely ineffectual in bringing about a greater prevalence of moral actions in the subject and to a decrease in the autonomy of the subject. I now turn to the last broad category of moral enhancement, namely the strengthening of the agent's will.

4.2.5 Alteration of the will

It may be that a subject already has certain moral higher-order or lower-order attitudes, but that these desires lack the volitional force to lead to action. A last category of moral enhancement would strengthen the will of the subject to make his higher-order attitudes effective in their attempt to alter his lower-order attitudes and/or to make the lower-order attitudes effective in leading to corresponding moral action. It may, for instance, be the case that the subject already desires to act morally on a more frequent basis, but is unable to make this desire effective.

The will is a mysterious concept and it is not clear whether it is an independent cognitive force that attaches itself to a given desire, an aspect of a desire (the classical view) or perhaps an actual attitude itself, such as a belief or a value that gives a desire the volitional force to move to action. Thus it may be that a specific lower-order desire to act in a given way only becomes effective if there is a corresponding lower-order value or higher-order belief that imbues it with such "volitional force". It is, for instance, a general trait of humans that they are far more inclined to act in ways that they expect will bring about pleasurable states than unpleasurable ones.

Now it may be that there is a biotechnical means of altering a specific desire so that it is effective in leading to action without the involvement of the agent. Such an action will not have been willed by the agent, however, and would be heteronomous in exactly the same way as the actions produced by the hypothetical device described in our discussion about the direct determination of actions. Whatever we have said about means of altering the actions of the subject while entirely circumventing the processes of the mind will also apply to cases where a means of moral enhancement gives volitional force to a desire independently from the attitudes in this regard of the subject himself. In order for the agent himself to will an action, there has to be some corresponding attitude on the part of the subject, otherwise these actions will be heteronomous.

To increase the will we may either increase the volitional force of specific kinds of mental events or attitudes or associate certain types of attitudes with other attitudes that already have such volitional force. An example of the first kind of "will-strengthening" would be somehow to increase the volitional power of moral judgments so that they are more effective in leading to action. An example of the second kind would be to "pair up" a certain kind of lower-or higher-order attitude with another mental state that already has very strong volitional force. Thus we might imagine a kind of hedonic moral enhancement that would make acting on the basis of moral judgment pleasurable for the agent

and immoral actions unpleasurable. By doing so we would effectively be punishing bad behaviour and rewarding good behaviour.

Something akin to this kind of moral enhancement is described in Anthony Burgess' seminal novel, *A Clockwork Orange* (1986), in which a violently asocial young man is conditioned by means of a form of aversion treatment known as "Ludovico's technique" to suffer extreme nausea whenever he desires to commit acts of violence. Such an approach would be consonant with certain theories of desire that posit pleasure as the (or a major) driving force that moves individuals to act.[8] This approach is also not dissimilar to certain conventional means of moral education, where an individual is initially driven to act in certain ways out of fear of punishment and expectation of reward, until the behaviour is habituated through repetition.

In the case of "Ludovico's technique," Alex, the protagonist of Burgess' novel, abstains from violent acts after enhancement not as a result of the withdrawal of his endorsement, based on critical reflection, of the desire to commit acts of violence but merely in order to avoid the unpleasant physiological consequences of such desires. Alex' desire to commit acts of violence is left intact; it is his ability to will those desires that is left diminished. A method such as Ludovico's technique in effect subverts the will of the subject. If autonomy, simply put, is the ability to have the will that you want to have, then any moral enhancement that alters the direction of the agent's will in a manner that is not endorsed on a higher level by the subject will effectively decrease his autonomy. Now it may be that the aversion to bad consequences eventually leads to the desires that the subject is compelled to act on to become habituated and for the subject to grow to accept these desires. Such

[8] If neural activity is somehow an indication of corresponding mental functioning, then tentative support for such a view of action is to be found in the work of Harbaugh et al. (2007), whose study has revealed neural activity indicative of hedonic stimulation when a subject makes charitable contributions. Whether the expectation of such hedonic stimulation is the cause of such contributions and is the only (or a major) contributing factor in motivation is far from certain, however.

acceptance will likely be of the fatalistic kind, rather than based on critical reflection[9] and undermine the agent's ability to alter his higher-order beliefs, desires or principles. Moreover, the trauma brought about by such negative physiological consequences will make the changes very difficult for the subject to integrate into his narrative identity pre- and post-enhancement.

4.3 Additional considerations

All other things being equal, moral enhancement will least controversially be able to safeguard the personal autonomy of the subject if the enhancement takes place in accordance with the higher-order desires of the subject prior to enhancement. We may also cautiously affirm the possibility that a subject could be autonomous even if the subject is enhanced as a result of the desires of another party, provided that (i) such enhancement is limited to the lower-order desires of the subject, (ii) the subject is able to endorse the new lower-order desires post-enhancement in a procedurally independent manner and (iii) the means by which they arose can be integrated by the subject into the diachronic narrative identity that straddles his life pre- and post-enhancement.

The outline given above has of necessity been schematic. The complexity of human psychology is rarely that simple, however. In the concluding section of this chapter I will therefore briefly touch on some additional considerations about moral enhancement that may well prove capable of affecting the personal autonomy of the enhanced subject.

---

[9]Such critical reflection that occurs in the wake of the forced deflection of the agent's will, will be based on the wrong reasons, namely the unpleasant consequences of immoral action for the self rather than on reasons about consequences for others. Much in traditional moral education operates in such a manner, but this does not mean that such education promotes the ability of an individual to reflect independently and authentically about the right things to desire.

4.3.1 Duration and reversibility

We have previously stated that an individual cannot be considered autonomous unless he is able to revise his higher-order attitudes post-enhancement. Given the complex way in which the various lower- and higher order attitudes interact, however, it may well be that the enhancement sets in motion a chain reaction that has unwanted long-terms effects on the ability of the individual to revise his higher-order desires. Some forms of enhancement may be more prone to this than risk than others. The risk may be mitigated, however, if the duration of such an intervention has a limited life span in the subject, such as in the case of a pharmaceutical intervention that is only effective for as long as the pharmaceutical medium is present within the body of the subject, or if the effects can be reversed with little negative consequence. Thus it is important to take into consideration that even if the intervention is reversible, the process of reversing the effects may be accompanied by mental or physical hardship or be potentially hazardous to the physical well-being of the subject, such as in the case of surgically induced alterations. A difficult, painful or potentially hazardous reversal may cause the subject to "bear with the consequences" even though they are not, or no longer, desirable.

Lastly, it should be kept in mind that even if the effects of the intervention are easily reversible, the memory of the event may continue to persist in the subject. If the effects of the intervention are severely demoralising or deleterious to the subject's sense of self-worth, for instance, the adverse effect on the psychological constitution of the subject, including such psychological traits that are beneficial to personal autonomy, may have a long or lasting influence on the subject. The actions of the post-enhancement subject will also continue to persist in the memories of other individuals whose opinions about the subject are important to him. If negative opinions persist about the subject in the minds of these individuals, this may affect his behaviour subsequent to a reversal of the physical

effects of the intervention in such a way as to exercise a limiting effect on his ability to act autonomously.

Irreversibility, prolonged, unpleasant or hazardous reversibility of the enhancement therefore have the potential to undermine the ability of the enhanced subject to alter his higher-order preferences. This may not be true for all cases of moral enhancement, however. If the effects of the enhancement can be pinpointed to a very large degree of certainty, and if the enhancement takes place as a result of the will of a subject who is thoroughly aware of the consequences, including the fact that there may be unforeseen consequences that he cannot predict from the vantage-point of his pre-enhanced self, then the consequences of the moral enhancement should be deemed to be in accordance with the will of the subject. We are often stuck in ordinary life with the far-reaching and lasting effects of decisions, even if those decisions were based on considered judgments.

4.3.2 The predictability of the effects

Given what we have stated above regarding possible limitations on the reversibility of a moral enhancement, the predictability of the effects of the moral enhancement is an important factor when assessing the influence of a moral enhancement on the personal autonomy of an individual. Without a high level of predictability, an agent may chose to undergo a specific kind of moral enhancement on the basis of his higher-order desires, but find that the intervention does not have the effect that he desired and be the less autonomous for it. This does not mean that the effects of a form of moral enhancement should of necessity be entirely or even to a large extent predictable, and it is likely that such predictability is in any case impossible to achieve. But the unpredictability itself is something that may be predictable and that the pre-enhanced subject should be made aware of and take fully into consideration before deciding to undergo such moral enhancement.

Moreover, where certain possible side effects are known to any parties involved in the administration of the intervention, these side effects would have to be made known to the

subject, otherwise he will not be able to make a considered decision regarding whether to undergo the enhancement or not. Withholding information or exaggerating certitude concerning the effects of the intervention by such parties will be tantamount to the deception of the subject.

4.3.3 The biotechnical factor

This is a thesis about moral improvement by biotechnical means. As previously indicated, I hold all moral enhancements to have one physical target, namely the brain of the individual, which is the physical region in the body that governs our mental activity, including the ability to make decisions. Biotechnical means of altering the brain are numerous, and might include genetic selection or modification, surgery, pharmaceutical interventions and deep brain stimulation (DBS).

All of the four means of intervention would alter the relative development of regions in the brain, the pathways between these regions and (possibly in some of the scenarios) the content that is held or represented in the neurons. While it may be technically possible to bring about changes in the brain that would be impossible to replicate through traditional means, it is possible to imagine means of moral enhancement that would merely duplicate, in the effects that they bring about, changes that could have been brought about through "conventional" means.

Thus it is commonplace in the general enhancement debate to note that non-enhanced human beings are also the result of gene selection; for when our parents chose each other, they did so, consciously or unconsciously, on the basis of certain preferences regarding what would constitute suitable genetic material for their offspring. It is also commonplace that everything that happens to and around us in our environment, including traditional forms of education, bring about physical changes in the brain, not only radically so during the first years of existence, but also in more modest fashions through the adaptation of neural pathways for frequently performed tasks, for instance, or the storage of memories.

If we were therefore to posit two identical human beings who undergo two different types of changes that result in both human beings having identically changed neural structure and content as a result of the changes, with the only difference being that the changes in the first case were effected through traditional exposure to the environment, whereas the change for the second human were caused through one of the biotechnical means listed above, what are the salient differences between the two scenarios, both from a general moral perspective and our specific area of enquiry, namely personal autonomy?

Both traditional means of moral education and biotechnical moral enhancement pose risks to the narrative identity of the subject as well as to his ability to form his higher-order attitudes and endorse or alter his lower-order attitudes in a procedurally independent manner. Both traditional moral improvement and biotechnical moral enhancement may have unpredictable results and prove, under certain conditions, to be impossible to reverse. Lastly, both traditional and biotechnical methods may circumvent the cognitive and affective mechanisms of the mind, leaving the subject unable to resist the alteration. The only reason why moral enhancement is proposed as an adjunct to traditional means of moral improvement, is that it may be more effective in reaching its goals, given the limitations, previously outlined, of the traditional means of improvement.

There is one possible difference with implications for the personal autonomy of the enhanced subject that I want to broach in conclusion, namely the concern that the very awareness of the artificiality of the moral enhancement will somehow alter his self-conception in such a way as to undermine his autonomy. This concern is expressed most cogently by the late German philosopher, Jürgen Habermas (Habermas 2005).

According to Habermas, the mere awareness that our nature has been shaped by "programmers" may end up making the enhanced person feel less free and less authentic, thereby instilling a kind of fatalism that ends up actually making us less free and less authentic. Habermas' argument is primarily centred on genetic engineering, but it could be applied to other forms of biotechnical interventions as well, as the spectre of scientists in

laboratory coats and flashing computer screens looming over the parental bed may indeed elicit such concerns on the part of an enhanced subject, especially if the enhancement occurred due to the will of parties other than the subject herself.

The Habermasian objection forms part of a prominent school of thought opposed to the enhancement project due to its "unnaturalness" and perceived attempt to dismantle human nature. Objections against enhancement from human nature and "naturalness" fall outside the scope of this paper, and I will only focus on possible implications for the autonomy of a subject as a result of the "artificial" nature of biotechnical means of enhancement.

Regardless whether the views on the sanctity of "human nature" of Habermas, as well as other notable proponents of this line of argumentation such as Michael Sandel (2009) and Leon Kass (2002) are correct or not, there is a real risk that if a person who holds such views discovers that he is the result of genetic selection for the purposes of moral enhancement or if he is subjected to such enhancement against his explicit will, the knowledge of this fact could have a detrimental effect on his self-conception and by extension his capacity for autonomy.

The concern seems to be less so for an individual who has willingly subjected himself to moral enhancement, given that such a decision would, if it were based on a whole-hearted, conscious and informed acceptance of the principle of biotechnical enhancement, be compatible with a particular view of human nature. It also seems to me as if this risk will decrease as our scientific understanding of the mind and its workings becomes "main-stream". Such knowledge may lead to greater understanding of the biological limits of our current morality and a growing acceptance of biotechnical means of improving that morality, effectively rendering the Habermasian objection increasingly obsolete.

This understanding may also have other implications for the moral enhancement project. It so happens that much in our gradual understanding of the human mind, how it functions and reaches decisions, is diametrically opposed to general common-sense intuitions. Many of these findings have not reached the "main stream" yet and are as a result

of their as yet predominantly theoretical status and counter-intuitive nature relatively easy to brush away by the non-specialist.

If it were to be conclusively established that the notion of a core "self" is a fiction, that much or most of our conscious experience is epiphenomenal at worst and at best very often a *post-hoc* justification of preconscious decisions, that "what we value" can often be equated to what satisfies the reward centres of the brain – what would the effect of such knowledge be on the autonomy of an individual? The technological advances on which the feasibility of moral enhancement relies, will presumably evolve together with our understanding of the workings of the human brain. This changing conception of our nature will eventually have to reach society and may have certain effects on society at large.

Thomas Metzinger expresses this concern in his book *The Ego Tunnel* (2009):

> One of the many dangers in this process is that if we remove the magic from our image of ourselves, we may also remove it from our image of others. We could become disenchanted with one another. Our image of Homo Sapiens underlies our every practice and culture; it shapes the way we treat one another as well as how we subjectively experience ourselves... Now that the neurosciences have irrevocably dissolved the Judaeo-Christian image of a human being as containing an immortal spark of the divine, we are beginning to realize that they have not substituted anything that could hold society together and provide a common ground for shared moral intuitions and values. An anthropological and ethical vacuum may well follow on the heels of neuroscientific findings." (Metzinger 2009)

Changing self-conceptions regarding human nature may therefore provide additional reasons for the necessity of moral enhancement.

Chapter 5

Conclusion

Moral enhancement need not lead to a diminishment of the personal autonomy of the subject of such an intervention provided it serves merely as a mechanism to help the subject overcome the deterministic limitations that prevent her from bringing her lower-order desires into conformity with the higher-order desires that she has arrived at through independent, thoughtful deliberation (to the extent that such deliberation is itself free of deterministic constraints). In such a case, the enhancement will actually serve to increase the personal autonomy of such a subject. It is important that the means by which the intervention is induced do not limit the subject from continuing to revise her higher-order desires and should preferably be reversible.

Whether such an intervention is technically feasible remains to be seen, but the conclusion entails that some forms of moral enhancement, at the very least, are compatible with personal autonomy and could even increase it.

We may also cautiously allow for the possibility that enhancement at the behest of another party may, under certain conditions, eventually lead to a restoration or even increase in the autonomy of the subject: such moral enhancement would need to have as its direct locus of intervention only the lower-order desires of the subject, lead to a predictable, procedurally independent endorsement of the new lower-order desires by the subject post-enhancement and be capable of coherent integration by the subject into the diachronic narrative identity that straddles his life pre- and post-enhancement. If these conditions are met, then such an enhancement cannot be deemed to have caused a decrease in the autonomy of the subject. Whether any imposed form of moral enhancement would

actually be capable of meeting these conditions is open to debate; and yet, many of the hopes for the improvement of human society that are placed on moral enhancement by its proponents would hinge on the moral permissibility of such involuntary imposition.

I opened the paper with an extended quotation from Bertrand Russell, in which he advocated the use of science to bolster our "kindly impulses." I did not at the time provide the reader with Russell's concluding thoughts in this respect, but will do so now. The scientists capable of creating such an intervention, Russell states,

> ...would first have to administer the love-philtre to themselves before they would undertake such a task. Otherwise, they would prefer to win titles and fortunes by injecting military ferocity into recruits. And so we come back to the old dilemma: only kindliness can save the world, and even if we knew how to produce kindliness we should not do so unless we were already kindly. (*ibid.*)

The very persons most in need of moral enhancement may therefore be the ones least likely to choose such enhancement for themselves and thus the ones most at risk of forsaking some autonomy if they were to be subjected to moral enhancement against their will.

Bibliography

Nicholas Agar. A Question about Defining Moral Bioenhancement. *Journal of Medical Ethics*, 2013. Published Online First: 25 January 2013. doi: 10.1136/medethics-2012-101153.

Aristotle. *Nicomachean Ethics*. Cambridge University Press. Transl. R. Crisp, Cambridge, 2004.

John A. Bargh and Tanya L. Chartrand. The Unbearable Automaticity of Being. *American Psychologist*, 54(7):462–479, 1999.

Simon Baron-Cohen. *The Science of Evil: On Empathy and the Origins of Cruelty*. Basic Books, New York, 2011.

Roy F. Baumeister, Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. Ego Depletion: Is the Active Self a Limited Resource? *Journal of Personality and Social Psychology*, 74(5):1252–65, 1998.

Bernard Berofsky. *Liberation from Self: A Theory of Personal Autonomy*. Cambridge University Press, New York and Cambridge, 1995.

Michael Bratman. *Structures of Agency*. Oxford University Press, Oxford, 2007.

Tommaso Bruni. The Ambivalence of Moral Psychology. *AJOB Neuroscience*, 2(4):13–15, 2011.

Anthony Burgess. *A Clockwork Orange*. W.W. Norton and Company, New York, 1986.

Sarah Chan and John Harris. Moral Enhancement and Pro-Social Behaviour. *Journal of Medical Ethics*, 37(3):130–1, 2011.

John Christman. Constructing the Inner Citadel: Recent Work on the Concept of Autonomy. *Ethics*, 99(1):109–124, 1988.

John Christman, editor. *The Inner Citadel: Essays on Individual Autonomy*. Cambridge University Press, Cambridge, 1989.

John Christman. Autonomy and Personal History. *Canadian Journal of Philosophy*, 21(1): 1–24, 1991.

John Christman, 2011. URL `http://plato.stanford.edu/archives/spr2011/entries/autonomy-moral/`. Accessed on 14 October 2012.

John Davenport. *Narrative Identity, Autonomy and Mortality: From Frankfurt and MacIntyre to Kierkegaard*. Routledge, New York, 2012.

Mark H. Davis, Carol Luce, and Stephen J. Kraus. The Heritability of Characteristics Associated with Dispositional Empathy. *Journal of Personality*, 62(3):369–91, 1994.

David DeGrazia. Enhancement Technologies and Human Identity. *The Journal of Medicine and Philosophy*, 30(3):261–83, 2005.

David DeGrazia. Moral Enhancement, Freedom, and What We (Should) Value in Moral Behaviour. *Journal of Medical Ethics*, 2013. First Published Online : 25 January 2013. doi:10.1136/medethics-2012-101157.

Daniel Dennett. The Self as a Center of Narrative Gravity. In F. Kessel, P. Cole, and D. Johnson, editors, *Self and Consciousness: Multiple Perspectives*. Erlbaum, Hillsdale, NJ, 1992.

Thomas Douglas. Moral Enhancement. *Journal of Applied Philosophy*, 25(3):228–245, 2008.

Thomas Douglas. Moral Enhancement Via Direct Emotion Modulation: a Reply To John Harris. *Bioethics*, 9702(3), 2011.

Gerald Dworkin. Autonomy and Behavior Control. *Hastings Center Report*, 6:23–28, 1976.

Gerald Dworkin. *The Theory and Practice of Autonomy*. Cambridge University Press, 1988.

Gerald Dworkin. The Concept of Autonomy. In *The Inner Citadel, op.cit.* 1989.

Laura Waddell Ekstrom. A Coherence Theory of Autonomy. *Philosophy and Phenomenological Research*, 53(3):599–616, 1993.

Joel Feinberg. *Harm to Self: The Moral Limits of the Criminal Law Volume 3*. Oxford University Press, New York, 1986.

John Martin Fischer and Mark Ravizza. *Responsibility and Control*. Cambridge University Press, Cambridge, 1998.

Harry G. Frankfurt. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68(1):5–20, 1971.

Angelo Franzini, Carlo Marras, and Paolo Ferroli. Stimulation of the Posterior Hypothalamus for Medically Intractable Impulsive and Violent Behavior. *Stereotactic and Functional Neurosurgery*, 83:63–66, 2005.

Jürgen Habermas. *The Future of Human Nature*. Taylor & Francis, Malden, MA, 2005.

Jonathan Haidt. *The Righteous Mind*. Random House, New York, 2012.

William T Harbaugh, Ulrich Mayr, and Daniel R Burghart. Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations. *Science*, 316(5831): 1622–1625, 2007.

John Harris. Moral Progress and Moral Enhancement. *Bioethics*. First Published Online: 19 June 2012. doi: 10.1111/j.1467-8519.2012.01965.x.

John Harris. Moral Enhancement and Freedom. *Bioethics*, 25(2):102–111, 2011.

John Harris. 'Ethics Is for Bad Guys!' Putting the 'Moral' Into Moral Enhancement. *Bioethics*, 2012. First Published Online: 2 February 2012. doi:10.1111/j.1467-8519.2011.01946.x.

David Hume. *A Treatise of Human Nature*. Oxford University Press, Oxford, 2000.

Fabrice Jotterand. 'Virtue Engineering'? and Moral Agency: Will Post-Humans Still Need the Virtues? *AJOB Neuroscience*, 2(4):3–9, 2011.

E.T. Juengst. What Does Enhancement Mean? In Erik Parens, editor, *Enhancing Human Traits: Ethical and Social Implications*. Georgetown University Press, Georgetown, 1998.

D Kahneman. Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice. In T Frangsmyr, editor, *Nobel Prize Lecture*, volume 8 of *Prize Lectures*, pages 449–489. Citeseer, 2002.

Leon R. Kass. Ethical Challenges from Biotechnology. In *Life, Liberty and the Defense of Dignity: The Challenge for Bioethics*. Encounter Books, San Francisco, 2002.

Ariel Knafo, Carolyn Zahn-Waxler, Carol Van Hulle, JoAnn L Robinson, and Soo Hyun Rhee. The Developmental Origins of a Disposition Toward Empathy: Genetic and Environmental Contributions. *Emotion (Washington, D.C.)*, 8(6):737–52, 2008.

Neil Levy. Autonomy and Addiction. *Canadian Journal of Philosophy*, 36(3):427–448, 2006.

Alfred R. Mele. *Autonomous Agents: From Self-Control to Autonomy*. Oxford University Press, New York and Oxford, 2001.

Thomas Metzinger. *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York, 2009.

Terrie E Moffitt. Genetic and Environmental Influences on Antisocial Behaviors: Evidence from Behavioral-Genetic Research. *Advances in Genetics*, 55(05):41–104, 2005.

Jorge Moll, Ricardo De Oliveira-Souza, and Roland Zahn. The Neural Basis of Moral Cognition. *Annals Of The New York Academy Of Sciences*, 1124(1):161–180, 2008.

Anna Pacholczyk. Moral Enhancement: What Is It and Do We Want It? *Law, Innovation and Technology*, 3(2):251–277, 2011.

Ingmar Persson and Julian Savulescu. The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity. *Journal of Applied Philosophy*, 25(3):162–177, 2008.

Ingmar Persson and Julian Savulescu. Getting Moral Enhancement Right: the Desirability of Moral Bioenhancement. *Bioethics*, 9702:16–17, 2011.

Ingmar Persson and Julian Savulescu. *Unfit for the Future: The Need for Moral Enhancement*. Oxford University Press, Oxford, 2012.

Bertrand Russell. *Icarus: Or, the Future of Science*. E.P. Dutton & Company, New York, 1924.

Michael Sandel. *The Case against Perfection: Ethics in the Age of Genetic Engineering*. Belknap Press of Harvard University Press, Harvard, MA, 2009.

Eric Schwitzgebel and Joshua Rust. The Self-Reported Moral Behavior of Ethics Professors. http://www.faculty.ucr.edu/ eschwitz/SchwitzPapers/EthSelfRep-110316.pdf. Accessed 17 June 2012, 2011.

John R. Shook. Neuroethics and the Possible Types of Moral Enhancement. *AJOB Neuroscience*, 3(4):3–14, 2012.

Sylvia Terbeck, Guy Kahane, Sarah McTavish, Julian Savulescu, Philip J Cowen, and Miles Hewstone. Propranolol Reduces Implicit Negative Racial Bias. *Psychopharmacology*, 222(3):419–24, 2012.

Jeremy Waldron. Moral Autonomy and Personal Autonomy. In John Christman and Joel Anderson, editors, *Autonomy and the Challenges to Liberalism: New Essays*, chapter 13, pages 307–329. Cambridge University Press, Cambridge, 2005.

Björn Wallace, David Cesarini, Paul Lichtenstein, and Magnus Johannesson. Heritability of Ultimatum Game Responder Behavior. *Proceedings of the National Academy of Sciences*, 104(40):15631–15634, 2007.

Timothy D Wilson. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. The Belknap Press of Harvard University Press, Cambridge, Massachusetts, and London, England, 2004.

Susan Wolf. Moral Saints. *The Journal of Philosophy*, 79(8):419–439, 1982.

Chris Zarpentine. 'The Thorny and Arduous Path of Moral Progress': Moral Psychology and Moral Enhancement. *Neuroethics*, pages 1–13, September 2012. First Published Online: 30 September 2012. doi: 10.1007/s12152-012-9166-4.