

Autonomy, Rationality, and Contemporary Bioethics



Jonathan David Pugh

St. Anne's College

A Thesis Submitted for Examination for the Degree of DPhil in
Philosophy

2014

ABSTRACT

Personal autonomy is often lauded as a key value in contemporary bioethics. In this thesis, I aim to provide a rationalist account of personal autonomy that avoids the philosophical flaws present in theories of autonomy that are often invoked in bioethics, and that can be usefully applied to contemporary bioethical issues. I claim that we can understand the concept of autonomy to incorporate two dimensions, which I term the ‘reflective’ and ‘practical’ dimensions of autonomy. I suggest that the reflective dimension pertains to the critical reflection that agents must carry out on their motivating desires, in order to be autonomous with respect to them. I begin by rejecting prominent desire-based and historical accounts of this dimension of autonomy, before going on to defend an account based upon a Parfitian analysis of rational desires. Following this analysis of the reflective dimension of autonomy, I argue that autonomy can also be understood to incorporate a practical dimension, pertaining to the agent’s ability to act effectively in pursuit of their ends. I claim that recognising this dimension of autonomy more comprehensively reflects the way in which we use the concept of autonomy in bioethics, and makes salient the fact that agents carry out their rational deliberations in the light of their beliefs about what they are able to do. I go on to argue that this latter point means that my account of autonomy can offer a deeper explanation of why coercion undermines autonomy than other prominent accounts. Having considered the prudential value of autonomy in the light of this theoretical analysis, in the latter half of the thesis I apply my rationalist account of autonomy to a number of contemporary bioethical issues, including the use of human enhancement technologies, the nature of informed consent, and the doctor-patient relationship.

[Word Count – 74’981]

ACKNOWLEDGEMENTS

I would like to begin by thanking the Wellcome Trust for funding the research that went into this thesis. Furthermore, I would also like to thank St. Anne's College for their financial support throughout my time at Oxford; as well as providing me with a generous scholarship, travel grants from the college also enabled me present my research at a number of conferences. I am also grateful to the philosophy department for also providing funds to finance these trips.

Whilst I would not have been able to embark on this project without the generous financial support detailed above, money will only get you so far. My work on this project would not have been possible if I had not been lucky enough to receive the support of a large number of people.

First, I would like to express my deep appreciation for the time and effort of my supervisors Prof. Julian Savulescu and Prof. Roger Crisp. Prof. Savulescu was instrumental in persuading me to embrace a new topic in my doctoral research when I first moved to Oxford. His belief in my abilities at this early stage gave me the confidence to stretch myself in my research, and this has undoubtedly enabled me to become a far better philosopher. He also provided me with invaluable advice on numerous written drafts of this thesis (and other work), as well as providing me with numerous professional opportunities.

Prof. Crisp has also been extremely generous with his time and expertise over the course of my research. Cumulatively, he has probably read this thesis ten time over, given the number of chapter drafts that he has provided comments upon. I thoroughly enjoyed our discussions, and I have learnt a huge amount from them; I often left his office trying to get my head around a devastating criticism of a claim that I had naively assumed to be self-evident! I shall also remember Prof. Crisp's kind words of encouragement when I was trying to publish my first academic papers. All in all, there is little else that a student could hope for from their two supervisors. I am sincerely grateful to them both.

I would also like to thank the friends who have supported me throughout this project. In particular, Conor, Tomas, Ian, Row, and John have all helped me to manage the stress of doctoral research, and I have enjoyed some great evenings with them. As a fellow philosopher, I would particularly like to thank John for the many informative discussions/diatribes we shared, and his insightful advice on Michael Sandel's views.

I reserve my deepest gratitude for my family. They have been a source of love, strength and support throughout this project. As former medics, my mum and dad have undoubtedly had a huge influence on my interest in applied ethics. Not only that, but they have taken such interest in my work that they now often pose far more difficult questions to me than many philosophers! This is also true of my brother Chris. As a fellow academic, albeit in a different discipline, he has been a huge source of inspiration to me, and I doubt that I would have had the courage to embark on a doctoral thesis in the absence of his fine example; he is, as always, a trail-blazer. He has also been instrumental in keeping my philosophy rooted in the 'real world' and as far as possible from the ivory tower!

Finally, I would like to thank Hannah. For more than I can adequately express here.

Table of Contents

Introduction	1
I Introducing Autonomy	2
II The Reflective Dimension of Autonomy	5
III The Practical Dimension of Autonomy	15
IV Local and Global Autonomy	16
Conclusion	20
 Chapter One – Two Prominent Views Concerning the Reflective Dimension of Autonomy.....	22
I Hierarchical Theories: Real Selves, Authenticity and Autonomy	22
II Two Objections to Hierarchical Theories	28
III Internalist Responses	31
IV Externalist Responses.....	40
Conclusion	48
 Chapter Two – A Rationalist Account of Reflective Autonomy	50
I Rationalist Understandings of Reflective Autonomy	51
II Objections to Ekstrom’s Theory	56
III Parfit on the Rationality of Desires.....	62
IV Incorporating Parfit’s Account into A Rationalist Theory of Autonomy	69
V Responding to the Problem of Manipulation.....	75
Conclusion	80
 Chapter Three – The Practical Dimension of Autonomy.....	82
I Introducing The Practical Dimension of Autonomy	83
II Positive and Negative Freedom.....	86
III Autonomy and Freedom at the Point of Action	89
IV True Beliefs and Autonomy	98
V Freedom at the Point of Decision.....	103
Conclusion	109

Chapter 4 – Coercion and Autonomy	110
I Introducing Coercion	111
II Why Does Coercion Undermine Autonomy?	115
III Coercive Offers?	125
Conclusion	135
 Chapter Five - On the Prudential Value of Autonomy	 138
I The Nature of Autonomy’s Prudential Value	139
II Valdman’s Objection	147
III The Value of Different Sorts of Autonomy	152
IV Autonomy and Conflicting Values in Bioethics	157
Conclusion	164
 Chapter Six - Enhancing Autonomy	 166
I Enhancing Autonomy– Theoretical Considerations	166
II Enhancing Autonomy With Biotechnologies.....	174
III Indirectly Enhancing Autonomy?	184
IV Counter-productive Enhancements, and the Voluntariness of Choice.....	187
Conclusion	196
 Chapter Seven - Informed Consent and Autonomy Part One: Voluntariness.....	 197
I Introducing Informed Consent and its Justification	198
II Rationality and Undue Influence	207
III The Shared Decision Making Model and Liberal Rationalism	215
Conclusion	224
 Chapter Eight- Informed Consent and Autonomy Part Two: Disclosure, Understanding and Competence	 226
I The Information Element of Informed Consent - What Should Be Disclosed?	226
II Material Information	231
III (Rational) Competence.....	241
 Conclusion	 257
 References	 264

Introduction

Personal autonomy is often lauded as a key value in contemporary bioethics.¹ Indeed, on their widely endorsed ‘four principles’ approach to biomedical ethics, Beauchamp and Childress propose that the principle of respect for autonomy is one of four fundamental ethical principles of biomedical ethics (alongside the principle of beneficence, the principle of non-maleficence, and the principle of justice).² The concept of autonomy is also commonly understood to undergird the doctrine of informed consent, a doctrine that is invoked ubiquitously in contemporary bioethics.

However, autonomy is an ambiguous concept that has lent itself to a plethora of different uses in moral philosophy.³ Moreover, the ambiguity of the concept has led contemporary bioethicists to reach divergent conclusions about bioethical issues in which autonomy related concerns are salient. In this thesis, I aim to develop an account of personal autonomy that avoids the philosophical flaws in the theories of autonomy that are often invoked in bioethics, and that I shall apply to contemporary bioethical issues. In this introductory chapter, I shall make some preliminary remarks about the nature of autonomy broadly construed, and delineate what has been termed the ‘standard view’ of autonomy in the bioethical literature. I shall conclude by explaining the framework that I shall adopt in developing my own account of personal autonomy.

¹ For example, see Gillon, “Ethics Needs Principles—Four Can Encompass the Rest—and Respect for Autonomy Should Be ‘First among Equals’”; Beauchamp and Childress, *Principles of Biomedical Ethics*. Smith, “The Pre-Eminence of Autonomy in Bioethics.”

² Beauchamp and Childress, *Principles of Biomedical Ethics*.

³ See Arpaly, *Unprincipled Virtue*, 118–125 and Dworkin, *The Theory and Practice of Autonomy*, 3–6 for surveys of the different understandings of autonomy in the philosophical literature.

I Introducing Autonomy

The term ‘autonomy’ is derived from the Greek ‘autos’ (self), and ‘nomos’ (law); accordingly, the concept that the term ‘autonomy’ aims to capture seems to be, broadly speaking, the property of self-government.⁴ Accordingly, as a preliminary observation, we might say that in investigating the nature of autonomy, we are investigating what it is for an agent to be self-governing.

Even this formulation might be understood as making an important presumption, since it presumes that autonomy is a property of *agents*. Dworkin points out that this is one of the few claims that autonomy theorists agree upon.⁵ However, in their discussion of autonomy in a biomedical context, Beauchamp and Childress explicitly reject this assumption, and instead claim that autonomy is a property of choices rather than agents.⁶ I shall claim below that Beauchamp and Childress’ rejection of the claim that autonomy is a property of agents’ results from a failure to distinguish autonomy in a *local* sense, and autonomy in a *global* sense. As such, for the purposes of this preliminary discussion, I shall assume that autonomy is a property of agents.

What then is it for an agent to be self-governing? Kant famously claimed that in order to be autonomous, an agent must be governed by her *noumenal* self, that is, the self as it is conceived as a member of the transcendent realm of pure reason, and not the self as a member of the phenomenal realm, in which it is subjected to external causes according to Kant’s dualist metaphysics. It is worth noting three features of the Kantian

⁴ Dworkin, *The Theory and Practice of Autonomy*, 12.

⁵ *Ibid.*, 6.

⁶ Beauchamp and Childress, *Principles of Biomedical Ethics*, 58.

account, as it is commonly understood.⁷ First, on Kant's view, the autonomous agent is not moved to act by their desires; on the contrary, this would be the paradigm of heteronomy on the Kantian account, since desires represent contingent external causes on the will in Kant's metaphysics.⁸ Second, autonomy is an inherently *moral* concept for Kant, since on his view pure reason demands that agents act in accordance with the Categorical Imperative. Third, autonomy is a property that undergirds the unique value of human life on the Kantian view; as autonomous agents, humans are understood to have a dignity beyond price.

I mention the Kantian account here only to set it aside. Although certain bioethicists appeal to broadly Kantian accounts of autonomy,⁹ the conceptions of autonomy that many bioethicists invoke in their discussions are decidedly un-Kantian. As Taylor points out, *pace* Kant, many contemporary theorists understand an agent to be autonomous if they direct their decisions in the light of their own desires and values, and without the controlling influence of others;¹⁰ notice that on this understanding, an autonomous agent's desires and values can have non-moral content. As I shall explain in more detail below, in this thesis I shall be interested in this latter sort of understanding of autonomy.

The second feature of the Kantian approach delineated above indicates that Kant's is a *substantive* account of autonomy, in so far as it stipulates that the choices of autonomous agents must have certain (and on Kant's account, *moral*) content. We may

⁷ These are at least features of Kant's account on orthodox understandings of his view; however, it should be acknowledged that some Kantian scholars might contest the sketch that I have given here. For instance, see Herman's analysis in Herman, *The Practice of Moral Judgment*.

⁸ See Hill, *Autonomy and Self-Respect*, 30.

⁹ O'Neill, *Autonomy and Trust in Bioethics*; Velleman, "A Right of Self-Termination?"

¹⁰ Taylor, *Practical Autonomy and Bioethics*, xiii

contrast substantive accounts of autonomy with *procedural* accounts; according to procedural accounts, an agent is autonomous if they direct their decisions in accordance with a certain procedure (such as the one that Taylor describes above). The precise details of the procedure in question will differ from theory to theory; however, the key point is that procedural theories do not claim that the autonomous agent's choices must have a particular *content*.

In this thesis, I shall be concerned primarily with procedural theories of autonomy. There has admittedly been a revived interest in substantive theories of autonomy in recent years. For instance, feminist philosophers have argued that procedural accounts are inadequate because agents may make their choices in accordance with the requirements of a procedural theory of autonomy, and yet (it might be claimed) still lack autonomy because they make their choices in accordance with values that are determined by patriarchal norms.¹¹ Some theorists have responded to this sort of problematic case by endorsing a substantive account that stipulates that there are normative restrictions upon what autonomous agents can desire; for instance, such theories might claim that an autonomous agent cannot choose a life of servitude.¹²

Despite this revived interest in substantive theories, I shall not directly consider them in this thesis. In order to at least partly justify this omission, it is prudent to highlight the main problem with these theories. It seems that the main problem with substantive accounts of autonomy is that, on such accounts, one cannot be autonomous with respect to those of one's choices that fail to comply with certain externally imposed norms. Yet, at a pre-theoretical level, we might be attracted to the claim that

¹¹ Westlund, "Selflessness and Responsibility for Self." See also Oshana, "Personal Autonomy and Society."

¹² For example, see Benson, "Freedom and Value."

autonomy should be a matter of acting in accordance with one's *own* values, rather than with externally imposed norms.¹³

Other philosophers have responded to the sort of hard case delineated above by suggesting that this sort of case indicates that autonomy is a 'relational' phenomenon, and that the values that shape our decision-making are in large a part a product of our socialization.¹⁴ Although I shall focus primarily on procedural theories of autonomy, it should be acknowledged that the theory that I shall endorse is compatible with a relational view of the autonomous agent. I shall say more about this in chapter two. In the next section, I shall consider what an adequate procedural theory of autonomy should aim to achieve, and suggest that procedural theories pertain to one of two dimensions of autonomy.

II The Reflective Dimension of Autonomy

I lack the space here to survey the many ways in which philosophers have understood the concept of autonomy.¹⁵ Moreover, given the diverse array of approaches to the concept, it seems unlikely that we will be able to capture the essence of autonomy by attempting to unite all the disparate accounts into a unified theory. Rather, as Levy

¹³ Frankfurt makes a similar objection in his comments on Kantian autonomy in Frankfurt, *Necessity, Volition, and Love*, 130–135. See also Haworth, *Autonomy*, 156–157 and Noggle, "Autonomy and the Paradox of Self-Creation," 96.

¹⁴ Noggle stresses this point in Noggle, "Autonomy and the Paradox of Self-Creation," 104. See also Mackenzie and Stoljar, *Relational Autonomy*.

¹⁵ See note 3 for references to work that surveys some of the prominent understandings of autonomy in the literature.

points out, it seems that in attempting to provide an adequate theory of autonomy we must “restrict the range of meanings that we attribute to the word”.¹⁶ Accordingly, in this thesis, I shall understand the concept of autonomy to denote a particular capacity to which we seem to attribute prudential value in bioethical contexts, namely, a person’s capacity to both:

- (i) Make decisions about what to do in accordance with their own desires and values
- And
- (ii) To act on the basis of those decisions.¹⁷

The second clause of this definition means that autonomy on the understanding that I shall employ is an inherently practical concept. I shall explain this position in the next section, and in greater detail in chapter three.

In view of the first clause of my definition, a theory of autonomy must be able to explain what it is for an agent to make decisions in accordance with their ‘own’ desires and values. We might also add that an adequate account ought to reflect at least some of our pre-theoretical intuitions about which agents are autonomous. Of course, it would be a mistake to claim that an adequate theory of autonomy should be able to justify *all* of our pre-theoretical intuitions about which agents might appropriately be deemed to be autonomous in bioethical contexts; after all, it may be possible to debunk some of

¹⁶ Levy, “Autonomy and Addiction”, 429.

¹⁷ For a similar understanding, see Brock, *Life and Death*, 28.

our intuitions in this context. However, it seems plausible to claim that we should aim for a reflective equilibrium between theory and our robust intuitions in our thinking about autonomy.¹⁸

According to what I shall follow Walker¹⁹ in calling the ‘standard view’ of autonomy in bioethics, an agent is autonomous with respect a particular decision if it is made:

(1) Intentionally,

(2) With understanding,

And

(3) Without controlling influences that determine their action.²⁰

I shall analyse this account in greater detail in chapters seven and eight. Here, I wish to draw attention to two inadequacies in the standard view.

The first is that, whilst we may agree that being subject to controlling influences that determine action can undermine our autonomy, this view fails to explain *why* this is the case. This is problematic because the mere fact that an influence can be understood to determine action is not sufficient to establishing that the influence in question undermines autonomy. To claim otherwise would be to beg the question against

¹⁸ See also Dworkin, *The Theory and Practice of Autonomy*, 9.

¹⁹ Walker, “Respect for Rational Autonomy,” 340.

²⁰ Beauchamp and Childress, *Principles of Biomedical Ethics*, 59. See also Faden and Beauchamp, *A History and Theory of Informed Consent*, Chapter Seven.

compatibilist views of autonomy of the sort that I shall consider in the first two chapters of this thesis. On these compatibilist theories, autonomy is understood to be compatible with the truth of causal determinism; on these views then, not all forms of determining influence are understood to undermine autonomy. Those who defend the standard view simply stipulate that coercion, non-rational persuasion and manipulation can all void acts of autonomy.²¹ Whilst this stipulation may be correct, it seems that an adequate theory of autonomy should be able to explain how and why these forms of influence undermine autonomy; listing examples of controlling influences is not satisfactory.

The second problem is that some agents who do not make their decisions in accordance with their own desires and values can still be autonomous on the standard view.²² Whilst the standard view is correct to stipulate that the controlling influences of other agents can undermine autonomy, it fails to give an adequate account of the *internal* impediments that agents can face to making their decisions in accordance with their own desires and values. Admittedly, in their defence of the standard view, Beauchamp and Childress mention that “. . . conditions such as debilitating disease, psychiatric disorders, and drug addiction”²³ can also undermine autonomy; however, in the absence of an account of why these things can undermine autonomy, this stipulation seems *ad hoc*. Moreover, as I shall explain in chapters seven and eight, defenders of the standard view explicitly reject conditions for autonomy that could provide such an explanation, on the basis of an objection that is far from convincing. To close this section I shall illustrate two cases in which agents seem to face internal impediments to making decisions in the light of their own desires and values.

²¹ Beauchamp and Childress, *Principles of Biomedical Ethics*, 94–95.

²² For a similar objection, see Kihlbom, “Autonomy and Negatively Informed Consent,” 147; Walker, “Respect for Rational Autonomy,” 348–351.

²³ Beauchamp and Childress, *Principles of Biomedical Ethics*, 93.

Even at a pre-theoretical level, it seems clear that being autonomous cannot always simply be a matter of ‘doing what one wants to do’. As Frankfurt points out, this will often not be sufficient for autonomous agency, since one’s motivating desire might be an imposter on one’s will. To illustrate, consider the following example:

Jane suffers from bulimia. She is aware that she is dangerously underweight and wants to return to a healthy weight. However, after every meal, she makes herself sick, even though she knows that this is damaging her health. Yet Jane feels alienated from her action whenever she does so; she believes that it is not a reflection of what she really wants.

It seems that Jane is not self-governing here, since she is moved to act by a desire that she feels alienated from; we might say that her motivating desire is ‘inauthentic’ in some sense. Although I use the example of bulimia to illustrate an inauthentic desire, there are various conditions that could cause an agent to be alienated from their desires in this manner. For instance, Frankfurt’s case of an “unwilling addict” is analogous to the example of Jane;²⁴ furthermore, some (although not all)²⁵ sufferers of clinical depression and anorexia nervosa could be understood as being motivated by a desire that they feel alienated from when they engage in self-harming behaviour.

It should be acknowledged here that the philosophical literature on authenticity and autonomy often assumes that the motivational states whose authenticity is in

²⁴ Frankfurt, “Freedom of the Will and the Concept of a Person,” 12.

²⁵ In complex cases, sufferers of these conditions may endorse their self-harming desires. I shall consider such cases in chapter eight.

question in cases such as Jane's are *desires*. Even Gary Watson, who claims to endorse a reason-based, non-Humean view of free action, suggests that autonomous agents act on the basis of *desires* that cohere with their rational evaluative judgements.²⁶ Accordingly, although Watson aligns his position with that of Plato rather than Hume, it seems that certain Humean assumptions concerning motivation are built into much of the discussion of authenticity and its relationship to autonomy. I shall not question those assumptions here, but they should be acknowledged at the outset.

Examples of alienation suggest that an adequate procedural account of autonomy in bioethics should be able to explain why agents such as Jane lack autonomy. However, in some case agents might lack autonomy even if they are acting in accordance with a motivating desire that they do not feel alienated from; they will lack autonomy if they have been *manipulated* to have that desire.

The standard view is able to explain why agents can lack autonomy if they have been externally manipulated to be motivated to act in some way. Consider the following case:

Manuel is a law-abiding citizen. However, someone covertly hypnotises Manuel to develop a motivating desire to rob a bank. Although Manuel does not form this desire as a result of any sort of deliberation, he does not feel alienated from it.

Again, at a pre-theoretical level, it seems that Manuel is not self-governing. *Prima facie*, at least, the problem here seems to be that an external process overrides Manuel's

²⁶ Watson, "Free Agency," 207.

own input into his motivating desire. In cases such as Manuel's, where the process that overrides the agent's own input into the development of their motivating desire is a result of third party interference, this process can be understood as being manipulative, in so far as the term 'manipulation' connotes a degree of agency on the third party's part.

Manuel lacks autonomy on the standard view due to condition (3). However, contrary to the impression that the standard view might leave us with, the fact that another agent has controlled Manuel does not fully account for what we find to be problematic in the example. To illustrate, compare Manuel's case with that of Phineas Gage. Gage was an upstanding member of society who was the victim of an accident in which a metal rod was blasted through his skull. Although he survived, his character completely changed as a result of the neurological damage that he sustained in the accident. Like Manuel, Phineas Gage lacked any input into the way he came to change his desires and values. However, the change in Gage's desires and values was not the result of third party interference; as such, it is inaccurate to claim that he was *manipulated*, in so far as this term connotes a degree of agency on the third party's part.

I shall explain the senses in which both Gage and Manuel can be understood to lack autonomy in more detail in chapter two.²⁷ For now though, this example points to an important distinction between two ways in which an agent may lack autonomy. When an agent lacks autonomy with respect to their motivating desires because they have been subjected to the controlling influence of another agent, we may say that such

²⁷ I shall also explain how such agents differ from normal agents who develop their desires and values in the light of social determining influences, but who should surely count as autonomous on a plausible compatibilist theory.

agents are *heteronomous*, in the sense that another agent rules them.²⁸ In contrast, when an agent lacks autonomy with respect to their motivating desire, but has not been subjected to the controlling influence of another agent, we may say that they are *non-autonomous*, but not heteronomous.

This is an important distinction to acknowledge in certain problematic cases in contemporary bioethics in which considerations of autonomy are salient. The reason for this is that we may want to at least question the autonomy of certain agents who are not subject to external agential control. For example, it seems that Jane, the sufferer of bulimia in my example above lacks autonomy in some sense, despite the fact that she is not subjected to external agential control. Furthermore, it seems that some sufferers of anorexia nervosa and clinical depression may be said to lack input into their formation of their motivating desires in a fashion that is similar to the process that overrides Gage's input into his motivating desires. In order to allow for the possibility that such agents might lack autonomy, it seems that a plausible theory of autonomy in bioethics should allow that both *heteronomy* and *non-autonomy* are failures of autonomy. Whilst the standard view purports to offer an account of the former, it does not offer an account of the latter.

We may also note that there is an important difference between the cases of Jane and Manuel. Jane's lack of autonomy seems to stem from the fact that she is motivated by a desire that she *herself* identifies as being inauthentic; she rejects her desire to make herself sick, even while she succumbs to the force of it. In contrast, if Manuel's desire is inauthentic, it is not so in the same way that Jane's is; *ex hypothesi*, Manuel accepts and

²⁸ This definition is partly stipulative. Note that heteronomy can be understood in a broader, Kantian, sense to mean 'ruled by *something* else'.

endorses his desire to rob the bank, and does not feel alienated from it. However, it is somewhat jarring to claim that Manuel's desire to rob the bank is authentic, that it is really 'his own'; this suggests that there may be different ways of understanding what it is for a desire to be inauthentic.

I shall explore these different understandings of authenticity in the next chapter. For the purposes of this introductory section, it is sufficient to acknowledge that the cases of manipulation and self-alienation both suggest that in order for an agent to be autonomous, they must bear a certain sort of relation to the motivational states that give rise to their actions. As I shall explain in the following two chapters, procedural theorists tend to cash this out by claiming that agents are only autonomous with respect to their motivating desires if they carry out some sort of *reflection* on these desires. In carrying out such reflection on one's motivating desires, it is believed that agents can have a greater degree of assurance that those desires are in some way 'their own', and not merely the outcome of determining forces of the sort that serve to undermine autonomy.

I propose that the above discussion suggests that an adequate theory of autonomy will incorporate what we may term a *reflective* dimension, so-called, because it pertains to the sort of reflection that autonomous agents must carry out on their motivating desires. This reflective dimension of autonomy can be understood as the *explanandum* of procedural theories of autonomy. Henceforth, when I intend to refer to agents who meet the conditions of a procedural theory of autonomy, I shall say that such agents are 'reflectively autonomous' on that theory.

It should be acknowledged that many of the questions that I shall consider in my investigation of the reflective dimension of autonomy have also been understood as

pertaining to the concept of moral responsibility, rather than autonomy. This is a reflection of the fact that these two concepts have often been conflated in the philosophical literature.²⁹ I lack the space here to consider the extent to which these two concepts differ. However, it is prudent to warn the reader against extrapolating the arguments that I shall make regarding autonomy to the concept of moral responsibility.³⁰ Where possible, I shall restrict my discussion of the reflective dimension of autonomy to works that ostensibly discuss autonomy as opposed to moral responsibility.

To close this introductory discussion of the reflective dimension of autonomy, it should be acknowledged that bioethicists who reject the standard view of autonomy in bioethics have appealed (either implicitly or explicitly) to a diverse range of theories of the reflective dimension of autonomy, often without acknowledging important philosophical objections to these theories.³¹ Moreover, the standard view itself eschews reference to what I have termed the reflective dimension of autonomy (for reasons I shall consider in chapters seven and eight). Accordingly, it seems crucial that an adequate theory of autonomy for use in contemporary bioethics should be informed by a philosophical discussion of the reflective dimension of autonomy.

²⁹ See Fischer, "Recent Work on Moral Responsibility," 98 for discussion of this point. For attempts to differentiate the two concepts, see Oshana, "The Misguided Marriage of Responsibility and Autonomy"; Mckenna, "The Relationship between Autonomous and Morally Responsible Agency."

³⁰ That said, the rationalist view of autonomy that I shall defend bears some broad similarities to the reasons-responsiveness approach to moral responsibility. See Fischer and Ravizza, *Responsibility and Control* for a prominent example of this approach.

³¹ For a limited sample, Doorn, "Mental Competence or Capacity to Form a Will" endorses a Frankfurtian hierarchical approach; Juth, "Enhancement, Autonomy, and Authenticity" endorses a historical approach; DeGrazia, *Human Identity and Bioethics*, 95–106 endorses a hybrid of these two approaches. Kihlbom, "Autonomy and Negatively Informed Consent," 147, endorses a coherentist approach. Walker, "Respect for Rational Autonomy" endorses a rationalist account.

III The Practical Dimension of Autonomy

Many philosophers purport to provide a comprehensive analysis of autonomy by giving an account of only the reflective dimension of autonomy.³² However, meeting conditions pertaining to the reflective dimension of autonomy is not sufficient for autonomy *in toto* on the understanding of autonomy that I am invoking here. Autonomy, on this understanding, involves not only being able to make decisions on the basis of one's own desires and values, but also being able to *act* in accordance with those decisions.

This sort of understanding of autonomy is implicit in the bioethical application of the principle of respect for autonomy. As Beauchamp and Childress point out, the principle of respect for autonomy incorporates a positive obligation that enjoins us to facilitate an agent's ability to make an autonomous decision; however, it also incorporates a negative obligation not to restrain the autonomous actions of others.³³ For instance, the principle might enjoin us to respect a patient's decision to refuse a treatment that is necessary for saving her life. In view of this negative obligation, we can be accused of undermining another agent's overall autonomy if we obstruct their pursuit of an end that they have chosen to pursue in a reflectively autonomous sense. Accordingly, this negative obligation implies that autonomy can be understood as having a *practical* dimension, pertaining to the agent's ability to act effectively in pursuit of their ends.

³² Again, for a limited sample, see Dworkin, *The Theory and Practice of Autonomy*; Ekstrom, "A Coherence Theory of Autonomy"; Christman, "Autonomy and Personal History." See Oshana, "Personal Autonomy and Society," 83–86, for an analysis of this tendency in the philosophical literature.

³³ Beauchamp and Childress, *Principles of Biomedical Ethics*, 64.

I shall defend this view in chapter three. However, I introduce the practical dimension here because I shall use the distinction between the reflective and practical dimensions of autonomy to frame my theoretical discussion of the nature of autonomy. It should be acknowledged here that I do not believe that we should recognise this dimension of autonomy simply because we need to be able to make sense of the negative obligation incorporated into the principle of respect for autonomy. In chapters three and five I shall claim that neglecting to incorporate a practical dimension into our overall theory of autonomy is also a theoretical deficiency. For the purposes of this introductory chapter though, I suggest that an adequate theory of autonomy *in toto* for use in bioethical contexts must incorporate conditions pertaining to both the reflective and practical dimensions of autonomy. I shall also argue that acknowledging the relationship between the two dimensions of autonomy is integral to explaining why deception and coercion can undermine autonomy (in chapters three and four respectively).

IV Local and Global Autonomy

I have delineated an understanding of autonomy that frames the concept in terms of both a reflective and practical dimension. To conclude this introductory chapter, I shall explain another distinction that I shall use throughout my discussion of autonomy, namely the distinction between local and global autonomy.

Our interest in being self-governing seems to stem from our interest in being in charge not only of our individual decisions and acts, but also of our diachronic projects,

and indeed, our own lives. Accordingly, when we consider the question of whether an agent is autonomous, it is possible to ask this question at both a global and local level. Conceived as a global concept, autonomy is, as Dworkin claims:

. . . a feature that evaluates a whole way of living one's life (that) can only be assessed over extended portions of a person's life.³⁴

Dworkin himself claims that autonomy is intuitively only a global concept, since he believes that it is odd to claim that people can switch back from autonomy to non-autonomy over short periods of time.³⁵ I do not share Dworkin's intuition here; it is not at all clear why it must be odd to suppose that an agent might be autonomous with respect to a particular decision but not to another one shortly after. This is particularly true in bioethical discussions of informed consent; for instance, it seems plausible that a physician could ensure that a patient was able to autonomously consent to some intervention by adequately informing them about the nature of the intervention, but fail to do so for another intervention shortly after. Accordingly, I shall follow Christman (and others) in claiming that we can also conceive of autonomy as a *local* property that an agent instantiates in a specific time-slice with respect to particular acts and decisions.³⁶

The question of whether an agent is locally autonomous is perhaps less complex than the question of whether an agent is globally autonomous. Although it might be

³⁴ Dworkin, *The Theory and Practice of Autonomy*, 16.

³⁵ Ibid.

³⁶ Christman, "Autonomy and Personal History", 3.

clear how to assess an agent's autonomy with regards to a particular decision in a certain specified set of circumstances, it is not immediately clear how we are to evaluate a person's autonomy as a feature that pertains to extended portions of their life, given the varied circumstances which 'a significant portion of one's life' can include.

One plausible way in which we might assess an agent's global autonomy is to consider whether the agent lives in accordance with diachronic plans of her own choosing, where a diachronic plan is understood to stipulate long-term goals that are based on the set of values that form the agent's conception of the good life. These diachronic plans may vary in length; for example, in a biomedical context, we may say that a patient might have a diachronic plan to overcome some health problem, and that they may make local decisions that will have an effect on their pursuit of that long-term goal; however, some diachronic plans may cover the agent's whole life. Furthermore, it seems that some diachronic plans may be of more importance to the agent than others; typically, it seems that an agent's life-plans will often be central to the agent's sense of 'who she is', whilst other diachronic plans may not represent goals that are particularly central to the agent.

There has been little discussion concerning the relationship between global and local autonomy. For Christman, global autonomy arises as a result of the aggregate of instances of local autonomy over a person's life.³⁷ I shall not employ this sense of global autonomy here. In my view, it seems plausible to claim that some instances of local autonomy can serve to *undermine* the pursuit of global commitments. For example, imagine an agent who values the pursuit of two mutually exclusive diachronic goals. Suppose that the agent continually changes her mind about which goal to

³⁷ Christman, "Autonomy and Personal History", 3.

prioritise. Here, it seems that the agent might make a locally autonomous decision to act in pursuit of one goal that will threaten the successful fulfilment of the other competing goal. The mere fact that the agent might be autonomous with respect to her local decisions does not seem to contribute to her global autonomy in this case, because her locally autonomous decisions to act in pursuit of alternating competing goals undermines her ability to successfully pursue either of them.

In stressing the importance of diachronic plans to global autonomy, I am not claiming that an agent's life must be unified by a certain single set of static diachronic plans throughout her life.³⁸ Clearly, people, and their circumstances, change over time, and they may change their diachronic plans accordingly. However, it seems that at least some threshold level of stability is required, so that the agent has sufficient time to commit to long-term goals that can confer an intelligible diachronic purpose to her decisions and actions. Furthermore, the nature of the way in which we change our plans is important. If an agent is to maintain their global autonomy despite a significant change in their plans, then they must be locally autonomous with respect to their decision to change their plans.

I mentioned above that Beauchamp and Childress' claim that autonomy is a property of choices rather than agents belies a failure to acknowledge the distinction between local and global autonomy; I am now in a position to explain this point. Beauchamp and Childress claim that the reason why autonomy should not be understood as a property of agents in a bioethical context is that:

³⁸ Raz, *The Morality of Freedom*, 37 raises this concern.

. . . even autonomous persons with self-governing capacities sometimes fail to govern themselves in particular choices . . . (and) some persons who are generally not capable of autonomous decision-making can, at times, make autonomous choices.³⁹

Pace Beauchamp and Childress, these cases do not demonstrate that autonomy should not be conceived of as a property of persons; rather, these cases just show the importance of distinguishing local and global autonomy. With respect to the first case, there is no reason to think that a person's failure to make a locally autonomous decision must necessarily undermine their status as a globally autonomous person; indeed, I shall suggest in chapter five that sacrificing our local autonomy with regards to a particular decision might sometimes be necessary for securing our global autonomy. Furthermore, we can also claim that a person might lack the capacities that are necessary to autonomously form and execute diachronic plans, and yet claim that they can be locally autonomous with respect to simple, synchronic decisions. As such, in contrast to Beauchamp and Childress, I shall claim that autonomy is a property of persons, and that a person's desires, intentions, actions and decisions are autonomous in a *derivative* sense; they are, I suggest, things that an agent can be autonomous *with respect to*.

Conclusion

In this introduction, I have attempted to map some of the contours of a plausible pre-theoretical understanding of autonomy, in preparation for the theoretical analysis

³⁹ Beauchamp and Childress, *Principles of Biomedical Ethics*, 58.

that I shall undertake in the following chapters. In the first two chapters, I shall investigate what I have introduced as the 'reflective' dimension of autonomy. This dimension of autonomy has received a great deal of attention in the philosophical literature; as such, it is incumbent upon me to consider some of the prominent views that have shaped the current debate, and to explain why they are inadequate. This shall be my aim in the first chapter. In the second chapter, I shall present and defend my own rationalist conception of the reflective dimension of autonomy. In chapter three, I shall go on to defend the inclusion of conditions pertaining to the practical dimension of autonomy in an overall theory of autonomy in bioethics, before going on to argue in chapter four that acknowledging the practical dimension of autonomy allows for a more nuanced explanation of why coercion can undermine autonomy.

I shall conclude the theoretical investigation of autonomy by considering the value of autonomy in chapter five. In this chapter, I shall also begin to apply my theory of autonomy to contemporary bioethical issues, and I shall continue to do so in the final three chapters. In chapter six, I shall consider ways in which it might be possible to enhance personal autonomy through the use of emerging biotechnologies. In chapters seven and eight, I shall consider the ramifications that my theory has for the doctrine of informed consent that is invoked in many different debates in contemporary bioethics.

Chapter One – Two Prominent Views Concerning the Reflective Dimension of Autonomy

In the introduction, I introduced what I termed the ‘reflective’ dimension of autonomy. In this chapter I shall consider two prominent types of compatibilist theory that I believe fail to give an adequate account of this dimension of autonomy. I shall first consider what we may term desire-based ‘hierarchical theories’ of autonomy, which claim that autonomous agents must endorse their first order motivating desires with a further sort of higher-order conative state. I shall then consider so-called ‘historical theories’, which claim that autonomous agents act only in accordance with motivating desires that have a certain sort of aetiology.

In the process of presenting the reflective dimension of autonomy in the introduction, I introduced the concept of an ‘authentic’ desire. In view of the prevalence of this terminology in the literature on autonomy, prior to beginning the exposition of the hierarchical theories of autonomy that often make use of this terminology, it is prudent to clarify how the concept of ‘authenticity’ and ‘the real self’ is used in these discussions.

I Hierarchical Theories: Real Selves, Authenticity and Autonomy

As I pointed out in the introduction, the term ‘autonomy’ is derived from the Greek ‘autos’ (self), and ‘nomos’ (law). Perhaps in view of this etymology, one way in

which philosophers have approached the question of what is for an agent to be autonomous has been to try to give an account of the ‘self’ that should be understood as the source of the ‘government’ in autonomous agency.¹ I also suggested in the introduction that it is possible to interpret the Kantian account of autonomy as claiming that the autonomous agent must be governed by their ‘*noumenal* self’, that is, the self understood as a member of the transcendental realm of pure reason in Kant’s metaphysics. However, the procedural theorists of autonomy that I shall consider in the following two chapters understand the concept of ‘the self’ in a broader fashion. They understand the self to refer to some central persisting element of the agent that incorporates their most deeply held desires, values and beliefs. On the understanding of the self that these theorists seem to invoke, the self can be understood as the metaphorical locus of what Mill terms the agent’s ‘character’ in his discussion of individuality,² or of the psychological continuities that ground personal identity on some theories.³

In my discussion, I shall follow other procedural theorists in invoking this understanding of the self, and the related concept of authenticity; an agent’s authentic desires may be understood as those desires that are true to the central elements of the agent’s self. It is worth noting that this understanding of the self is in tension with some

¹ For instance, Ekstrom writes “ . . . in order to understand autonomous action . . . we need a working conception of what constitutes the ‘self’” Ekstrom, “A Coherence Theory of Autonomy,” 599. In contrast, Berofsky argues against conceiving of autonomous agency as that which proceeds from some extant metaphysical self. See Berofsky, *Liberation from Self*.

² Mill, *On Liberty*, Chapter Three.

³ See Parfit, *Reasons and Persons*, (Part Three) for a classic psychological theory of personal identity. As I shall discuss below, Bratman explicitly points out that the self-governing policies that undergird autonomy on his view are inextricably related to the agent’s identity, since they concern plans that are constituted by psychological continuities. See Bratman, “Planning Agency, Autonomous Agency,” 41.

other prominent understandings of the concept. For instance, on this understanding it not merely a ‘grammatical error’ to claim that agents can have a self in some sense, *pace* Kenny.⁴ Moreover, in holding that the self is something that both persists over time, and can undergird the intelligibility of the agent’s long-term diachronic plans, this understanding of the self is not compatible with those theories that deny that the self exists in a diachronically continuous sense,⁵ or those that deny that the self can persist over long periods of time.⁶

However, this understanding of the self is compatible with a number of claims that are incorporated into a diverse range of theories of the self. For instance, the understanding that I shall invoke here is compatible with Dennett’s deflationary account that denies that the self has a physical locus, and which instead claims that the self is the centre of an agent’s ‘narrative gravity’.⁷ Furthermore, the understanding that I invoke is not committed to the contentious claim that the self is static; on the understanding that I invoke, the self can persist even if the desires, beliefs, and values that constitute it change over time, as long as the agent changes them in accordance with the sorts of procedure that the theories of autonomy that I survey here attempt to explicate.

With this understanding in mind, it seems that one way in which one could attempt to explicate the reflective dimension of autonomy is to explicate what it is for an agent’s motivating desires to reflect their ‘real self’. Frankfurt’s theory of freedom of

⁴ Kenny, *The Self*, 4.

⁵ For example, see Hume’s exposition of his so-called ‘Bundle Theory of the Self’; Hume, *A Treatise of Human Nature* (section entitled ‘Of Personal Identity’).

⁶ For example, see Strawson’s ‘Pearl View of the Self’ in Strawson, “The Self”. For an explanation of how Strawson’s and Hume’s views differ, see Strawson, “Hume on Himself.”

⁷ Dennett, *Consciousness Explained*.

the will can be viewed as such an attempt.⁸ It should be acknowledged that Frankfurt's intention in the work I discuss below was to provide a theory of freedom of the will and its relation to personhood, rather than autonomy *per se*. However, this has not prevented many commentators regarding his theory as a prominent example of a theory of autonomy.⁹ Moreover, Frankfurt's influence is apparent in Dworkin's original theory of autonomy, in so far as Dworkin claimed that a Frankfurtian conception of authenticity was a necessary, although not sufficient,¹⁰ condition of autonomy.¹¹

According to Frankfurt, conscious entities have 'first order desires' to do or have certain things. Some of these desires are the ones that actually motivate them to act; for Frankfurt, it is these 'effective' first order desires that constitute 'the will'.¹² So, on Frankfurt's view, if I have a desire to *x* and I end up *x*-ing, this particular first order desire is effective, and thereby constitutes my will. In so far as we are creatures that have such first-order desires, nothing separates humans from other members of the animal kingdom. However, according to Frankfurt, 'persons' are unique in that they can also have 'second order desires'; these desires are 'higher order' desires that have as their object a certain first order desire.¹³ Further, persons can have second order *volitions* that a particular first order desire become effective in moving them to act.¹⁴ To illustrate, suppose Jones has two conflicting first order desires; he wants to eat a

⁸ Frankfurt, "Freedom of the Will and the Concept of a Person."

⁹ Taylor is a notable exception. See his arguments against this interpretation in Taylor, *Practical Autonomy and Bioethics*, Chapter Three.

¹⁰ Dworkin also claims that 'Procedural Independence' is also a necessary condition on autonomy. Dworkin, "Autonomy and Behavior Control," 25. I shall consider this below.

¹¹ *Ibid.*, 24–25. As I point out below, Dworkin came to abandon this Frankfurtian claim.

¹² Frankfurt, "Freedom of the Will and the Concept of a Person," 8.

¹³ *Ibid.*, 10.

¹⁴ *Ibid.*

gourmet meal, but he simultaneously wants to refrain from indulgent eating because he is on a diet. It seems that Jones could have a second order volition for his first order desire to refrain from indulgent eating to be effective in moving him to act.

The relationship between the agent's effective first order desires, and their second order volitions is integral to freedom of the will for Frankfurt. He writes:

. . . it is in securing the conformity of his will to his second-order volitions . . . that a person exercises freedom of the will.¹⁵

Understood as a theory of autonomy, it seems that Frankfurt's hierarchical approach can explain why being alienated from one's motivating desire can undermine autonomy. Recall the example of Jane from the introduction.¹⁶ We can understand Jane as having two conflicting first order desires; she feels compelled to make herself sick after meals, but she also harbours a desire not to do this. We can understand her as also having a second order volition for the desire not to make herself sick to constitute her will. Nonetheless, Jane's first order desire to make herself sick becomes effective. We may contrast Jane with another bulimia sufferer Beatrice: suppose that Beatrice has only one first order desire to make herself sick. However, unlike Jane, suppose that Beatrice's second order volition is that this desire *should* come to constitute her will. She embraces her bulimia, and would reinstate her first order desire should it wane.¹⁷

¹⁵ Ibid., 15.

¹⁶ Introduction, 9.

¹⁷ These examples correspond to Frankfurt's examples of the unwilling addict and the willing addict. Frankfurt, "Freedom of the Will and the Concept of a Person," 12–15.

On a Frankfurtian model, Jane is not autonomous with regards to her motivating desire to make herself sick. The reason that Jane lacks autonomy on such a theory is that her second order volition is incongruous with her effective first order desire. Accordingly, we may claim that she does not identify with her first order desire to make herself sick, and its role in motivating her to act. In contrast, according to these hierarchical theories, Beatrice is autonomous because her second order volition endorses her first order desire to make herself sick; her desire is, according to the early hierarchical theories, authentic.

Frankfurt's desire-based hierarchical theory accounts for the reflection that autonomy requires by appealing to the agent's higher order conative attitudes towards their motivating desires; these higher order attitudes are understood to have the authority to speak for the agent's 'real self' on these views. One plausible explanation for why Frankfurt and Dworkin appeal to the authority of conative attitudes in their desire-based hierarchical theories is that they endorse a Humean model of practical reasoning of the sort that I mentioned briefly in the introduction.¹⁸ On Humean models, reason is motivationally inert; in order to *x*, an agent must have a desire to *x*; beliefs by themselves cannot motivate. As such, reason is (famously) for Hume, "the slave of the passions";¹⁹ although it has a role in deliberations concerning matters of fact and relations among ideas, it does not, and cannot motivate us to act.²⁰

¹⁸ Introduction, 10. See also Watson, "Free Agency," 207.

¹⁹ Hume, *A Treatise of Human Nature*, Book II, Section III.

²⁰ Roger Crisp suggested to me that Frankfurt's appeal to conative attitudes might also be grounded by a general commitment to non-cognitivism in ethics. I cannot pursue the relationship between these two explanations here, although both might plausibly be correct.

The debate concerning whether or not the Humean conception of practical reason is correct has received considerable attention,²¹ and I do not intend to consider it in great depth here (although I shall say a little more about how rationalist views of reflective autonomy relate to this debate in the next chapter); the above discussion is intended to be purely illustrative. As such, I shall consider Frankfurt's desire-based hierarchical theory on its own merits, regardless of the plausibility of the Humean model of practical reasoning it seems to implicitly assume. In the next section, I shall delineate two objections to desire-based hierarchical theories that I consider to be particularly powerful.

II Two Objections to Hierarchical Theories

i) *The Problem of Manipulation*

As I suggested in the previous section, desire-based hierarchical theories can explain why self-alienated agents like Jane lack autonomy. However, this is not the only sort of threat to autonomy that can occur on the level of the reflective dimension of autonomy. As I suggested in the introduction, it seems that being manipulated to have a motivating desire can undermine autonomy, and it is not clear that desire-based hierarchical theories can adequately explain why this is so.

²¹ For a defence of the Humean model see Smith, *The Moral Problem*, 92–129. For arguments against, see Nagel, *The Possibility of Altruism*, Part Two and Platts, *Ways of Meaning*, Chapter Ten.

To illustrate, recall the example of Manuel from the introduction; Manuel was (non-consensually) hypnotised to form the desire to rob a bank.²² A proponent of the Frankfurtian model might claim that Manuel's desire is not authentic because he does not endorse it at a second order level. However, suppose that Manuel was hypnotised to also have the second order volition that his first order desire to rob the bank become effective. It seems counter-intuitive to claim that Manuel is autonomous with regards to his desire here, even if, by Frankfurt's lights, the desire is authentic.

At first, this objection seems less problematic for Dworkin's hierarchical theory. As I explained above, in his early work, Dworkin endorsed Frankfurt's claim that an agent's identifying with their first order desire is necessary for autonomy. Crucially though, he argued that it was not sufficient. In fact, he supplements his account with a further necessary condition that seems to be in place to rule out this sort of manipulation case. On Dworkin's account, an agent will fail to be autonomous if she does not have "procedural independence"²³ in coming to identify with her desire.

However, procedural independence is something of a slippery notion. In a later discussion, Dworkin explains that spelling out the conditions of procedural independence would involve, *inter alia*:

. . . distinguishing those ways of influencing people's reflective and critical faculties which subvert them, from those which promote them.²⁴

²² Introduction, 10.

²³ Dworkin, "Autonomy and Behavior Control," 25.

²⁴ Dworkin, *The Theory and Practice of Autonomy*, 18.

Dworkin fails to go beyond this sketch to provide a detailed account of procedural independence. Accordingly, as Taylor suggests, the problem of manipulation still remains for Dworkin's account at an indirect level, since Dworkin only avoids the problem of manipulation "by fiat";²⁵ his theory fails to sufficiently explain what demarcates failures of procedural independence or *why* such failures undermine the agent's autonomy. Thus, even if the procedural independence requirement is a plausible condition for autonomy, it lacks both substance and explanatory power.

ii) *The Authority Dilemma*

The second objection to hierarchical theories can be phrased as a dilemma concerning the authority of the agent's higher-order conative attitudes to represent the agent's real self. When we consider the second order volitions through which the agent identifies with their first order desires on desire-based hierarchical theories, a question that naturally arises is why the second order endorsement of a first order desire should be sufficient evidence that the agent truly identifies with their first order desire.

It seems that two equally unappealing responses are possible. On the first horn of the dilemma, the hierarchical theorist could claim that an *even higher* order volition authenticates one's second order volitions as being one's own. However, this reply is implausible because it seems to lead inexorably to a regress of increasingly higher order conative attitudes. On the second horn, the hierarchical theorist might claim that at some level, a higher order desire cannot be authenticated, and does not require authentication.

²⁵ Taylor, "Introduction," 5.

However, this reply leads to what Christman terms the *ab initio* problem,²⁶ since it implies that the authenticity of one's first order desires can only be ensured by a second (or higher) order desire that is not itself authentically the agent's. As Christman puts it, this would involve the claim that ". . . desires can be autonomous without foundations"²⁷, and this, he claims, renders the second response "implausible".²⁸

III Internalist Responses

In light of the two problems highlighted above, some theorists have responded by maintaining the hierarchical structure of Frankfurt and Dworkin's early theories, but changing the content of the relevant higher order attitudes in order to explain how we are to distinguish those of an agent's desires that are authentically theirs, whilst responding to the authority dilemma. Others have abandoned this hierarchical structure, and argued that an adequate account of autonomy must incorporate historical conditions that are more detailed than Dworkin's appeal to 'procedural independence'. I shall consider these latter theories in section IV.

The difference between these two approaches is borne out by a distinction that Mele draws between 'internalist' accounts of authenticity, and 'externalist' accounts of authenticity.²⁹ According to internalist accounts, the fact that some desire satisfies a certain sort of internal psychological scrutiny by the agent is both necessary and

²⁶ Christman, "Autonomy and Personal History," 7.

²⁷ Ibid.

²⁸ Ibid.

²⁹ Mele, *Autonomous Agents*, 146–147.

sufficient for establishing that the desire is authentic; Frankfurt's early theory is internalist in this way, and so are the 'neo-hierarchical' theories that I discuss in this section. In contrast, although this psychological scrutiny may still be necessary on externalist accounts, it cannot be sufficient; a further necessary requirement of authenticity on the externalist account is that the causal history of the desire meets certain conditions.

Although an adequate theory of autonomy must respond to both the problem of manipulation *and* the authority dilemma, it seems that internalist and externalist accounts of authenticity differ with regards to the problem that they are intended primarily to address. Internalist accounts seem to respond primarily to the authority dilemma, whilst externalist accounts seem to respond primarily to the problem of manipulation. In this section, I shall consider three internalist accounts offered by Frankfurt, Dworkin, and Bratman, before going on to consider the externalist accounts of Christman and Mele.

Frankfurt responded to what I have called the authority dilemma by claiming that an agent's second order desires must be adopted *decisively* if they are to count as having the authority to speak for the agent's real self.³⁰ The thought here is that the agent's decisive second order desires have the authority to speak for the agent's real self in so far as she commits to them wholeheartedly, and without reservation. With the decisiveness condition in place, Frankfurt claimed that his theory was able to escape what I have called the authority dilemma. He argued that the *ab initio* problem does not

³⁰ See Frankfurt, *The Importance of What We Care About*, 21.

arise on his revised account because an agent's second order desire has authority just by virtue of its being decisively adopted.³¹

However, this condition seems problematic for two reasons. First, we may ask if the decisiveness condition can even adequately respond to the authority dilemma; the stipulation of this condition immediately raises the question of what grants authority to the agent's decision to commit wholeheartedly to a desire.³² At this point, the hierarchical theorist might claim that the decisive commitment should be understood as being reflexive; as such, the commitment should be understood as being self-justifying, and thereby authoritative.³³ However, this smacks of bootstrapping; the 'justification' of the decisive commitment seems to have been produced 'out of thin air', and thus does not seem to present a convincing response to the authority dilemma. Moreover, even if one claimed that a reflexive decisive condition *could* provide an adequate response to the authority dilemma, the second problem is that Frankfurt's theory is still vulnerable to the problem of manipulation. As Christman claims, it seems plausible to imagine an agent who is manipulated to decisively adopt a second order preference, as Frankfurt's condition requires; yet we would not claim that such an agent is autonomous.³⁴

In light of this second problem, Frankfurt claimed in later work that the agent's second order identification with their effective first order desires does not need to be a

³¹ Ibid.

³² On the rationalist account that I shall defend, the fact that an agent has decided in accordance with what I term their personally authorized preferences would confer authority on such a decision. However, this view is not available to a desire-based hierarchical theory, since the view I defend relies on the claim that only an agent's *rational* desires have agential authority.

³³ Lehrer, "Reason and Autonomy" makes this sort of move.

³⁴ Christman, "Autonomy and Personal History," 9.

“deliberate psychic event”.³⁵ Instead, Frankfurt argues that if an agent is “satisfied”³⁶ with their first order desires (when being satisfied amounts to lacking any desire to change her first order desires), then this is sufficient for establishing the authenticity of these desires. It seems plausible to claim that part of Frankfurt’s motivation in appealing to the concept of satisfaction is to attempt to avoid the problem of manipulation, since it might be claimed that it is not possible to manipulate someone to be satisfied with their desires, just because satisfaction is not a deliberate psychic event.

However, the appeal to satisfaction cannot solve the problem with Frankfurt’s theory. As Bratman points out, sufferers of depression are likely to be satisfied with their first order desires, since they will lack the desire to change them;³⁷ however, this is hardly evidence of the agents’ autonomy with respect to these desires. The problem here is that the satisfaction that clinically depressed agents have with regards to their desires is a result not of their positive endorsement of their first order desires, but rather of their pathological apathy. Moreover, it seems plausible that an agent could be manipulated (by way of a neurological intervention) to experience this sort of pathological apathy. As such, even if we concede that Frankfurt’s satisfaction criterion provides him with the means to respond to the authority dilemma, his revised theory is still vulnerable to the problem of manipulation.

In contrast to Frankfurt, it is noteworthy that Dworkin came to abandon the identification requirement on autonomy that stipulates that the autonomous agent must identify with their effective first order desires. Instead, he came to argue that “the capacity to raise the question of whether I will identify with or reject the *reasons* for

³⁵ Frankfurt, *Necessity, Volition, and Love*, 104.

³⁶ *Ibid.*, 103–104.

³⁷ Bratman, “Identification, Decision, and Treating as a Reason,” 7–8.

which I now act”³⁸ is necessary for autonomy. Notice here that the emphasis on Dworkin’s revised account is on the agent’s reasons for acting, rather than the relationship between the agent’s motivating desires and higher order desires.³⁹ I believe that Dworkin’s revised account is a step in the right direction, and I shall endorse an account that shares the sentiments of Dworkin’s revised view. However, at this stage it is illuminating to point out a flaw in Dworkin’s revised view, because it concerns the strategy that he employs to avoid the authority dilemma, a strategy that could also be employed by a neo-Frankfurtian account.

Dworkin argues that his revised account does not face the authority dilemma because he is only providing an account of *global* autonomy. As such, he argues that there is “no conceptual necessity”⁴⁰ to ask whether the second order capacity to reflect on one’s motivating desires is endorsed at a yet higher order level. He also goes on to claim that it is a “contingent fact”⁴¹ about humans that their reflections cease at the second order level; to reflect at this second order level simply equates to being globally autonomous on Dworkin’s view. As such, he claims that the internalist element of his theory does not fall into the regress horn of the authority dilemma.

³⁸ Dworkin, *The Theory and Practice of Autonomy*, 15. Emphasis added.

³⁹ In his revised theory Dworkin claims that also autonomy requires the ability to both “alter one’s preferences and to make them effective in one’s actions” (Ibid., 18). Christman objects to this aspect of Dworkin’s theory, claiming that it is possible to lack the ability to alter preferences that one cares most deeply about, but one is still autonomous when one acts in accordance with them. Christman, “Autonomy and Personal History,” 6. Whilst I agree with Christman’s observation, it is not a valid objection to Dworkin’s theory, since Dworkin a closer reading reveals that Dworkin stipulates that autonomy is the capacity to *accept or attempt* to change one’s first order preferences. Dworkin, *The Theory and Practice of Autonomy*, 20.

⁴⁰ Dworkin, *The Theory and Practice of Autonomy*, 20.

⁴¹ Ibid., 19.

The problem with Dworkin's use of this strategy in his revised account is that it is not at all clear how his appeal to global autonomy is supposed to allow his account to escape from the authority dilemma. Dworkin just seems to assume that the relevant sort of second order reflection is sufficient for autonomy; however, it still seems appropriate to ask *why* such reflections would have agential authority, even on a theory of global autonomy. Dworkin just denies this, since he simply equates global autonomy with this second order reflection; no explanation is given as to why the fact that his theory concerns global (and not local) autonomy takes away the conceptual necessity to investigate the authority of the agent's higher order reflection. However, since such an investigation is deemed conceptually necessary for accounts of local autonomy, it seems that Dworkin owes us at least a more detailed explanation of how the appeal to global autonomy relinquishes him from this conceptual necessity. Indeed, in his assessment of Dworkin's revised theory, Christman still charges that Dworkin's theory is susceptible to what I have called the authority dilemma, and does not refer to Dworkin's preemptive appeal to global autonomy;⁴² I suspect that Christman is as dubious about Dworkin's appeal as I am.

Therefore, in addition to the problems concerning the condition of procedural independence mentioned above, the internalist element of Dworkin's revised theory is also problematic, since Dworkin fails to adequately explain why a conception of global autonomy should not be susceptible to the authority dilemma. We might also point out that in providing a theory of only global autonomy, it is not clear that Dworkin's theory

⁴² Christman, "Autonomy and Personal History."

will be applicable to all practical questions concerning autonomy, since many of these will concern local, rather than global autonomy.⁴³

Michael Bratman provides an internalist hierarchical account that *does* seem to avoid these pitfalls whilst responding to the authority dilemma. Unlike Dworkin's account, Bratman's account tackles the authority dilemma head-on by stipulating that the higher order conative states that govern the agent's first order desires will have agential authority if they represent plan-like attitudes that are constitutive of an agent's personal identity. According to Bratman, our planning agency is central to our personal identity. In forming plans that are stable over time, humans “. . . pursue complex forms of cross-temporal organization”⁴⁴, since these plans refer the agent's future action back to their earlier self. Crucially for Bratman's account of autonomy, this cross-temporal aspect of planning makes an agent's planning attitudes central to their personal identity, since he claims that plans are “constituted by psychological continuities”⁴⁵ of the sort that can constitute identity on Lockean views of personal identity.⁴⁶

Although Bratman phrases his argument in terms of personal identity, his use of this terminology overlooks an important distinction between ‘numerical’ and ‘narrative’ identity. According to Schetmann, the former sense of personal identity can be understood as the sense we use when we ask questions concerning re-identification, that is to say questions concerning what it is that makes an entity at time *t1* the same thing as itself at a later time *t2*. In contrast, narrative identity is the sense of identity at work

⁴³ Taylor raises this sort of objection. See Taylor, “Introduction,” 8.

⁴⁴ Bratman, “Planning Agency, Autonomous Agency,” 41.

⁴⁵ Ibid.

⁴⁶ See An Essay Concerning Human Understanding. Book II Chapter XXVII. For a modern discussion of Lockean theories of personal identity, see Parfit, *Reasons and Persons*, Part Three.

when we attempt to answer questions concerning the psychological features that constitute a person's identity in the sense of their character.⁴⁷ With this distinction in mind, it seems clear that if what matters with regards to the agential authority of an agent's hierarchical conative states is the relation of those states to the agent's identity, then the notion of personal identity at work here is the agent's narrative identity, and not the sense of numerical identity which undergirds issues concerning re-identification. However, I shall follow Bratman in using the broader term 'personal identity' in my discussion.

To return to Bratman's account, Bratman defines a self-governing policy to be:

. . . a higher order policy that favours the functioning of a desire as having a justificatory weight in effective practical reasoning.⁴⁸

Given the role of these policies in constituting the agent's identity, it is clear why this sort of higher order conative attitude is not susceptible to the authority dilemma. As Bratman explicitly states, self-governing policies have agential authority just because they are tied to the agent's identity.⁴⁹ The authority of these higher order attitudes does not require a regress; nor do they spring out of nowhere as per the *ab initio* horn of the dilemma. Rather, the authority of self-governing policies arises from their being constitutive of the agent's identity.

⁴⁷ See Schechtman, *The Constitution of Selves*, 1-2.

⁴⁸ Bratman, "Planning Agency, Autonomous Agency," 43.

⁴⁹ *Ibid.*, 42.

Bratman goes on to flesh out his theory in more detail by introducing an altered conception of Frankfurtian satisfaction in order to ensure that autonomous agents cannot be alienated from their planning attitudes.⁵⁰ He also stipulates that these self-governing policies must be reflexive, that is to say self-endorsing, so as to prevent the threat of a regress at the level of the agent's endorsement. I have already suggested that this sort of appeal to reflexivity smacks of bootstrapping. However, we can consider the merits of Bratman's position without considering this problematic aspect, which I believe can be circumvented in other ways.⁵¹ What is important for my discussion is that Bratman's theory allows the internalist to respond to the authority dilemma, by tying the higher order conative attitudes that authenticate one's first order desires to the agent's identity. Moreover, it does so in a manner that allows for autonomy at both the global and local level.

The main problem though, as Bratman himself recognises, is that although his theory seems able to respond to the authority dilemma, it too remains susceptible to the problem of manipulation.⁵² Although self-governing policies might have the authority to speak for the agent's 'real self', it still seems that an agent could have been manipulated to have a certain set of self-governing policies. The reason for this is that Bratman does not rule out the possibility that the agent could have been manipulated to have her particular identity-constituting planning attitudes.

However, Bratman's theory also faces another problem, since he neglects to stipulate the basis upon which the autonomous agent should endorse a policy that

⁵⁰ Ibid., 43–44.

⁵¹ The theory that I develop in the next chapter shall respond to a similar problem by appealing to a holistic, rather than foundationalist justification of one's motives or conception of the good.

⁵² Ibid., 35 and 58.

favours a desire as having justificatory weight in practical reasoning. This is problematic because if it is the case that the autonomous agent need only endorse a self-governing policy on the basis of the mere fact that it concerns a desire that the agent simply happens to have, then autonomy would require very little agential input on Bratman's theory; all that autonomy would require is that we formalise the desires we find ourselves having into appropriate self-governing policies. However, we may question whether this minimal degree of input is really sufficient for autonomy. It seems that Bratman thus fails to offer an explanation for *why* we should adopt the self-governing policies that ground our autonomy; in lacking such an explanation, it is not clear that the adoption of self-governing policies is in fact sufficient for self-governance.⁵³

This latter consideration partly motivates the move that I make towards stipulating a rational requirement in the theory of autonomy that I shall propose in the next chapter. To conclude this chapter, I shall consider externalist theories of authenticity which aim to provide an answer to the problem that plagues all of the internalist theories considered here; namely, the problem of manipulation.

IV Externalist Responses

⁵³ Bratman expands on this elsewhere by claiming that the autonomous agent 'decides' to treat a desire as reason-giving. Bratman, "Identification, Decision, and Treating as a Reason." However, this does little to answer my objection, since it is not clear what the basis for this decision must be. Indeed, in this paper, Bratman seems to implicitly endorse subjectivism about reasons; in chapter two, I shall follow Parfit in claiming that we ought to reject subjectivism about reasons in favour of objectivism.

According to externalist accounts of authenticity, the *aetiology* of one's motivating desire is the crucial consideration in questions concerning autonomy. Christman argues that in order to assess an agent's autonomy, we must ask "... if the person would have, or did resist the adoption of a value or desire, and for what reasons".⁵⁴ An agent will fail to be autonomous if she is (or would hypothetically be) unable to resist adopting a certain desire, or if she is unable to judge the process by which she came to have the desire in a "minimally rational"⁵⁵, and "self-aware"⁵⁶ manner. For Christman, minimal rationality requires that one's judgemental deliberation must be consistent, and free from any other logical failing,⁵⁷ whilst being self-aware of the processes of one's belief formation requires that one does not deceive oneself about the nature of these processes.⁵⁸

To cash out this account, consider again the manipulated agent Manuel from the introduction. I suggested that a problem with desire-based hierarchical theories is that they must conclude that Manuel's desire for *x* is authentic, even if the second order conative attitude that endorses it was formed as a result of his being manipulated. The strength of the externalist account is that it can explain why Manuel's desire would be inauthentic in this manipulation case. On Christman's externalist account, the desire is inauthentic, because Manuel would have resisted the desire if he had been able to, and if he had known that he was being manipulated to have it.

⁵⁴ Christman, "Autonomy and Personal History", 10.

⁵⁵ Ibid., 12.

⁵⁶ Ibid., 16–17.

⁵⁷ Ibid., 12.

⁵⁸ Ibid., 16–17.

Can Christman's externalist account provide a response to the problem of manipulation? *Prima facie*, it seems that it could; as we saw with the example of Manuel, on Christman's account, an agent lacks autonomy with respect to those desires whose origin he believes he should reject. However, it seems that the problem of manipulation remains in a different form for Christman's account, since it seems possible for one to lack autonomy with respect to a desire even if one endorses the aetiology of that desire in a self-aware and minimally rational manner.

To explain why, it is important to observe that an agent's judgement about whether or not to approve the aetiology of a particular desire is likely to be informed by values and desires that she already has, and that she may have acquired (and sustained) in a manner that seems to be inimical to autonomy. In view of this, an agent may just fail to comprehend how some causal pathways of attaining a desire are not suitable, without failing to be minimally rational or self-aware; however, this does not entail that the aetiology of the desire is suitable or should be endorsed.

This is best illustrated by way of example.⁵⁹ Consider an adolescent who has been brainwashed by her parents to become a Jehovah's Witness, and who harbours a desire to continue living in this way of life. We can imagine asking her whether that is really what *she* wants to do, or whether it is in fact her parents' desire. Such an adolescent might maintain that living this way is what *she* wants to do. Suppose that we suggest to her that she only has this desire and these values because she was born into a strict religious family, and she was brainwashed to hold these values. Crucially, this information need not (and we might assume *would* not in our example) affect the adolescent's endorsing the means by which she came to have her desire to continue

⁵⁹ For a similar case, see Wolf's case of Jo-Jo in Wolf, "Sanity and the Metaphysics of Responsibility," 53.

living as a Jehovah's Witness; rather, in so far as she holds the values of a Jehovah's Witness, she is likely to endorse the type of strict upbringing that led to her desire to live this way of life, and to lead it unquestioningly.

The problem with Christman's model is that his stipulation that autonomous agents must meet standards of minimal rationality and self-awareness does not seem to render the Jehovah's Witness non-autonomous. First, we may assume that her endorsement of her strict upbringing meets standards of minimal rationality, in so far as it does not involve inconsistency or any logical failure. Second, we can even assume that the Jehovah's Witness is self-aware of the fact that she has been indoctrinated, but just does not regard this as a reason to reject her desire; she may just believe that this is how desires *ought* to be formed. However, it still seems jarring to claim that the adolescent autonomously desires to live as a Jehovah's Witness, even if she endorses the aetiology of the desire.

A similar case that reinforces this conclusion can be taken from Huxley's 'Brave New World'. In Huxley's novel, all infants are subjected to Skinnerian behavioural conditioning, and to the 'hypnopaedic'⁶⁰ indoctrination of certain values appropriate to their caste. These hypnopaedic suggestions come to constitute the child's value set, as the character of the Director in the novel makes clear in this passage:

Til' at last the child's mind *is* these suggestions, and the sum of the suggestions *is* the child's mind. And not the child's mind only. The adult's mind too – all their life long.

⁶⁰ I.e. sleep-hypnotism.

The mind that judges and desires and decides – made up of these suggestions. But all of these suggestions are *our* suggestions!⁶¹

The infants who are subjected to hypnopaedic indoctrination, and the adults into whom they develop, seem to be a paradigm instance of agents lacking autonomy. However, it is not clear that these Brave New Worlders (henceforth BNWs) would lack autonomy on Christman's account; after all, they are likely to endorse the aetiology of their desires on the basis that *all* BNWs acquire their desires through hypnopaedic interventions.

Mele suggests an alternative externalist account which avoids this sort of pitfall by stipulating that the legitimacy of a desire's aetiology is not to be decided by the agent's own attitude towards it. Rather, Mele suggests that a necessary condition of an agent's possessing an authentic desire is that she was not 'compelled' to have that desire in a manner that means she is practically unable to shed it.⁶² For Mele, to be compelled is not merely to be *caused* to have some desire; rather it is to be caused to have a desire in a manner that bypasses the subject's capacities for control over their mental life.⁶³

Mele's externalist account can explain why BNWs and the above adolescent lack autonomy, since it can claim that children brought up in such circumstances have had those values imputed to them in a way which bypasses their control over their mental life. Even if the adolescent *now* approves of the aetiology of their desires as part of their mental life, the externalist might argue that at the time when the adolescent appropriated the values of a Jehovah's Witness, she lacked the critical facilities to

⁶¹ Huxley, *Brave New World*, 23.

⁶² Mele, *Autonomous Agents*, 166.

⁶³ *Ibid.*, 171. Mele also points out that an agent is not compelled to have some desire if she herself consciously arranges that she should form a desire in a way that otherwise bypasses her mental control.

approve of the cause of her coming to have these values; so too in the case of BNWs. As such, it might be claimed that Mele's externalist account can explain why BNWs and the adolescent Jehovah's Witness do not autonomously form their desires to continue living in the way that they do. The externalist might claim that the reason they lack autonomy is that they are not able to critically reflect on the cause of their initial desires and values, desires and values that now partly constitute the self which critically judges the aetiology of new desires.

Whilst I believe that Mele's account is close to the truth, it is problematic in its current formulation, but not for the same reasons that Christman's theory is flawed. The problem is that if Mele's theory is correct, then the externalist account seems to prove too much, since everybody (whatever their background) is brought up to have *some* set of values that will influence their later judgements, and upon whose source they cannot critically reflect at the time of appropriation. As those who espouse relational views of autonomy point out, we are all, at least in part, an outcome of social and environmental forces that determine many of our values and desires at a pre-critical stage of our development. As Noggle puts it, "there is no self before the socialization that creates it"⁶⁴ in pre-critical childhood development. In a sense then, by Mele's lights, we are all victims of manipulative processes that serve to undermine our autonomy, in so far as we have all had certain values and desires imputed to us during the pre-critical stages of our development, some of which we are now practically unable to shed.

Therefore, although Mele's externalist account might be able to account for why BNWs and the Jehovah's Witness lack autonomy by stipulating that the autonomous agent cannot have any desires that have been formed in a manner which bypasses their

⁶⁴ Noggle, "Autonomy and the Paradox of Self-Creation," 104.

mental control, his theory is in danger of ruling out the possibility of *anyone* being autonomous, once we acknowledge that we are all subject to socializing forces of the sort that Noggle highlights in our pre-critical development.

One reply to this problem would be to return to Christman's theory and to just bite the bullet and claim that our intuition that the above Jehovah's Witness and BNWs lack autonomy is erroneous. However, this would be a huge concession to make; indeed, Lindley goes so far as to suggest that someone could only call BNWs autonomous as "a joke".⁶⁵ Alternatively, one might claim that the desires and values that we develop in pre-critical stages of childhood development represent a special case, and that what is important for these desires and values is that we consider them as part of our mental life once we have developed certain critical competencies, regardless of the fact that we had no input into their formation. This compatibilist move is, I believe, the response that we should make to these problematic cases, and it is also the response that Mele himself seems to make in his own discussion of similar cases.⁶⁶ However, it should be noted that if one makes this move, then *pace* Mele, it is not clear why the *causal history* of one's desires is so important to autonomy. Rather, this move suggests an account of authenticity with a far more internalist flavour, whereby what matters with regards to the authenticity of an agent's desires is that they critically assess them as part of their mental life *once they have them*.

This last claim is also supported by another argument that can be raised against externalist theories. The above objection I made against Mele's theory suggested that his externalist condition cannot be *necessary* for autonomy, since it seems to rule out

⁶⁵ Lindley, *Autonomy*, 45.

⁶⁶ Mele, *Autonomous Agents*, 168. Noggle also makes this move in his defence of his own internalist theory. Noggle, "Autonomy and the Paradox of Self-Creation," 99–102.

the possibility of most people being autonomous. Prior to that, I had suggested that Christman's externalist conditions were not *sufficient* for autonomy, since it was possible for an agent to endorse the aetiology of her desires and yet lack autonomy because she endorsed an illegitimate aetiology. In view these objections, it might be suggested that the best way to proceed in our analysis of autonomy would be to return to Christman's theory and attempt to refine the externalist criterion to cover the thought that the autonomous agent must endorse the *correct* sort of aetiologies.

However, this strategy seems flawed, since even refined externalist conditions do not seem to be necessary for autonomy. As Berofsky points out, it seems plausible to suppose that an agent could reject the aetiology of a particular desire, and yet still hold that desire autonomously if she endorses the desire on other grounds.⁶⁷ To illustrate, suppose that I was brainwashed by subliminal advertising to form the desire to give money to charity. Even if I later came to reject the aetiology of this desire (having been made aware of it), it still seems plausible to claim that I could choose to sustain the desire without thereby lacking autonomy with respect to it; I could do so because I now endorse the *content* of the motivating desire. That is, I might know that I only formed the desire because I was brainwashed, but I might come to re-evaluate the desire *itself* on the back of this knowledge, and nevertheless endorse the desire because I choose to do so on the basis of the nature of the desire's content. As Berofsky puts it:

One may reasonably have objections to the (aetiological) process without having any qualms about the results.⁶⁸

⁶⁷ Berofsky, *Liberation from Self*, 211–212.

⁶⁸ *Ibid.*, 212.

In this case, it is not clear why the externalist conditions are necessary, let alone sufficient for autonomy. This suggests that externalist theories are in fact focusing on the wrong aspect of our desires, since these arguments suggest that it is not the aetiology of our desires that really matters with regards to our autonomy. Rather, we seem to be drawn back towards an internalist picture of authenticity, whereby the authenticity of an agent's motivating desires depends on their reflective endorsement of the *content* of that desire, rather than the aetiology of the desire itself.

Conclusion

To conclude this chapter, I have argued that the two prominent theories of autonomy that I have surveyed here are flawed. I suggested that desire-based hierarchical theories face two key objections; the authority dilemma and the problem of manipulation. I suggested that Bratman's neo-hierarchical internalist theory is able to give an account of why certain second order attitudes have the authority to speak for the agent's 'real self', in so far as these attitudes are linked to the agent's identity. However, his theory fails to give an adequate account of why agents should give certain desires justificatory weight in practical reasoning, and it is unable to explain why manipulation is inimical to an agent's autonomy.

I also argued that appealing to an externalist historical requirement cannot circumvent this latter problem. I argued that such requirements are not necessary for autonomy, let alone sufficient; the aetiology of one's motives is not as central to one's

autonomy as these externalist theories would have us believe. As such, I have suggested that we are drawn back towards an internalist picture. In the next chapter, I shall develop a rationalist internalist theory which claims that the autonomous agent must sustain their motivating desires on the basis of a belief that the object of their desire is good in a reason-implying sense, and because it can be said to cohere with the agent's other rational desires and (minimally rational) beliefs.

Chapter Two – A Rationalist Account of Reflective

Autonomy

In the previous chapter, I argued that desire-based hierarchical theories and historical theories of the reflective dimension of autonomy are both inadequate. In this chapter, I shall consider what may be termed *rationalist* theories of the reflective dimension of autonomy. On the understanding that I shall employ here, rationalist theories of the reflective dimension of autonomy may be distinguished by the fact that they attempt to cash out the reflection that autonomous agents must carry out on their motivating desires by claiming that an agent is only autonomous with respect to those desires that they believe to be, in some sense, rationally warranted. I shall delineate what I believe to be the most plausible rationalist theory, namely Laura Waddell Ekstrom's coherence account, and suggest that some flaws in Ekstrom's theory (and other indeed other rationalist theorists before her) are attributable to the fact that her theory fails to give a sufficiently detailed explanation of what it is for a desire to be rational.

In an attempt to rectify this, I shall (in section three) delineate Derek Parfit's recent work on the rationality of desires, before going on to explain (in section four) how we might apply Parfit's understanding to Ekstrom's theory of autonomy in a manner that would allow her to respond to the problems raised in section two. In section five, I shall argue that an amended version of Ekstrom's theory can respond to the problem of manipulation that plagued the theories surveyed in chapter one.

I Rationalist Understandings of Reflective Autonomy

Philosophers have long appealed to our capacity for rationality in order to explicate the self-government that the concept of autonomy connotes. For instance, Plato claimed that the appetitive and passionate parts of the mind should be subordinate to reason, since reason “. . . looks out for the whole of the mind”.¹ Similarly, Aristotle claimed that the desiderative elements in the irrational part of one’s soul ought to be obedient to reason.² Perhaps most famously, as I mentioned in the introduction, Kant went so far as to suggest that we are only autonomous when we act in accordance with our *purely* rational nature. Kant’s theory of autonomy may be viewed as a *substantive* rationalist theory in so far as the concept of autonomy itself is understood by Kant to have substantive moral content.

Some modern rationalist theorists have sought to move away from this Kantian picture of the relationship between rationality and autonomy by moving towards *procedurally* rationalist accounts of autonomy, which make the broader claim that agents are autonomous with respect to their motivating desires, if they cohere with the agent’s own rational evaluative judgments. Gary Watson proposed one such sort of rationalist account. Watson claimed that an agent is autonomous if her motivating attitudes cohere with her evaluative judgments, which in turn originate in “. . . the rational part of the soul”.³ In contrast to the Frankfurtian model influenced by Hume, Watson suggests that on his Platonic picture, although our desires can be said to motivate us, they are silent with regards to the good; on Watson’s view, our desires are

¹ Plato, *The Republic*, 441e and 434d – 441c.

² Aristotle, *The Nicomachean Ethics*, 1102a26 – 1103a.

³ Watson, “Free Agency,” 208.

“blind or irrational”⁴ impulses. In contrast, reason allows us to identify the good. As such, Watson claims that “. . . the desires of Reason are desires for ‘the good’,⁵ terming those desires of reason, our ‘values’. Given this difference between our values and our non-rational desires, Watson equates autonomous action with those acts that cohere with one’s values, since (on his view) our values represent those of our desires that have a rational basis.

Whilst Watson’s theory represents, I believe, a step in the right direction in linking personal autonomy to the agent’s evaluative judgments, it has been criticised on the grounds that the evaluative judgments that ground autonomy on his view occur at the first order level. Opponents of Watson’s theory have claimed that if one’s theory of autonomy fails to incorporate *any* sort of hierarchical structure, then it will be unable to account for a number of important features about practical deliberation. For instance, Bratman argues that theories of autonomy require hierarchy in order to account for how an agent’s decision in the face of competing rational desires can have agential authority, and in order to explain how agents can act autonomously in the light of motivational pressures that are contrary to their central commitments.⁶

Although I am not convinced by these objections, rather than responding to them at length here, I shall take as my focus for the remainder of this chapter Laura Waddell Ekstrom’s coherence theory of autonomy, which may be regarded as developing Watson’s rationalist conception into character-based model of autonomous agency that can incorporate a hierarchical structure; it is thus not susceptible to the objection to Watson’s theory explained above. Although I shall henceforth consider

⁴ Ibid.

⁵ Ibid., 209.

⁶ Bratman, “Planning Agency, Autonomous Agency,” 38–40 and 49.

only Ekstrom's account, it should be acknowledged that many of the arguments that I make below could be extended to other rationalist theories.

On Ekstrom's theory, an agent is autonomous when they act on a first order desire because they have a 'personally authorized preference' for that desire to be effective. This terminology requires some explanation. On Ekstrom's account, a preference is understood to be a second order desire for a certain first order level desire to be effective in moving the agent to act. However, Ekstrom's understanding of a preference moves away from a Frankfurtian picture of second order volitions, since a preference on Ekstrom's account " . . . is formed in the search for what is good".⁷ Notice that Ekstrom's notion of 'preferences' thus bears a strong resemblance to Watson's conception of rational desires or values, although Ekstrom's preferences are understood to presuppose the existence of higher order mental states, unlike Watson's account of rational desires.⁸

A preference is *personally authorized* if it coheres with what Ekstrom calls the agent's "character system",⁹ which on Ekstrom's view is the agent's set of preferences at time *t*, in conjunction with the set of propositions that the agent accepts at *t*; Ekstrom terms the latter the agent's "acceptances".¹⁰ With this in mind, Ekstrom suggests that a preference will have agential authority in so far as it *coheres* with, and reflects the agent's character. In turn, Ekstrom offers the following definition of coherence, where the phrase "S has a preference for desire *x*" is to be understood as "S has a preference

⁷ Ekstrom, "A Coherence Theory of Autonomy," 603.

⁸ *Ibid.*, 604.

⁹ *Ibid.*, 606.

¹⁰ *Ibid.*

that the first order desire x be the desire that leads S all the way to action when or if S acts’’:¹¹

Df, coherence: a preference for desire d coheres with the character system of S at t if and only if, for any competing preference for desire g , it is either (i) more valuable for S to prefer that desire d than it is for her to prefer that desire g , on the basis of the character system of S at t , or (ii) as valuable for S to prefer that the conjunction of g and a neutralizing desire n as it is for S to prefer that g alone on the basis of character system of S at t .¹²

Ekstrom’s terminology of it being ‘more valuable for an agent to prefer d ’ here is slightly ambiguous. It may be more valuable for an agent to prefer one thing to another for two different kinds of reason. First, an agent’s preferring d might be more valuable for her because she believes that the object of desire d is more valuable than the object of desire g ; such an understanding appeals to object-given reasons. On this understanding, it is the satisfaction of the preference that would be valuable for the agent, rather than simply having the preference *per se*. This, I believe, is the sense that Ekstrom intends to invoke. However, it could be more valuable for an agent to prefer d because she has *state-given reasons* to hold a particular preference; to illustrate, suppose that someone threatened to torture S unless she held the preference for d . This would give me S (state-given) reason to be in the state of holding a preference for d , even if S had no (object-given) reason to want the object of d itself. Whilst Ekstrom’s choice of

¹¹ Ibid., 611.

¹² Ibid.

terminology might lead one to think that she is appealing to state-given reasons, I think it is most natural to understand her view as appealing to object-given reasons.¹³ I shall discuss these reasons in more detail in section III.

There is much to be said in favour of Ekstrom's theory. First, it can respond to the authority dilemma by claiming that those of an agent's preferences that cohere with her character system can be understood as having agential authority, in so far as the agent's coherent preferences and acceptances may plausibly be understood as representing the agent's real self. As Ekstrom points out, there is a strong case in favour of this view, since cohering elements of the self are likely to be "particularly long lasting",¹⁴ since they are "well-supported with reasons".¹⁵ By virtue of this support, they will also be "fully defensible against external challenges",¹⁶ as well as being preferences that the agent feels "comfortable owning".¹⁷

Furthermore, Ekstrom's theory avoids a problem that I raised against Bratman's account. I suggested that a problem with Bratman's account is that it does not stipulate the basis upon which the autonomous agent should endorse a self-governing policy that favours a desire as having justificatory weight in practical reasoning. Unlike Bratman's self-governing policies, Ekstrom's preferences are formed in the search for the good. Accordingly, the agent can give reasons for why she has certain preferences; the desires that she prefers to be effective are those that are likely to lead to the attainment of that which she believes to be good.

¹³ See Parfit, *On What Matters*, 50–52 for further discussion of state-given reasons.

¹⁴ Ekstrom, "A Coherence Theory of Autonomy," 608.

¹⁵ Ibid.

¹⁶ Ibid.

¹⁷ Ibid., 609.

Despite these strengths of Ekstrom's account, in the next section, I shall explain two problems with Ekstrom's theory as it stands, before going on to argue (in section III and IV) that these problems can be overcome by reconsidering Ekstrom's theory through the lens of Derek Parfit's account of rational desires.

II Objections to Ekstrom's Theory

i) *The True and the Good: An Objective or Subjective Standard?*

According to Ekstrom, an agent's preferences and acceptances are those of her desires and beliefs that she forms in her pursuit of the good and the true respectively. She also claims that the good and the true here are to be understood as the agent's *subjective* conception of the good and the true.¹⁸ *Prima facie*, this might seem plausible; after all, if one were to claim that an agent's preferences and acceptances must be formed in her pursuit of an objective conception of true and the good, then this would seem to lead one back to a substantive picture of autonomous agency, rather than a procedural account.

In fact, Berofsky challenges Watson's rationalist account on this score. Berofsky claims that Watson's theory endorses an 'evaluation conception of value', according to which an agent's values must track some objective good.¹⁹ However, Berofsky suggests that our non-rational impulses might be better candidates for our true selves than

¹⁸ Ibid., 606.

¹⁹ Berofsky, *Liberation from Self*, 81 and 89–93.

evaluative judgments that one may feel alienated from.²⁰ Indeed, Watson himself came to repudiate his initial theory on a similar basis in a later paper.²¹ Notably, in his own theory Berofsky endorses a reductionist motivation conception of value, according to which an agent's values are just those of their desires that would be adopted by "a fully informed, minimally rational agent".²² One of the reasons that he endorses such a position is that he believes that the evaluation conception of value is unable to account for the undeniable fact that we "... sometimes place value on senseless or masochistic ends, that is, ends that have no objective value".²³

It seems plausible to claim that this sort of thought may underlie Ekstrom's appeal to a subjective conception of the true and the good. However, notice that Ekstrom also implicitly seems to reject Berofsky's reductionist approach. In appealing to a purely subjective conception of the true, it seems that on Ekstrom's account, an agent's preferences could include desires that they would *not* have under idealized conditions, but which are congruous with their (non-idealized) subjective conception of the true. That said, Ekstrom herself does not fully explain what it is for an agent to form a desire in the pursuit of their subjective conception of the good on her account; I shall attempt to rectify this below. However, although I believe that Ekstrom is right to implicitly reject Berofsky's reductionist account, I do not believe that the conception of the good that a rationalist theory of autonomy appeals to must be purely subjective, for reasons that I shall explain below. Here though, I shall press the point that Ekstrom's appeal to a purely subjective conception of the true is problematic.

²⁰ Ibid., 101.

²¹ Watson, "Free Action and Free Will," 150.

²² Berofsky, *Liberation from Self*, 85.

²³ Ibid., 80.

To explain why, consider cases of agents who are alienated from their motivating desires. On Ekstrom's account, such agents lack autonomy because they act in accordance with a first-order motivating desire that is not endorsed by a personally authorized preference; such agents do not believe that they have a reason to act in accordance with their first-order motivating desire, given their subjective conception of the good. Accordingly, Ekstrom can explain why agents who act on compulsive desires lack autonomy.

Notice that an agent's preferences must be rationally warranted on Ekstrom's theory, in so far as the agent must believe that they have reasons to have those preferences, reasons that are based upon their subjective beliefs concerning the good. However, in appealing to a purely subjective conception of the true, Ekstrom is denying that the autonomous agent's beliefs, notably including their beliefs about the good, must be in any way rationally warranted. To use the Parfitian terminology that I shall introduce below, Ekstrom thus claims that the autonomous agent must be practically, but not epistemically rational. However, this asymmetry is problematic, because it seems plausible to claim that some agents might lack autonomy with regards to their motivating desire, not because the motivating desire itself is incongruous with their subjective conception of the good, but rather because the agent's *beliefs* about the good are irrational, in the sense that the agent fails to meet certain epistemic standards in holding these beliefs.

To illustrate, consider sufferers of clinical depression. In many cases of clinical depression, the sufferer may have a suicidal desire that they personally authorize; but this authorization may stem from a belief about the disvalue of their own life that they

may hold irrationally.²⁴ For instance, such agents may not be able to offer any cogent reasons as to *why* they hold this belief about their own self-worth, or respond to any epistemic reasons that they are presented with to not hold such beliefs; it is simply a belief that they hold unshakably. It seems plausible to claim that such an agent lacks autonomy with respect to their suicidal desire; I suggest that just as we believe that an agent can lack autonomy if they are compelled by a motivating desire from which they feel alienated, so too can an agent lack autonomy if their endorsement of their motivating desire is based upon an irrational belief about the good. In cases such as the one I am considering here, it seems possible for agents to have compulsive evaluative beliefs, as well as compulsive first-order desires.²⁵

ii) *The Problem of Competing Desires and Coherence*

I have already explained that on Ekstrom's account, when agents have to decide which of two competing preference should win out and cohere with their other central preferences, in the absence of a neutralizing desire, an autonomous agent will decide that one preference defeats another on the basis that it is *more valuable* for her to prefer the object of desire *d* to the object of desire *g*.

A problem with this view is that it is unable to account for the possibility that an agent could be autonomous with respect to a desire to act in manner that they themselves believe to be sub-optimal. Consider the following example. Suppose that

²⁴ See Beck, *Depression*, 3. Notice that I am not claiming here that one *cannot* possibly hold such desires and beliefs rationally.

²⁵ I shall discuss similar cases in the final section of chapter eight.

Jim values having a career in philosophy, but also values spending more time with his family. Following a great deal of consideration, let us suppose that Jim decides that it would be slightly more valuable for him to prefer that his desire to spend more time with his family be effective in moving him to act. Would it really be the case that, having made this assessment about what is more valuable (and sticking to it), Jim would no longer be autonomous if he became motivated to instead pursue a career in philosophy? Admittedly, he would be doing so in the knowledge that he could be doing something else that he believed to be slightly more valuable; however, it still seems plausible to claim that Jim could nonetheless still be autonomous with respect to this decision.²⁶ Notice that this claim is compatible with the claim that Jim would have been *more* autonomous if he had chosen to act in accordance with what he believed to be his strongest reasons. The point that I am making here is that it is more plausible to make these two claims, rather than to rule out the possibility of Jim's autonomy here.

If one is not convinced by this example, imagine a case in which there was just *no* way of deciding which of two competing preferences *A* and *B* it would be more valuable for one to prefer; *A* and *B* might, for example, concern desires to pursue goods which are incommensurable. *Ex hypothesi*, there is no rational way of deciding which of *A* and *B* it would be more valuable to prefer. Paul Hughes has argued that when a person acts from volitional ambivalence like this:

²⁶ Sher discusses a similar example. Sher, "Liberal Neutrality and the Value of Autonomy," 143. See also Raz, *The Morality of Freedom*, 304.

. . . she is not autonomous either with respect to the desire that prompts her action or the action itself . . . (since) . . . in cases of volitional ambivalence there is no single conative ‘self’ directing the agent’s actions.²⁷

Pace Hughes, it is not clear to me why an agent would not be autonomous with respect to their action once they had elected to act in accordance with, say, preference A rather than preference B. Once the agent has plumped for A, it seems plausible to claim that they will be autonomous with respect to acting in pursuit of A *in so far as the preference for A is itself still rationally warranted*. Although A is no better or worse than B, this only means that the agent may lack a rational basis for their choice of A over B; but this does not mean that they lack autonomy with respect to their acting in pursuit of A, since that act itself is still rationally warranted.²⁸

This interpretation suggests another way of framing the problem with Ekstrom’s account that I am highlighting here. In the sorts of cases that I am considering here, it is not clear that the preference for one course of action has defeated or neutralized the preference for an alternative in the manner that is required for coherence on Ekstrom’s account; there are just *no* reasons that could undergird an agent’s claim that it was *more* valuable for her to have one preference rather than another. Yet it seems that one of these preferences could still cohere with her other central preferences and acceptances.

To conclude this section, I have suggested that there are two problems with Ekstrom’s theory as it stands. In order to respond to these problems, I suggest that we should revisit the building blocks of procedurally rationalist theories of autonomy, and

²⁷ Hughes, “Ambivalence, Autonomy, and Organ Sales,” 238–239.

²⁸ Sher, “Liberal Neutrality and the Value of Autonomy,” 144 makes a similar point.

re-consider the question of what it is for a preference to be formed in the light of one's evaluative judgments, or, we might say following Watson, what it is for a desire to be rationally warranted. In the next section, I shall delineate Parfit's work in this area from Part One of his *On What Matters*. Although Parfit's theory of rational desires was not developed as part of a theory of autonomy, I shall argue that his account can provide rationalist theories of autonomy with the conceptual apparatus to respond to the problems that I have raised here. In the following, I shall not provide an in depth defence of Parfit's arguments in favour of his objectivist account of rational desires, or his arguments against what he calls subjectivism.²⁹ Rather, I shall assume that Parfit's account is correct, since my aim is to investigate the implications that this account might have for a theory of autonomy.

III Parfit on the Rationality of Desires

According to one prominent understanding of what it is for a desire to be rational, a rational desire is one that is *causally dependent* upon beliefs that the agent has attained in accordance with appropriate rational standards. This seems to be similar to the view that Berofsky endorses in his reductionist motivational concept of value. However, Parfit rejects this view of rational desires since it conflates practical and epistemic rationality; whilst an agent's coming to have a belief without employing certain rational standards may render her *epistemically* irrational, Parfit points out that this irrationality need not transmit to her *practical* rationality.

²⁹ Parfit, *On What Matters*, Part One.

To illustrate this distinction, Parfit asks us to consider two cases. First, he asks us to consider an agent who has the irrational belief that smoking will improve her health, and who forms a desire to smoke on the basis of this belief. Second, he asks us to consider an agent who has the rational belief that smoking will damage her health, and who forms the desire to smoke on the basis of this belief. According to Parfit, the agent in the first case is epistemically irrational because she holds the belief that smoking will improve her health, despite the overwhelming evidence she has against the veracity of this claim.³⁰ However, her desire to smoke *given that she has that belief* is practically rational, since she wants “what, if (her) beliefs were true, (she) would have strong reason to want”.³¹ On the other hand, although the agent in the second case holds an epistemically rational belief, she is practically irrational because she has strong reason not to want to smoke, in so far as she believes that doing so will damage her health.

Parfit’s assessment of these cases is indicative of his endorsement of an objectivist account of reasons. According to such an account, it is not the agent’s subjective desires that give her reasons to act in certain ways; rather it is facts about the *object* of the agent’s desire that provide her with reasons to act. On the objectivist account, there are facts that make certain outcomes worth pursuing and that “ . . . give us reasons both to have certain desires or aims, and to do whatever we can to fulfil them”.³² One of Parfit’s main arguments in defence of endorsing an objectivist account is that on subjectivist theories, we have no basis for explaining why an agent who has no desire to avoid a period of agony in the future after ideal deliberation is being

³⁰ Ibid., 115.

³¹ Ibid., 113.

³² Ibid., 45.

practically irrational; this, Parfit claims, is surely implausible.³³ We might notice here that Berofsky's reductionist approach to value is a subjectivist account on Parfit's schema, since it reduces values to that which an agent would desire under certain ideal conditions of deliberation.

With the above discussion in mind, it should be no surprise that Parfit makes the following claim:

What makes our desires rational or irrational is not the *rationality* of the beliefs on which these desires causally depend, but the *content* of these beliefs.³⁴

On this view, the rationality of a desire depends on whether the belief upon which our desire causally depends concerns facts that make the desire's outcomes worth pursuing; the reasons for having such desires are *object-given*. It should be acknowledged that the appellation 'object-given reasons' does not imply that such reasons must be objective, in the sense that Berofsky suggests is problematic on Watson's view. To see why, it is useful to draw a distinction between what we may call *personal* self-interested reasons and *impersonal* self-interested reasons.³⁵

³³ Ibid., 73–74.

³⁴ Ibid., 113.

³⁵ This choice of terminology is somewhat confusing, since Parfit himself uses the term 'impersonal' to describe a type of goodness that contrasts with goodness for a *particular* person. Ibid, p. 41. However, I believe this confusing choice of terminology is unavoidable, since Parfit also uses other terms that one could plausibly use to clearly make the distinction that I draw above for other purposes. For example, he uses the term 'impartial reasons' to refer to reasons that we have to care for *anyone's* well-being. Ibid.,

According to Parfit, an outcome is worth pursuing for a particular person if “. . . there are certain facts that give this person self-interested reasons to want this event to occur”.³⁶ In turn, ‘self-interested reasons’ are reasons provided by facts concerning the person’s well-being.³⁷ Parfit seems to endorse the view that all agents share certain self-interested reasons. As I mentioned above, his main argument against subjectivist accounts of reasons is that such accounts deny that we have any reason to want to avoid some period of future agony if we did not want to avoid it despite having full awareness of the relevant facts.³⁸ This seems to suggest that Parfit endorses a theory of well-being that incorporates objective elements; on such a view, we all have some self-interested reason to avoid agony because it is simply bad for us to be in the conscious state of having a sensation that we dislike.

Notice though that even on a theory of well-being that incorporates *only* objective elements, agents may differ with regards to what they have self-interested *instrumental* reasons to want. To illustrate, assume that agents have self-interested reasons to want to be in the conscious state of having a pleasurable experience. On this assumption, although all agents might share the (telic)³⁹ self-interested reason to want this outcome, agents will not necessarily achieve this goal in the same ways. For instance, if the sensation of eating ice-cream were pleasurable for Ben, then this fact would give Ben a self-interested reason to want to eat ice-cream; however, if the

40. Furthermore, he uses the term ‘objective’ to refer to the theory that facts concerning the objects of our desires give us reasons. Ibid. 45.

³⁶ Ibid., 41.

³⁷ Ibid., 39–40.

³⁸ Ibid., 73–82.

³⁹ Ibid., 44.

sensation of eating ice-cream were painful for Jerry (say because he has tooth-ache), Jerry would not have a self-interested reason to eat ice-cream.

Although Parfit seems to endorse an account of well-being that incorporates objective elements, he also explicitly points out that his objectivist theory of reasons is compatible with *subjective* accounts of welfare.⁴⁰ I shall consider this in greater detail in chapter five. At this point though, I suggest that we call those reasons that are understood to depend on facts concerning objective elements of an agent's well-being their '*impersonal* self-interested reasons'. In contrast, I shall call those reasons that are understood to depend on facts concerning subjective elements of well-being (including those contingent facts about what particular individuals have instrumental reasons to want on purely objective theories of well-being) '*personal* self-interested reasons'. Notice though, that what I call personal self-interested reasons are still object-given on the Parfitian account. For the remainder of this thesis, I shall be concerned primarily with these two types of self-interested reasons. Although I leave open the possibility that an agent's desires might be rationally warranted by other sorts of reasons (such as impartial reasons that they might have to care about the well-being of others), it seems that in the context of *personal* autonomy, an agent's self-interested reasons will be particularly salient.

At this point, it is important to acknowledge the distinctions that Parfit draws between normative and motivating reasons, and real and apparent reasons. When we have normative reason to *x* there are facts about *x* that give one sufficient reason to pursue the outcome of that desire; one has a *normative* reason regardless of whether one

⁴⁰ Ibid., 74. See also Parfit, *Reasons and Persons*, Appendix I for different accounts of well-being.

is aware of these facts. Such facts concern what Parfit calls “real reasons”.⁴¹ To illustrate, suppose that Shelia is about to drink the liquid in a glass that she believes to contain gin, but which in fact contains acid.⁴² Here, Sheila has *normative* reason to *not* drink the liquid in the glass because it will harm her. On the other hand, the motivating reasons an agent has depend on what the agents believes. For example, although Shelia does not have a normative reason to drink the liquid in the glass handed to her, she has a motivating reason to do so because she believes (rationally, but incorrectly) that drinking the liquid would serve as a means to an end that she values (quenching her thirst). In this case, Shelia’s motivating reasons do not track her normative reasons.

Shelia’s case raises the point that agents often lack epistemic access to certain normative reason-giving facts. In such scenarios, agents have to decide how to act on the basis of their “apparent reasons”,⁴³ that is to say the reasons that they understand themselves as having *given* their beliefs. Whether or not these apparent reasons amount to real reasons depends on the truth status of the beliefs upon which the apparent reason causally depends. If the beliefs are true, the apparent reason is also a real reason; if not, the apparent reason is what Parfit terms “merely apparent”.⁴⁴ With this framework in mind, we can understand why Parfit claims that the smoker who wants to smoke in order to improve her health has a rational desire (despite being epistemically irrational), whilst the smoker who wants to smoke in order to damage her health has an irrational

⁴¹ Ibid, 35.

⁴² I adapt this from Williams, *Moral Luck*, 102. Notice that Williams uses this example in defence of his thesis that agents have only what he calls internal reasons; this is a subjectivist thesis about reasons of the sort that Parfit rejects. For Parfit’s specific comments on Williams’ view, see Parfit, *On What Matters*, 65 and 77.

⁴³ Ibid., 35.

⁴⁴ Ibid.

desire (despite being epistemically rational). The content of the former's belief gives the agent a strong, but merely apparent, reason to have the corresponding desire, whilst the content of the latter's belief gives the agent a strong and real reason *not* to have the corresponding desire.

Accordingly, we can also understand the way in which reasons can be weighed against each other on Parfit's account. Parfit himself explains this in terms of normative reasons, and I shall do the same; however, one could give an account of how to weigh one's apparent reasons in a similar fashion by substituting the term 'reason(s)' below with the phrase 'apparent reason(s)'.

According to Parfit:

If our reasons to act in some way are stronger than our reasons to act in any of the other possible ways, these reasons are *decisive*, and acting in this way is what we have *most reason* to do. If such reasons are much stronger than any set of conflicting reasons, we can call them *strongly* decisive.⁴⁵

In view of this he goes on to offer the following schema of how we might classify the rationality of some possible act of ours. According to Parfit, some possible act of ours would be:

rational if we have beliefs about the relevant facts whose truth would give us sufficient reasons to act in this way,

⁴⁵ Ibid., 32.

what we *ought rationally* to do if these reasons would be decisive,

less than fully rational if we have beliefs whose truth would give us clear and decisive reasons not to act in this way,

and

irrational if these reasons would be strongly decisive.⁴⁶

IV Incorporating Parfit's Account into A Rationalist Theory of Autonomy

I suggest that Parfit's account of rational desires can be used to enrich Ekstrom's account of autonomy so that it is able to respond to the two problems that I raised in section II. I shall consider each in turn.

I suggested above that Ekstrom implicitly seems to endorse an evaluation conception of value rather than Berofsky's reductionist approach. This is fortunate, since if Parfit's agony argument is a decisive argument against what he terms

⁴⁶ Ibid., 34.

subjectivist theories, it seems that it would also be decisive against Berofsky's reductionist conception of value. However, Ekstrom herself does not spell out in detail what it is for an agent to form a desire in the pursuit of her subjective conception of the good. This is problematic, since one might worry that Ekstrom's theory is vulnerable to Parfit's objections to subjectivism that I surveyed above. In order to make it clear that Ekstrom's theory is not vulnerable to these objections, I suggest that when Ekstrom claims that preferences are desires that are formed in pursuit of the good, it is best to cash this out in the following way:

Parfitian Preference Formation: When forming a preference in the pursuit of the good, the agent forms their preference on the basis of a belief, whose truth, would make the object of their first-order desire good in a reason-implying sense.

On this view, the rationality of an agent's preference will depend on her beliefs about facts concerning what she has reasons to do given her understanding of the good; and these reasons do not arise simply because she forms the preference under certain idealized conditions as the reductionist would claim.

Notice that this formulation does not entail that autonomous agents must always choose to do what they believe is *objectively* good in the way that Berofsky finds problematic. On Parfit's view, an agent's self-interested reasons are provided by facts about the agent's welfare. Whilst agents can have what I called impersonal self-interested reasons if one endorses a theory of well-being that incorporates objective elements, I also suggested that agents can also have what I called personal self-interested reasons. One can have personal *instrumental* self-interested reasons even on

purely objective theories of well-being. Crucially though, as Parfit himself explicitly points out, his objectivist view of reasons is compatible with theories of well-being that incorporate subjective elements.⁴⁷

Furthermore, even on purely objective theories of well-being, it seems plausible to claim that there could be some scope for subjectivity, in so far as it seems plausible to claim that rational agents can disagree to a significant extent about the weight they assign to different objective goods. Indeed, Parfit himself claims, “(t)hough there are truths about the relative strength of different reasons, these truths are often very imprecise”⁴⁸. Indeed, in his discussion of Frankfurt’s views on rationality, Parfit suggests that the following are plausible claims:

. . . there are many intrinsically good ends, but no ends have supreme value. Nor are there precise truths about which ends are most worth achieving. We often have to choose between many good ends or aims, none of which is clearly better than the other, and in such cases there is no end that reason requires us to choose.⁴⁹

I shall return to these issues in chapter five. At this point though, I shall now turn to Ekstrom’s account of acceptances. I suggested above that Ekstrom’s claim that an agent’s acceptances need only be developed in light of their own subjective conception of the true is problematic, because agents might lack autonomy with respect

⁴⁷ Ibid., 74.

⁴⁸ Ibid., 33.

⁴⁹ Ibid., 100.

to a desire, not because it is incongruous with their subjective conception of the good, but because the agent's *beliefs* about the good are irrational. In view of the Parfitian analysis given above, it might be thought that we should instead say that agents must develop their acceptances in an epistemically rational manner. However, such a claim seems to be too demanding a condition of autonomy, since agents will often fail to be fully epistemically rational in their beliefs. Accordingly, we should phrase the relevant condition here in a negative sense, as follows:

Parfitian Acceptance Formation: an agent's belief will only qualify as one of their acceptances if they form it in a manner that is not epistemically *irrational*.

Parfit claims that the rationality of a belief will depend on, *inter alia*, its relation to our other beliefs, and also its relation to our sensory evidence. He also claims that the rationality of our beliefs depends on whether, in holding them, we are “. . . responding well to epistemic or truth-related reasons or apparent reasons to have these beliefs”.⁵⁰ In view of these brief remarks, we might say that an agent fails to meet a minimum threshold of epistemic rationality, or that they are epistemically irrational, if they form a new belief *x* in a manner that is unresponsive to the apparent reasons that they have (given their pre-existing beliefs) to believe or not believe *x*. Moreover, we might add that an agent will be epistemically irrational if they form a new belief *x* which they themselves realise is incompatible with one of their pre-existing beliefs *y*, without ceasing to believe *y*.

⁵⁰ Ibid., 117.

By incorporating these two supplementary conditions concerning preference and acceptance formation, I have explained how an amended version of Ekstrom's theory can answer the first set of worries I raised in section II. I shall now explain how a revised version of Ekstrom's theory should respond to cases of competing desires in view of Parfit's account of rational desires.

To begin, let us return to my earlier example of Jim. Recall that following deliberation, Jim decided that it would be slightly more valuable for him to spend more time with his family than to pursue a career in philosophy given his character system at the point of decision. In Parfit's terms, Jim believes that his reasons to spend more time with his family are *decisive*, although not *strongly* so. In this situation, it seems that if Jim chose to pursue a career in philosophy then he would not be choosing in what Parfit calls a 'fully rational manner'. Jim's decision to pursue a his career would be less than fully rational here, because he believes that he has a clear and decisive reason to pursue an alternative outcome, and pursuing this career would prevent him from pursuing that alternative outcome. I suggested above that we could nonetheless plausibly believe that Jim could be autonomous in acting this way.

As Sher points out, if this is right, then it seems that what matters for autonomy is that the agent acts in the light of a reason that is *sufficiently* strong to warrant action; it does not need to be the agent's strongest reason. Autonomy then seems to be compatible with what Parfit terms being 'less than fully rational', but not with practical *irrationality*. As I explained above, in contrast to a possible act's being less than fully rational, Parfit claims that a possible act is irrational if we have beliefs about the relevant facts whose truth would give us a clear and *strongly* decisive reason not to act in this way.

It should be acknowledged that since we normally act on the basis of apparent reasons, it seems that our weighting of reasons will often involve a consideration not only of the goodness of the outcomes at stake, but also the *probability* that our act will achieve that outcome. For instance, suppose that Beth is considering whether to go base-jumping. Although she believes that she would enjoy the experience, she is also aware that there is a chance of something going catastrophically wrong, leading to her death. In order to establish whether it would be rational for Beth to go base-jumping, we would need to assess both the comparative value of the possible projected outcomes of her going base-jumping or not, and the probability of each outcome occurring given certain courses of action. In this case, it might be rational for Beth to go base-jumping if she believed that the probability of the catastrophic outcome was sufficiently low, and if she believed that the value of the experience of base-jumping would be sufficiently high.

This discussion helps to explain why it is perhaps simplistic to suggest that the autonomous agent must always act in accordance with their strongest reasons, as Ekstrom's theory seems to suggest. An agent can be autonomous without being *fully* rational, and they can act in ways that put them at small risk of outcomes that they have strongly decisive reasons to avoid, if that act is also more likely to lead to an outcome that they have reason to want. Furthermore, as my discussion above suggested, agents can rationally disagree about the weight they assign to the same reasons. Accordingly, I conclude that coherence, in Ekstrom's terms, should not be viewed as simply a matter of which preferences it is *more valuable* for an agent to have, given her character system at the point of decision. Rather, when two preferences compete, we should allow for the possibility that preferences which it would be less than fully rational, but not irrational, for the agent to have can defeat a competing preference that she has stronger

reason to have, in view of her beliefs about the good. It is not only the *most* rational desires that can be said to cohere with the agent's character system.

As such, we might re-define coherence as follows:

Df, coherence (II): a preference for desire d coheres with the character system of S at t if and only if, for any competing preference for desire g , it is either (i) *not irrational* for S to prefer that desire d than it is for her to prefer that desire g , on the basis of the character system of S at t , or (ii) as valuable for S to prefer that the conjunction of g and a neutralizing desire n as it is for S to prefer that g alone on the basis of character system of S at t .

V Responding to the Problem of Manipulation

To conclude this chapter, I shall explain how a revised version of Ekstrom's theory can respond to the problem of manipulation that plagued the theories of autonomy that I considered in the previous chapter.

The Parfitian account of preference formation that I outlined above suggests a way to identify those interventions that might appropriately be termed manipulative. First, we may say that an agent has been manipulated if, as a result of third party interference, they sustain a motivating desire that they do not personally authorize. However, manipulation may occur at a deeper level. I suggested above that when an agent forms a preference in the pursuit of the good, we should understand the agent as forming a preference on the basis of a belief, whose truth, would make the outcome that

is the object of her first-order desire good in a reason-implying sense. Notice that on this account, there is a cognitive element to preference formation; our preferences are desires that we form in the light of our (non-irrational) *beliefs* about the good. Accordingly, I suggest that we define deeper manipulative interventions as those that intervene in the agent's decision-making process to sustain a certain preference in a manner that serves to bypass the cognitive element of their decision-making process. The intervention thus serves to ensure that the agent will not make their decision in the light of their (non-irrational) *beliefs* about the good.

However, we might wonder if Ekstrom's theory will, like Mele's historical theory, prove too much. As I discussed in the previous chapter, Mele's historical account stipulates that a necessary condition of an agent being autonomous with respect to a desire is that she was not *caused* to have that desire in a way that bypasses her mental control. I claimed that a major problem with this theory is that we seem to form many of our desires in ways that bypass our mental control. However, it seems that a similar charge could be raised against Ekstrom's theory, since she claims that autonomous agents must endorse their first order motivating desires with a preference that they have *formed* in the pursuit of the good.

Fortunately, this is not a serious problem for rationalist theories in the same way that it is for historical theories. Instead of claiming that preferences must be *formed* in the pursuit of the good, rationalist theories should instead claim that a preference can constitute part of an agent's character system *whatever* its causal history, as long as the agent herself now consciously makes a (non-manipulated) choice to *sustain* that preference, a choice based on the fact that it coheres with her other preferences and

acceptances.⁵¹ Of course, the preferences we *form* on the basis of our beliefs about the good are perhaps more likely to cohere with our other preferences and acceptances; however, it seems plausible to claim that I can come to rationally embrace a preference that has a dubious causal history. This, I take it, is the point underlying Berofsky's objection to historical theories of autonomy, when he claims that an agent could reject the aetiology of a particular desire, and yet still hold that desire autonomously if she endorses the desire on other grounds.

Now, this might mean that our autonomy is to some extent limited. We might worry that if we later consciously choose to sustain a preference that we have developed uncritically, we might only be doing so because it coheres with other preferences and acceptances that we have also developed uncritically. Indeed, the problematic cases of BNWs and Jehovah's Witness that I considered in chapter two seem to fit this description.⁵² Does this not undermine our autonomy in a fundamental and unavoidable manner? The key to replying to this worry is acknowledging that the acceptances that partly constitute our character system will include our *beliefs about the good*; and these particular acceptances are, I suggest, integral to our character systems, since they also undergird the rationality of our preferences.

As I argued above, it seems that our acceptances must reach some minimal threshold of epistemic rationality; in order for this to be the case, our acceptances must be responsive to the apparent epistemic reasons that we have to hold (or reject) those beliefs. In view of this, I suggest that even if all agents initially develop their

⁵¹ Notice that one may thus be autonomous with respect to a desire even if one did not critically reflect upon it at the time at which it was formed. See Savulescu, "Rational Desires and the Limitation of Life Sustaining Treatment," 199; Young, *Personal Autonomy*, 8 for defences of a similar position.

⁵² Chapter One, 43-44.

acceptances about the good in an uncritical manner, these acceptances can still undergird their autonomy, as long as the agent later chooses to sustain those acceptances in a manner that is responsive to their apparent epistemic reasons. However, if they sustain them dogmatically and without reflection, then I suggest that these acceptances cannot be said to undergird autonomous agency.⁵³

It might finally be argued that the rationalist theory that I have defended here is still vulnerable to the problem of manipulation in one way, since it seems possible that an agent could be manipulated into having a completely *different* set of coherent preferences and acceptances; or in other words, a different character system. The objection here is that if autonomy only requires that one's preferences and acceptances cohere, then an agent who undergoes such a universal manipulation would be autonomous; and this, it seems, is implausible.

In response to this charge though, we might point out that if an agent has been manipulated into completely changing their whole character system, then it is not clear that they would in fact be the *same* agent.⁵⁴ On certain psychological approaches to personal identity, an agent's identity over time is only maintained if their mental states at time $t+1$ are psychologically contiguous (in the right causal way) to their prior mental states at t .⁵⁵ In the extreme manipulation case that I am considering here, this does not seem to be the case. Accordingly, we might concede that such agents could be autonomous following their manipulation, but point out that the agent in question is not

⁵³ These reflections explain how my account is compatible with a relational view of the autonomous agent.

⁵⁴ Taylor suggests that Ekstrom could make this sort of response. Taylor, "Introduction," 15.

⁵⁵ Parfit, *Reasons and Persons*, 204–209. For other psychological accounts, see Shoemaker, "Personal Identity: A Materialist's Account"; Noonan, *Personal Identity*.

the same one as the agent who was manipulated. This, I suggest, is the best way to interpret the case of Phineas Gage that I delineated in the introduction; the problem pertains to the agent's identity across time, rather than their autonomy *per se*.⁵⁶

One might worry that this response to the extreme manipulation case relies on the endorsement of a psychological account of personal identity that is not universally accepted.⁵⁷ I lack the space to adequately defend such an account here. However, it should be acknowledged that this response does not rely on a psychological account of personal identity in the *numerical* sense; all that this response requires is that certain psychological continuities are integral to identity in some valuable sense; yet this need not be numerical identity. For instance, as I delineated in chapter one, when Bratman claims that autonomy is inextricably linked to the agent's personal identity, I suggested that we should understand Bratman to be appealing to the concept of *narrative*, rather than numerical, identity. It seems highly plausible that even if certain psychological continuities are not necessary for numerical identity, they will still be so for the agent's narrative identity; and this is all that my reply to the objection under consideration requires.

Finally, we might also observe that those who reject the claim that psychological continuity is a necessary element of identity in *any* valuable sense owe us an account of why extreme manipulation cases of the sort that I have considered here are problematic. In view of the arguments of this section, and in view of my objections to the way in which historical theories of autonomy reply to the problem of manipulation, this will not be an easy task.

⁵⁶ Introduction, 11.

⁵⁷ For example, see Carter, "How to Change Your Mind"; Snowdon, "Persons, Animals, and Ourselves"; Olson, *The Human Animal* (especially Chapter Three).

Conclusion

This analysis concludes my investigation of the reflective dimension of autonomy. I have followed Ekstrom in claiming that an agent is autonomous with respect to her motivating desires if those desires are endorsed by a preference that coheres with their central preferences and acceptances. However, by incorporating a Parfitian account of rational desires into conditions pertaining to the nature of preferences, acceptances, and coherence, I have suggested that it is possible to overcome some of the problems faced by Ekstrom's original theory. Moreover, they can help explain how a rationalist theory can overcome the problem of manipulation.

In view of this, I am now in a position to offer the following necessary condition for what it is to be reflectively autonomous to a minimum threshold level with respect to one's first-order motivating desire:

Minimum threshold condition of reflective autonomy: An agent is minimally autonomous with respect to her first order motivating desire if she has a preference to pursue the object of that desire which:

- i) . . . she sustains on the basis of a non-irrational belief that the object of that desire is good in a reason-implying sense.

And

- ii) . . . coheres with her character system.

In the next chapter, I shall consider the practical dimension of autonomy, which concerns the agent's ability to act effectively in pursuit of their ends.

Chapter Three – The Practical Dimension of Autonomy

In the previous chapter, I proposed an account of the reflective dimension of autonomy. However, as I suggested in the introduction, the concept of autonomy (as it is invoked in contemporary bioethics) also seems to incorporate a practical dimension, pertaining to an agent's ability to act effectively in pursuit of their ends. My aim in this chapter is to explain both what it is for an agent to be practically autonomous in this way, and how this dimension of autonomy relates to the reflective dimension of autonomy.

I shall begin by defending the claim that an adequate theory of autonomy should incorporate conditions pertaining to the practical dimension of autonomy. In section two, I shall consider some prominent understandings of the nature of freedom, before going on to suggest, in section three, how much freedom an agent must have in order to be minimally practically autonomous. In section four, I shall argue that practical autonomy requires holding certain true beliefs, and I shall consider the implications that this claim has for how we should understand deception to undermine autonomy. Finally, in section five, I shall explain an important relationship between the reflective and practical dimensions of autonomy, and argue that an agent's beliefs about what they are free to do can have an important influence on their decisions.

I Introducing The Practical Dimension of Autonomy

In the introduction to this thesis, I pointed out that the bioethical principle of respect for autonomy incorporates a negative obligation that enjoins us to not restrain the autonomous actions of others. As Brock points out:

. . . interference with self-determination can involve interference with people's deciding for themselves, *but also interference with their acting as they have decided they want to act*".¹

In the introduction, I claimed that this negative obligation suggests that there is a practical dimension to the concept of autonomy as we understand it in bioethical discussion, a dimension that pertains to the agent's ability to act effectively in pursuit of their ends.²

Some initial clarifications of this point are necessary. First, in claiming that autonomy requires that agents must be able to act effectively in pursuit of their ends, I do not mean to claim that they must be *successful* in their endeavours; one can of course fail to achieve one's ends and yet still be autonomous. Rather, the point undergirding the practical dimension of autonomy is that being unable to act effectively in pursuit of one's ends is inimical to one's autonomy all things considered, if we understand 'being able to act effectively' to simply mean that an agent is able to act in a manner which has

¹ Brock, *Life and Death*, 29. Emphasis added.

² Introduction, 15.

some positive influence on her pursuit of the goal that she autonomously wants to achieve. I shall say more about this below.

Furthermore, at this point it should also be acknowledged that the reflective dimension of autonomy is theoretically prior to the practical dimension. If an agent lacks reflective autonomy with respect to their decision about what to do, then they still lack autonomy all things considered, even if they have the freedom to act effectively on the basis of that non-autonomous decision. For the purposes of this chapter, I shall assume that the agents I discuss are reflectively autonomous with respect to their decisions.

As I pointed out in the introduction, philosophers tend to be wary of the claim that an overall theory of autonomy should incorporate conditions pertaining to the practical dimension of autonomy.³ However, the failure to recognise the importance of the practical dimension of autonomy leads to an impoverished discussion of autonomy in bioethics. Three points speak in favour of this view. The first follows on from my above discussion of the principle of respect for autonomy; the way in which we use the concept of autonomy in bioethical contexts suggests that we implicitly understand it to incorporate a practical dimension. If we believe that a theory of autonomy for use in contemporary bioethics should be congruous with our use of the concept in that context, then it seems that our theory of autonomy should accommodate a dimension of the concept that it is implicitly understood to incorporate in our bioethical discussions.

The second point in favour of this view is that acknowledging the practical dimension of autonomy seems to be necessary if we are to account for the high prudential value that we place on autonomy. I shall consider the value of autonomy in

³ See Introduction, note 32.

greater detail in chapter five; at this point though we might observe that there would seem to be little prudential value in being reflectively autonomous with respect to one's motivating desires if one was perpetually frustrated in one's attempts to pursue one's ends. If we believe that autonomy bears high prudential value because we have a fundamental interest in 'living a life that is our own' (as I shall claim in chapter five), then it seems that we should be able to *act* effectively on the basis of our decisions, as well as making those decisions in a reflectively autonomous way.

The third and strongest point in favour of this view is one that I shall develop over the course of this chapter. To put it simply here though, if we fail to acknowledge the practical dimension of autonomy in our overall theory of autonomy, it is not clear that we can adequately account for the way in which our beliefs about what we are free to do can have crucial effects on our choices; *we choose to sustain our motivating desires in the light of our beliefs about what is practically realisable*. In order to account for this crucial theoretical point, I claim that an adequate theory of autonomy cannot ignore considerations pertaining to the practical dimension of autonomy.

To begin my investigation into the practical dimension of autonomy, it is important to consider different understandings of liberty or freedom (like Berlin, I shall use the terms interchangeably).⁴ The reason for this is that if an agent is to be able to act effectively in pursuit of their ends, it seems clear that they will need to have certain sorts of freedoms.

⁴ Berlin, "Two Concepts of Liberty," 34. See Pitkin, "Are Freedom and Liberty Twins?" for a discussion of ways in which one might distinguish between the two terms.

II Positive and Negative Freedom

Following Berlin, it is commonly claimed that there are two separate understandings of freedom. It is claimed that freedom understood as the absence of constraint represents a negative conception of freedom. Negative freedom may broadly be construed as freedom from interfering external forces that prevent the agent from acting. Berlin captures this thought in his claim that we use the concept of negative freedom in answering the following question:

What is the area within which the subject — a person or group of persons — is or should be left to do or be what he is able to do or be, without interference by other persons? ⁵

Sometimes though, we are unable to pursue an end, not because we are being restrained from doing so, but rather because we lack certain abilities. For instance, consider an individual with locked-in syndrome who wishes to end their own life; suppose further that no-one would restrain the individual from ending their own life if she were able to do it herself. In this scenario, we might say that such an individual seems to lack practical autonomy because she lacks an ability that is necessary for acting in pursuit of her chosen end.

Cases such as these suggest that we also have a positive conception of freedom, in which freedom is constituted not by the absence of restraint, but rather by the

⁵ Berlin, "Two Concepts of Liberty," 34.

presence of capacities or conditions that enable the agent to be effective in the pursuit of their ends.⁶ Berofsky captures this sort of thought in his claim that positive freedom is constituted by those personal traits that are “ . . . essential or highly useful to the satisfaction of a wide range of activities and decisions”.⁷

In my view, Berofsky’s characterisation of positive freedom is too broad. Whilst it is true that many abilities are generally useful for the pursuit of a wide range of goals, an agent’s ability to pursue her ends may require very specific freedoms that are not essential to widely pursued activities. For instance, although having 20/20 unaided vision is not necessarily useful for the pursuit of a *wide* range of goals, a person with slightly impaired vision who wants to become a military fighter pilot nonetheless lacks a physical trait that means that they are precluded from achieving their goal. Accordingly, agents can lack freedoms that are important for the pursuit of *their* goals but that are not essential for the pursuit of a wide range of activities.

Conversely, agents might lack freedoms that are important for the pursuit of a wide range of goals, and yet still have the freedom to act in pursuit of what it is that they want to do. To illustrate this, consider the case of Epictetus. Despite his being enslaved, and thus lacking many freedoms, Epictetus was nonetheless free to pursue his goal of living a life of philosophical reflection.⁸ In view of these observations, *pace* Berofsky, I shall claim that an agent’s positive freedom is constituted by those traits and

⁶ Of course, Berlin famously understood positive freedom in a broader sense; however, as Miller points out, Berlin’s concept of positive freedom incorporates “a number of quite different doctrines” (Miller, “Introduction,” 10). In order to avoid a lengthy exegesis of Berlin’s essay here, I shall instead consider Berofsky’s narrower conception of positive freedom.

⁷ Berofsky, *Liberation from Self*, 16.

⁸ I thank Jeff McMahan for this example.

capacities that she requires in order to pursue an end that she herself is motivated to achieve.

Although the distinction between positive and negative freedom is widely adopted, it is somewhat problematic. As Berofsky points out, it may often be unclear whether some factor is an element of positive or negative freedom; for example, he points out that it is not clear whether we should understand intelligence as a constituent of positive freedom, or stupidity as a barrier to negative freedom.⁹ Feinberg has also questioned the utility of the distinction, arguing that we can have a comprehensive understanding of freedom as being constituted by freedom from preventative causes, given a sufficiently nuanced analysis of preventative causes, and the constraints to which they give rise. Feinberg suggests that we should analyse preventative causes as giving rise to the following two sorts of constraint:

- 1) A negative constraint = A preventative cause constituted by the absence of some enabling factor.

- 2) A positive constraint = A preventative cause constituted by the presence of some debilitating factor.¹⁰

⁹ Berofsky, *Liberation from Self*, 42. See also MacCallum, "Negative and Positive Freedom."

¹⁰ See Feinberg, *Freedom and Fulfillment*, 5–6.

Once these distinctions are made, Feinberg claims that we obviate the need for a distinction between positive and negative freedom; freedom is just constituted by freedoms from different sorts of constraints.¹¹

I am sympathetic to Feinberg's arguments here (although space does not allow for a defence of them). Nevertheless, in light of its prevalent use, I believe that the clarity of the following discussion will be best served by adhering to the vocabulary of positive and negative freedom. However, I shall use this language in the attenuated sense that Feinberg suggests is "harmless",¹² whereby positive freedom is characterised as the absence of a negative constraint, and negative freedom is characterised as the absence of a positive constraint.

III **Autonomy and Freedom at the Point of Action**

The question of how much freedom autonomy requires is a complex one, not least because of the difference between the two conceptions of freedom identified in the previous section. A further difficulty arises due to the fact that the question can be raised at two salient points.¹³ First, we might raise it at what we may term 'the point of action', when the agent has *already* decided to act in some way. Raised at this point, the question of freedom is primarily relevant to the practical dimension of autonomy. However, the question may be raised prior to the point of action, at what we might term

¹¹ This sentiment is shared by MacCallum's account in MacCallum, "Negative and Positive Freedom."

¹² Feinberg, *Freedom and Fulfillment*, 7.

¹³ This distinction maps onto Berofsky's distinction between freedom of action and freedom of decision. Berofsky, *Liberation from Self*, 26–27.

the ‘point of decision’, that is, prior to when the agent has decided what it is she will do. Raised at this point, an agent’s beliefs about what she is free to do may also impinge on the reflective dimension of their autonomy, as I shall go on to explain.

As such, in order to answer the question of how much freedom autonomy requires, we must carry out two different investigations. In this section, I shall begin by considering how much freedom is required at the point of action for an agent to be able to act effectively in pursuit of their ends. I shall consider how much freedom may be required at the point of decision in section V.

We can begin by observing that practical autonomy surely cannot require absolute negative freedom to do *anything* at the point of action, since we can be positively constrained from doing something without that constraint being inimical to our ability to achieve our ends. For instance, consider this example:

Harry has been asked by Jane to look after her dog. Suppose that Harry would instead like to visit the nearby pub. However, Harry decides to stay and look after the dog because he wants to prove his dependability to Jane. Now, suppose that Jane locks Harry in the house with the dog, because she is aware that Harry will have spotted the pub on his way in. However, Harry does not realise he is locked in, having *already* resolved to stay in the room looking after the dog.¹⁴

In this example, Harry is positively constrained from leaving the room. However, although he seems to lack a significant negative freedom, Harry still has the negative

¹⁴ This is a Lockean variant of a so-called Frankfurt example. See Frankfurt, “Alternate Possibilities and Moral Responsibility” and Locke, *An Essay on Human Understanding*. Book II, Chapter XXI.

freedom to *do what he is motivated to do*; he is not positively constrained from looking after the dog. Now, although it might be right to claim that Harry would have greater freedom if he were not locked into the room, it seems peculiar to claim that he would be more able to effectively pursue his end. Assuming, as is the case in the above example, that prior knowledge of a lack of negative freedom is not impinging on Harry's decision about what to do, his lacking the freedom to do something which he has already resolved *not* to do does not seem to reduce his autonomy in any way.

The contrast between freedom at the point of action and autonomy is also highlighted by cases in which agents sacrifice certain negative freedoms as an *expression* of their autonomy. To illustrate this, we may follow Dworkin in considering the case of Odysseus and the Sirens. Dworkin describes the case as follows:

Not wanting to be lured onto the rocks by the sirens, (Odysseus) commands his men to tie him to the mast and refuse all later orders he will give to be set free. He wants to have his freedom limited so that he can survive.¹⁵

Here, if the crew removed the positive constraints preventing Odysseus from leaving the ship, this would hinder Odysseus' ability to pursue his goal of hearing the sirens' song without being lured from his ship. Accordingly, Odysseus' case suggests that agents may autonomously decide to limit their negative freedom to do certain things, if having such freedoms would hinder their pursuit of their goals. Far from enhancing his autonomy, freeing Odysseus whilst the ship sailed past the sirens would have been inimical to it.

¹⁵ Dworkin, *The Theory and Practice of Autonomy*, 14-15

The cases of Odysseus and Harry suggest that what is important with regards to the negative freedom that practical autonomy requires at the point of action is not the number of options that one has the negative freedom to pursue, but rather whether one has the negative freedom to pursue the end that one has decided to pursue; we may say that in order to be able to act effectively in pursuit of their end, an agent cannot be positively constrained from doing so.

Of course, there are perhaps some limits to this; for instance, we might claim that we should not allow an agent the negative freedom to completely abandon their future negative freedom, by selling herself into slavery say. One particularly salient problem with this is that in doing so, the agent abdicates their negative freedom to act in accordance with a *future* desire that they might develop to not live as a slave.¹⁶ In view of this, whilst we might claim that respecting the agent's locally autonomous decision here requires that we do not positively constrain her from becoming a slave, we might still positively constrain her from doing this in the name of her *global* autonomy. This, I suggest, is a case in which respecting local and global autonomy might require different things of us; I shall consider other such cases in chapter five.

Like negative freedom, it seems that lacking certain positive freedoms need not always be inimical to our practical autonomy. After all, we all lack certain capabilities, but this does not necessarily preclude the possibility of our practical autonomy. Most obviously, some freedoms are just irrelevant to ability to pursue our ends. For example,

¹⁶ This is how Dworkin explains the wrongness of selling oneself into slavery. See Dworkin, "Paternalism". For other discussions, see Mill, "On Liberty," 116; Sneddon, "What's Wrong with Selling Yourself into Slavery?". Those who endorse substantive theories of autonomy might claim that an agent simply cannot autonomously desire to live as a slave. I do not have space to consider this point here. See Oshana, "Personal Autonomy and Society," 86–89.

if I do not enjoy listening to or playing music, the fact that I lack perfect pitch does not seem to prevent me from being practically autonomous. My above discussion of positive freedom also suggests that different agents might require different positive freedoms to act in pursuit of their goals. Whilst there may be certain abilities that most agents require to do this, it seems that an agent with suitably esoteric goals could require very different sorts of positive freedoms from other agents.

With these reflections in mind, I am now in a position to explain what it means for an agent to be able to act effectively in pursuit of their ends in the sense that I invoked when introducing the practical dimension of autonomy. In most cases, positive constraints that take away an agent's negative freedom will debilitate the agent from pursuing a certain goal in *any* sense; for example, an agent cannot effectively pursue their goal of travelling the world if they are incarcerated. As such, the question of whether an agent has the requisite negative freedom for practical autonomy is often a discrete question; it is either the case that the agent is debilitated from acting in pursuit of a goal by a positive constraint, or it is not. An analogous claim could be made with regards to *some* positive freedoms; if an agent lacks certain enabling factors, they may be precluded from acting effectively in pursuit of their goals in *any* sense. For instance, I shall argue below that an agent may lack practical autonomy if they are informationally cut-off from achieving their goals by virtue of holding certain false beliefs. Call these sorts of freedoms *discrete* freedoms. With regards to discrete freedoms, we may say that an agent is only able to act effectively in pursuit of a goal whose achievement requires certain discrete freedoms, if they actually have those freedoms.

However, many of our freedoms admit of degrees. For instance, it seems plausible to claim that the pursuit of different goals might require different degrees of intelligence. Scalar freedoms such as intelligence present something of a theoretical problem with regards to practical autonomy, since we clearly cannot say that an agent must have the *maximum* degree of some particular scalar positive freedom in order to be able to act effectively in pursuit of their ends; this would make autonomy too demanding. Therefore, in cases in which the pursuit of some goal requires a scalar freedom x , it seems that we must stipulate that there is some threshold level of x that the agent must have in order to be practically autonomous. However, a problem with stipulating the relevant threshold here is that we must also allow for the possibility that an autonomous agent could have the threshold level of this scalar positive freedom and yet fail to achieve their goal; again, if practical autonomy is not to be too demanding, it cannot require that the practically autonomous agent must always *succeed* in their endeavours.

In view of these problems, I suggest that we cash out the notion of ‘having the necessary positive freedom to be able to act effectively in pursuit of some goal’ by saying that we may appropriately be said to have such freedom if there is some nearby possible world in which we have the proposed threshold degree of positive freedom, and in which we *do* successfully achieve our goal. However, if an agent’s failure to achieve their goal is wholly attributable to their lacking a degree of freedom that is necessary (although not sufficient) for the successful pursuit of the goal in question, then the agent lacks practical autonomy. This formulation allows us to give some substance to what it means to have the necessary scalar positive freedom to be able to act effectively in pursuit of some goal, without committing us to the view that being practically autonomous requires that the agent must *succeed* in the pursuit of her goals,

or that she has the maximum degree of a particular scalar positive freedom. I suggest then that, at the point of action, the freedom (in both the positive and negative sense) that is required for practical autonomy is the freedom to act effectively in pursuit of one's own ends in the manner that I have delineated above.¹⁷

I shall develop this point in section five. At this point though, it is important to acknowledge the exact extent of the claim that I am making in this section. First, my claim that practical autonomy requires the freedom to act effectively in pursuit of one's ends pertains only to the freedom required at the point of action, and only to those desires that the agent is reflectively autonomous with respect to. Second, in making the above claim, I am seeking only to give an account of the freedom required for practical autonomy, and not an account of the nature of freedom itself. This is important, since defining freedom *itself* as relative to an agent's desires or motives seems to involve a conceptual confusion.

¹⁷ One potential objection to this account is that it might be understood to entail that agents who have a preference to achieve an outcome that cannot possibly be achieved (say of flying unaided) can be said to lack practical autonomy. I am prepared to accept this point, but only because it has limited force. The reason for this is that on the account of autonomy that I developed in chapter two, agents will not be autonomous with respect to such preferences, in so far as preferences are understood to be action-guiding. Recall that on the theory that I developed in chapter two, preferences are understood to be rational desires for a certain motivating desire to be *effective* in moving one to act. I also argued in chapter two that an agent's preferences must cohere with their non-irrational acceptances. The problem then with the preferences that I am considering here is that they will fail to cohere with an important set of the agent's acceptances; namely, their beliefs concerning their freedom at the point of decision. I shall discuss this in section V. Notice that this view is compatible with the claim that agents may autonomously harbour 'pipe-dreams' in a non-action-guiding sense, and the claim that they can be autonomous in pursuing these goals if they (non-irrationally) believe that they can be achieved.

To see why, consider the example of Tom Pinch discussed by Feinberg.¹⁸ Tom Pinch is gifted with the freedom to do everything but act effectively in pursuit of the one end that truly matters to him. Feinberg points out that Tom Pinch does not lack freedom *per se*; after all, *ex hypothesi*, Tom Pinch enjoys almost every conceivable freedom. Rather, Feinberg claims that Tom Pinch lacks only contentment.¹⁹ In view of my arguments above, although Feinberg is right to claim that Pinch is free, I believe that he conflates contentment and the practical dimension of autonomy here. In light of my comments above, we might also acknowledge another point that speaks against Feinberg's claim here is that one can fail to achieve one's goal despite having the freedoms necessary to its pursuit.²⁰

In contrast to Feinberg, I claim that Pinch lacks a freedom that is necessary, at the point of action, for his practical autonomy. Of course, the freedom that autonomy requires here does not exhaust the concept of freedom; the nature of freedom goes beyond the freedoms that are necessary for practical autonomy. Although we may say that Tom Pinch is generally free, he is not practically autonomous because he lacks the freedom to act effectively in pursuit of the one end that he actually wants to achieve. If this conclusion is correct, then we might observe one of its corollaries, namely the implication that our freedoms can be increased in ways that are inconsequential to our practical autonomy at the point of action. For example, recall the example of Harry above. Suppose that Jane returned to her room after half an hour and, again unbeknownst to Harry, unlocked the room that Harry was in. It seems that this would

¹⁸ Feinberg, *Freedom and Fulfillment*, 38.

¹⁹ *Ibid.*, 38–39.

²⁰ We might also point out that one may be mistaken in thinking that achieving a certain goal will bring contentment.

increase Harry's freedom, but it is far from clear that it would increase his autonomy in staying in the room.

This suggests something interesting about the relationship between practical autonomy and freedom. If, as the above discussion suggests, all that matters at the point of action is whether the agent has the freedom to act effectively in pursuit of the ends that they have decided to achieve, then the agent's freedom to do otherwise is inconsequential to their practical autonomy *at the point of action*.²¹ One might worry that this claim is in tension with another popular view, namely the view that autonomy requires freedom of choice. For example, Hurka assumes that "autonomy involves choice from a wide range of options",²² and Raz claims that an autonomous person must have "adequate options available for him to choose from".²³ Although I believe that Raz and Hurka are not entirely correct here (as I shall explain in section five), their claims above are not in tension with my conclusion that the agent's freedom to do otherwise is inconsequential to their autonomy *at the point of action*. Indeed, Raz and Hurka might agree with this conclusion; instead, they would claim that freedom of choice is crucial for autonomy at the *point of decision*.

Before considering the relationship between autonomy and choice at the point of decision, it is prudent to address an important aspect of positive freedom at the point of action that I have not yet considered, namely, the way in which holding certain true beliefs seems to be necessary for the effective pursuit of many of our ends. This has important implications for the way in which we should understand how deception can undermine autonomy.

²¹ This point echoes Frankfurt's claims in Frankfurt, "Alternate Possibilities and Moral Responsibility."

²² Hurka, "Why Value Autonomy?," 362.

²³ Raz, *The Morality of Freedom*, 373.

IV True Beliefs and Autonomy

On the account of autonomy that I have been developing, an agent's beliefs are important with regards to her autonomy in two ways. First, on the reflective dimension, in order for the agent to regard an outcome x as good in a reason-implying sense, she must have beliefs about the descriptive features of x and about the good. In chapter two, I argued that an agent must meet a minimum threshold of epistemic rationality in holding these beliefs if she is to qualify as being reflectively autonomous with respect to the motivating desires that she sustains on their basis.

However, beliefs also play a role on the practical dimension of autonomy, since in order for an agent to be able to act effectively in pursuit of the end that she is motivated to achieve, she must have certain true beliefs about how to go about achieving that end. Indeed, if an agent acts in a manner that she is incorrect in believing will serve as a means to achieving the end that she is motivated to achieve, her action will be importantly disconnected from that motive. As Killmister puts the point:

No matter how autonomous an agent's motivations are, the action itself cannot be autonomous unless it is appropriately connected to the motivation behind it.²⁴

To further illustrate this thought, consider the following example that I considered in the previous chapter (but for a different purpose this time). Suppose that Sheila wants to quench her thirst, and spots a glass on her kitchen table containing a clear liquid

²⁴ Killmister, "Autonomy and False Beliefs," 521.

standing next to a bottle of gin. Assume then that Sheila has decisive reason to believe that the glass contains gin and that the gin will quench her thirst. However, suppose that she is mistaken in her belief; the glass contains acid.

In view of the arguments that I considered in chapter two, it seems that Shelia's desire to drink the contents of the glass here is rational;²⁵ so too are her beliefs upon which her desire causally depend (recall, she has no reason to doubt that the liquid in the glass is gin). However, even if we assume that Shelia is reflectively autonomous with respect to her motivating desire here, it seems plausible to claim that she lacks autonomy in some way, because her action itself is not appropriately connected to the motive underlying it.

We may cash out this intuition by pointing out that Shelia's rational (but false belief) undermines her practical autonomy, by severing the connection between her motivation to alleviate her thirst and how she attempts to achieve that end. She is not wholly self-governing in her conduct, because the false belief prevents her from acting effectively in pursuit of her desired end.²⁶ This sort of thought is captured in Mele's claim that being 'informationally cut-off' precludes one from autonomous agency. He writes:

²⁵ Notice that the desire is rational in a motivating, rather than normative sense.

²⁶ One can describe Shelia's actions in ways that make it appear that she is practically autonomous. If we describe the end that Shelia is hoping to achieve as 'getting the liquid in the glass to her mouth', then her act of picking up the glass and putting it to her mouth is clearly connected to her motive in an appropriate way. However, to the extent that we identify the end that 'getting the liquid in the glass to her mouth' itself serves as a means to (i.e. alleviating her thirst), we see that this act is not appropriately connected to her motive, given that the glass contains acid. This illustrates the importance of precision in how we individuate the agent's acts when we are assessing her autonomy.

(A) sufficient condition of S's being informationally cut-off from autonomous action in a domain in which S has intrinsic pro-attitudes is that S has no control over the success of his efforts to achieve his end, *owing to his informational condition*.²⁷

Contrary to what some opponents have claimed,²⁸ stipulating an informational condition such as Mele's does not entail that one must be *successful* in one's actions in order to be autonomous with respect to them. As Killmister points out in her analysis, Mele's condition allows that an agent can be autonomous when she fails to achieve her end; the condition merely stipulates that the agent's poor informational condition should not thwart the possibility of her being successful in achieving her end, by virtue of disconnecting her act from her intention.²⁹ To illustrate why this is not problematic, Killmister discusses the example of a woman, Jill, who attempts to intentionally kill a man, Jack, by hitting him over the head with a crowbar. Although she fails, Killmister claims that Jill is nonetheless autonomous, because her failure to achieve her end is not *due* to her poor informational condition.³⁰ Killmister herself does not elaborate further on how Jill differs from someone like Sheila with regards to her poor informational condition, and why Jill is autonomous when Sheila is not. One way of explaining the difference is to appeal to a modal claim of the sort I made above regarding positive freedom. We might say that Jill's failure is not attributable to her false beliefs because there is a nearby possible world in Jill holds the same beliefs, and in which she successfully achieves her end. Whilst this is true of Jill, it is not true of Sheila.

²⁷ Mele, *Autonomous Agents*, 181, emphasis added.

²⁸ See Mckenna, "The Relationship between Autonomous and Morally Responsible Agency."

²⁹ Killmister, "Autonomy and False Beliefs," 525.

³⁰ *Ibid.*

The above reflections offer a suggestion as to why philosophers are divided over the question of whether holding false beliefs undermine autonomy.³¹ In this section, I have explained that having false beliefs can render an agent ineffectual in pursuing the ends that she is motivated to achieve. However, these false beliefs need not affect the agent's *reflective* autonomy with regards to her decision about what to do. As such, it seems that we can explain the diverging intuitions concerning cases of false beliefs as follows. If we believe that the reflective dimension of autonomy can tell the whole story of autonomous agency, then having false beliefs concerning the means that are necessary to achieving one's desired end need not undermine one's autonomy. On the other hand, if we claim that an adequate theory of autonomy also includes a practical dimension then it is clear why even epistemically rational false beliefs can undermine autonomy; they will do so when they render the agent ineffectual in her pursuit of the ends that she is motivated to achieve.

This has important implications for how we should understand the way in which deception undermines autonomy. Taylor claims that deception only undermines autonomy if it is *intentional*; he claims that an agent is not autonomous with respect to her decision if she fails to meet the following 'threshold condition':

If the information upon which the agent bases her decision has been affected by another agent with the end of leading her to make a particular decision . . . and if she is not

³¹ For two examples of theorists who claim that false beliefs do not undermine autonomy, see Mckenna, "The Relationship between Autonomous and Morally Responsible Agency," 208–209; Arpaly, "Responsibility, Applied Ethics, and Complex Autonomy Theories," 175.

aware of the way in which this information has been affected *then* she did not make the decision that (her deceiver) intended her to make.³²

Pace Taylor, on the account that I have delineated here, deception that renders the agent informationally cut-off from achieving their ends can undermine autonomy, even if the deception itself is unintentional.³³ However, there are still important differences between intentional and non-intentional deception. In cases of non-intentional deception, the deceived agent is not being *controlled* in any sense; in contrast, an intentionally deceived agent is led to form false belief by another so that their behaviour will serve the deceiver's ends rather than her own. As such, the intentionally deceived agent's will is subjugated to another's in a manner that makes the affront to autonomy more serious; the deceived agent is *heteronomous*, that is, ruled by another.³⁴ In contrast, we may say that agent who is merely informationally cut off from successfully achieving their end by virtue of non-intentional deception simply lacks practical autonomy.

Furthermore, an agent who provides testimony that leads another to form a false belief will normally be culpable for this if the deception is intentional; in contrast, if the deception is non-intentional, and if the testifying agent has fulfilled her epistemic

³² Taylor, *Practical Autonomy and Bioethics*, 7.

³³ Taylor rejects the claim that autonomy incorporates a practical dimension, and instead claims that being able to act effectively in pursuit of one's ends may increase the *value* of autonomy. This view not only seems incongruous with the way in which we use the term autonomy in bioethics, but it also fails to account for the way in which our beliefs about our freedom can impinge upon the autonomy of our decisions. Moreover, in chapter five, I shall also suggest that this is an inadequate account of the value of autonomy. I also object to Taylor's account in Pugh, "Ravines and Sugar Pills: Defending Deceptive Placebo Use."

³⁴ Introduction, 12.

responsibilities in forming the beliefs upon which she bases her testimony (by gathering adequate evidence for her claims say), she will not be culpable for undermining the other agent's autonomy by causing her to have a false belief.

V Freedom at the Point of Decision

At the end of section III, I suggested that practical autonomy requires that the agent has, at the point of action, the positive and negatives freedoms that are necessary for them to act effectively in pursuit of their own ends. In this section, I shall consider the freedom that autonomy requires *at the point of decision*. In considering the agent's freedom at the point of decision, we are considering the freedom that she believes herself to have prior to making a decision about what to do; as I shall explain, it is important to consider the agent's freedom at this point, because an agent's beliefs about their freedoms can impinge on their reflective autonomy.

To begin this argument, it is important to acknowledge that we make our decisions about what to do in the light of what we believe to be practically realisable. Griffin puts this point as follows:

We do not, as a matter of fact, form our plans of life as if they were, in effect, choices from a Good Fairy's List – 'whatever you want, just say the word'. Our desires are shaped by our expectations, which are shaped by our circumstances.³⁵

³⁵ Griffin, *Well-Being*, 47.

Accordingly, when we are in the process of deciding whether to sustain certain motivating desires, our decision is informed by what we believe we are free to do. For example, when I consider which career path I want to pursue, I make my decision having assessed the capacities that constitute my positive freedom to pursue certain careers, and having considered any positive constraints on my negative freedom to pursue others. It may be the case that my beliefs about what I am and am not free to do are false; however, it is not the freedom that I actually have that I take into consideration in my practical deliberations, but rather my *beliefs* about the freedoms that I have.

Colburn, following Elster,³⁶ refers to the sort of phenomenon explained above as “conscious character planning”, describing it as the process of:

. . . being aware of the limitations in one’s options and moulding one’s projects and inclinations so as to settle on preferences which one can fulfil.³⁷

As both Elster and Colburn argue, conscious character planning does not undermine autonomy.³⁸ Although a perfectly autonomous agent might not have any limitations on their freedoms, the fact that normal humans have limited freedoms and tailor their preferences in accordance with them is not inimical to their autonomy with respect to those preferences. To claim otherwise would be to rule out the possibility of autonomy at the very outset, given the nature of the world we live in, in which environmental

³⁶ Elster, *Sour Grapes*, 117–119.

³⁷ Colburn, “Autonomy and Adaptive Preferences,” 55.

³⁸ See Ibid, and Elster, *Sour Grapes*, 117–119.

forces contribute to the shape and limits of our freedoms. The absence of certain freedoms at the point of decision merely shapes the contours of our choice domains.

In conscious character planning, an agent's awareness of the limitations of their freedoms *informs* their decisions about what to do, but it does not preclude their reflective autonomy with respect to that decision. However, an agent's beliefs about their freedom can threaten their reflective autonomy if they believe themselves to have extremely limited freedoms. To see how, it is illustrative to contrast the case of Tom Pinch (considered above) with the example of Martin Chuzzlewit:

Suppose that Martin Chuzzlewit finds himself on a trunk line with all of its switches closed and locked, and with other 'trains' moving in the same direction on the same track at his rear, so that he has no choice at all but to continue moving straight ahead to destination *D* . . . But now let us suppose that getting to *D* is Chuzzlewit's highest ambition in life and his most intensely felt desire. In that case, he is sure to get the thing in life he wants most.³⁹

Whether or not Chuzzlewit is autonomous here depends on the extent to which his lack of freedom is the *reason* that he came to sustain his motive to go to *D*. On the theory that I have developed, for Chuzzlewit to be reflectively autonomous with respect to his motive to get to *D*, he must have come to adopt it on the basis of a (non-irrational) belief that his getting to *D* would be good in a reason-implying sense. However, our disposition to adopt motivating desires on the basis that their content is good in a reason-implying sense can be compromised in situations in which we believe that our

³⁹ Feinberg, *Freedom and Fulfillment*, 38.

freedoms at the point of decision are severely restricted. If we believe that only one course of action is available to us, our lack of alternative possibilities may dissuade us from engaging in any sort of reflection about the value of the available outcome; rather, we may adopt the motive to pursue that outcome on no other basis than the fact that it is the only option available to us.

This phenomenon is known as adaptive preference formation.⁴⁰ In contrast to conscious character planning, it may be defined as the “unconscious altering of our preferences in light of the options that we have available”.⁴¹ To illustrate the phenomenon of adaptive preference formation, let us alter the case of Harry the dog-sitter above so that Harry forms an adaptive preference:

Suppose that Harry forms the desire to leave the house and go to the pub upon Jane’s departure but that he then hears Jane lock him into the house. Upon hearing this, Harry resigns himself to staying in to look after the dog, but convinces himself that this was really his preference all along.

As Elster and Colburn both suggest (albeit for different reasons), unlike conscious character planning, adaptive preference formation *does* seem inimical to autonomy.⁴² On the theory that I have developed, the reason for this is that in cases of adaptive preference formation, the agent no longer endorses their motivating desire on the basis of a belief that the outcome of the desire is good in a reason-implying sense;

⁴⁰ Elster, *Sour Grapes*; See also Sen, *Development as Freedom*; Nussbaum, *Women and Human Development*.

⁴¹ Colburn, “Autonomy and Adaptive Preferences” 52.

⁴² Elster, *Sour Grapes*, 20; Colburn, “Autonomy and Adaptive Preferences,” 61–71.

rather they sustain this desire because the outcome it concerns is the only option available to them. However, the fact that an outcome is the only one available does not make that outcome good in a reason-implying sense. Moreover, the self-deceptive nature of the way in which this preference is formed may preclude later critical reflection on the value of the outcome.

Although lacking freedom at the point of decision is an obvious causal factor behind adaptive preference formation, it is not clear that lacking such freedoms must *necessarily* lead to adaptive preference formation. After all, the fact that only one option is available to an agent does not make it impossible for them to endorse that option on the basis of its reason-giving content (rather than its mere availability). For example, it is (to all intents and purposes) practically impossible for a passenger to jump out of a commercial airplane in mid-flight. Now, even if I believe that I have no alternative to staying in a plane for the duration of a flight, this does mean that I cannot regard the outcome of staying in the plane as reason-giving. As long as I (non-irrationally) believe that the content of my motivating desire is good in a reason-implying sense, then I can be reflectively autonomous with respect to that desire even if I also believe that I lack the freedom to do otherwise. However, I will lack autonomy if I adopt the motivating desire to do something, *just because* I believe I lack the freedom to do anything else.

Let us return to the question of Martin Chuzzlewit's autonomy. The way in which Feinberg phrases the example makes it ambiguous as to whether it is best to interpret Chuzzlewit's being motivated to get to *D* as an instance of adaptive preference formation. The fact that Chuzzlewit 'finds himself' on the particular trunk line does not tell us whether he regarded getting to *D* as being good in a reason-implying sense prior to finding himself on the track, or whether he forms the motive to get to *D* on the basis that he has found himself on the particular trunk line that leads to *D*. In the latter case, I

would suggest that Chuzzlewit lacks autonomy because he does not adopt his motive on the basis of its reason-giving content, but rather on its mere availability; he has thus formed an autonomy undermining adaptive preference. However, if Chuzzlewit had formed a preference for *D* prior to finding himself in this curious position, and his lack of freedom had not otherwise impaired his reflective autonomy with respect to his motivating desire, I would claim that Chuzzlewit is reflectively and practically autonomous.

The above reflections speak against Raz and Hurka's view that having a variety of choices is a necessary condition of autonomy. In chapter six, I shall argue that increasing an agent's choice set can serve to *enhance* their reflective autonomy. However, I believe that it is a mistake to claim that one *cannot* be autonomous with respect to one's decision if one lacks a variety of options to choose from. This was the point of my airplane passenger example above; we can be autonomous with respect to our practical decisions, even if we lack the freedom to do otherwise. In fact, to hold the opposite view is particularly problematic in a bioethical context, since in this context there are many agents who have severely restricted choices that we want to be able to say can make autonomous decisions. For instance, if we claim that agents are autonomous only if they have an adequate variety of options available to them, then we may be committed to the view that an agent who must receive a life-saving medical intervention cannot autonomously consent to it; after-all, if anybody lacks an adequate variety of options available to them, surely this patient does. Presumably, an adequate theory of autonomy in bioethics should allow that such a patient could autonomously consent to a medical intervention in this sort of scenario.

Conclusion

In this chapter, I have argued that an adequate theory of autonomy in bioethics should incorporate a practical dimension pertaining to the agent's ability to act effectively in pursuit of their ends. Two of the reasons for holding this view are that an account of autonomy that acknowledges this dimension of autonomy reflects our beliefs about what respecting autonomy requires in bioethics, and why we prudentially value autonomy. I argued that in order to be practically autonomous, agents must have the positive and negative freedoms to act effectively in pursuit of what they want (in a reflectively autonomous sense) to do. Whilst different agents may require different sorts of freedoms to pursue their goals, it seems likely that there will be some freedoms that are necessary for the pursuit of most goals. In particular, I claimed that having certain sorts of true beliefs will normally be necessary for practical autonomy; this helps to explain why deception is normally inimical to the deceived agent's autonomy.

I also claimed that there is an important relationship between the reflective and practical dimensions of autonomy, in so far as agents decide to sustain their motivating desires in the light of their beliefs about what is practically realisable for them. This suggests a third reason why an adequate theory of autonomy in bioethics should incorporate a practical dimension; a theory that does not incorporate a practical dimension cannot adequately explain the effect that our beliefs about what we are free to do can have on our decision-making. I explained how these sorts of beliefs can undermine autonomy in cases of adaptive preference formation. In the next chapter, I shall argue that recognising the practical dimension of autonomy is also crucial for giving an adequate account of why coercion undermines autonomy.

Chapter 4 – Coercion and Autonomy

As I pointed out in the introduction to this thesis, the standard view of autonomy in bioethics stipulates that a necessary condition of an autonomous decision is that the agent makes her decision about what to do in the absence of coercion (amongst other controlling influences that determine action). However, whilst this seems intuitively plausible, it leaves open two important questions. The first is the question of how coercion should be understood to undermine autonomy when it does. The second is the question of what should count as an instance of coercion. Whilst clear-cut cases of coercion spring easily to mind, there are many cases in which it is unclear whether a proposal ought to be understood as a coercive one. Perhaps most salient amongst these unclear cases are those that involve so-called ‘coercive offers’. I shall consider some such cases below; at this point though, it is sufficient to note that the reason that these cases are understood to be problematic is that many theorists claim that a necessary condition of an agent P’s coercing another, say Q, is that P *threatens* Q.¹

The issue of whether offers (as well as threats) can be coercive is not merely of theoretical interest. The issue of whether agents could autonomously consent to allegedly coercive offers has been raised in the context of bioethical debates concerning, *inter alia*, markets for human organs,² and offering sex offenders reduced sentences in return for undergoing surgical or chemical castration.³

¹ Stevens, “Coercive Offers,” 83.

² Richards, *The Ethics of Transplants*, 60–64.

³ McMillan, “The Kindest Cut?”.

In this chapter, I shall suggest that our answer to the second question should be informed by our answer to the first; in order to explain whether or not a certain proposal is coercive, I suggest that we must first have an account of why coercion undermines autonomous decision-making. Drawing on my distinction between the reflective and practical dimensions of autonomy, I shall argue that coercion can undermine autonomous decision-making in so far as it involves one agent subjugating the will of another by exerting control over what is practically realisable for them. I shall then use this analysis to explain the difference between threats and offers, and argue that although offers cannot be coercive in themselves, they can constitute part of a coercive situation which involves the illegitimate reduction of the recipient's freedoms.

I Introducing Coercion

To begin, it is prudent to highlight an important delimitation concerning the understanding of coercion that I shall invoke. Some theorists claim that directly forcing an agent to do something against their will is an instance of coercion.⁴ However, I shall *not* understand such cases as examples of coercion. The reason for this is that if we claim that actions carried out as a result of direct force are coerced, then we are in danger of making the concept of coercion too wide. Doing so would serve to obscure an important distinction made by McCloskey that highlights the fact that under direct force, it will usually be true that in some sense the forced agent does not act, but rather

⁴ See Bayles, "A Concept of Coercion"; Lamond, "Coercion, Threats, and the Puzzle of Blackmail."

is *acted upon*;⁵ in contrast, the agent who is subject to a conditional threat can still appropriately be deemed to be acting (though perhaps not autonomously). As such, I shall only be interested in coercion that involves techniques that alter the victim's motives.

Accordingly, let us take the following as a paradigm case of coercion:

Terry is a bank teller whose bank is targeted by an armed robber. The robber tells Terry to open the vault. Terry refuses to cooperate. The robber then grabs a customer and claims that he will shoot her unless Terry cooperates. Terry opens the vault.

Of course, agents can also be coerced into refraining from actions; suppose that after agreeing to open the vault in order to prevent the execution of the hostage, Terry begins to call the police. Here the robber might feasibly tell Terry that unless he refrains from doing so, then he will kill a hostage; this too, would amount to coercion.

Following Raz,⁶ we might claim that the following five elements of the above examples are necessary (but not sufficient)⁷ conditions for the robber's coercing Terry, where the act *X* can be understood as either an uninitiated act that the coercer aims to

⁵ McCloskey, "Coercion," 336. We should note that agents who are forced in this way also lack autonomy; however, they do not lack autonomy for the same reason that coerced agents do.

⁶ These conditions correspond to conditions that Raz appeals to in his definition of coercion, which in turn may be understood as a somewhat simplified version of those that Nozick offers. See Raz, *The Morality of Freedom*, 149; Nozick, "Coercion."

⁷ These elements are not *sufficient* for coercion because there are some further nuances that one might draw to distinguish coercion via threats from conditional warnings.

make his victim perform, or the act of ceasing an initiated action that the coercer aims to make his victim refrain from performing:

- 1) Prior to the robber's communication, Terry believes he has decisive motivating reasons to refuse to *X*.
- 2) The robber's communication makes it clear to Terry that he will bring about some consequence *C* that otherwise would not have occurred if and only if Terry refuses to *X*.
- 3) Terry has sufficient reason to believe that the robber will bring about *C* if he fails to *X*.
- 4) Part of the intention underlying the robber's communication here is to get Terry to *X*.
- 5) The fact that the non-performance of *X* will lead to consequence *C* is part of the reason that Terry comes to believe that he now has a strongly decisive reason to perform *X*.⁸

There are a few points worth remarking upon briefly here. The first thing to acknowledge is that it is important that we carefully individuate the act *X* in our description of a coercive situation; condition (1) might not be necessary if we

⁸ Whilst I have followed Nozick in analysing the nature of coercion from the perspective of the coerced party, an alternative strategy is to instead analyse it from the perspective of the coercer, and view coercion as one party taking advantage of pre-existing power imbalances. For example, see McGregor, "Bargaining Advantages and Coercion in the Market".

individuate the act *X* in too broad a manner. To illustrate, in the above example, suppose that we understood *X* to represent the act of simply ‘opening the vault’. On this understanding, it is clearly not a necessary condition of the robber’s threat being coercive that Terry must have believed that he had a decisive motivating reason to refuse to *X* prior to the robber’s threat; for instance, Terry may have believed that he had good reasons to open the vault for a customer. However, if we individuate *X* more precisely, as the act of opening the vault to allow the robber access to the money, then it seems right to claim that condition (1) will be a necessary condition of why the robber’s threat is coercive.

Condition (1) also makes it clear that a threat will not be coercive if the threatening party only threatens to bring about a bad consequence if the recipient fails to do something that the recipient herself believe she has reason to do anyway. For instance, suppose that I am driving to A and come to a fork in the road, with one road leading to A, and the other leading to B. Suppose that just before I come to the road, a highway-man tells me that he will shoot me unless I drive down the road leading to A. Whilst this is a credible threat, it does not undermine my autonomy in driving to A; my decision about what to do was not affected by the threat of a bad consequence if I chose otherwise. As such, the threat is not coercive in this example.

Conditions (3) and (5) explain why some other threats are not coercive. As (3) states, if the recipient of the threat believes that the threatening party will not actually bring about the threatened consequence, then such a threat is not coercive. Furthermore, condition (5) stipulates that the threat must be successful in order for it to be coercive; that is, the threat must be operative in changing the recipient’s belief about what they now have decisive reason to do.

Finally, it should be noted that the robber's act of communication *per se* is not coercive. The communication is important in so far as it is the means by which the robber alters Terry's beliefs about what the consequences of his action will be; however, the coercion seems to occur at the level of the robber's forming the *intention* to manipulate those consequences (at least partly) in order to influence Terry's behaviour, and not merely at the level of his *informing* Terry of those now probable consequences.⁹ Condition (4) aims to capture this sort of thought.

In the above case of coercion, the robber's communication to Terry is a conditional threat. Part of my aim in this chapter is to establish whether certain offers can be coercive. However, since this would introduce complications at this point, I shall, for the purposes of the next section, assume that only threats can be coercive, where a threat is to be understood as a communication of the sort identified in condition (2), in which the consequence that the communicator announces a conditional intention to bring about is something that they believe would make the other party worse off. Although I shall argue against this particular account of threats in section III, this rudimentary understanding is sufficient for my discussion at this point.

II Why Does Coercion Undermine Autonomy?

Although it seems intuitively plausible to claim that coercion can undermine autonomy, many theories of autonomy lack the conceptual apparatus to explain why this

⁹ This suggests a rough way in which to begin to draw a distinction between conditional threats and warnings. See Nozick, "Coercion," 453–458.

is the case. I shall first consider how Frankfurt and Dworkin attempt to explain why coercion undermines autonomy on their desire-based hierarchical accounts of autonomy, before going on to offer my own account of the relationship between coercion and autonomy. Although I have already argued that desire-based hierarchical theories of autonomy are subject to damning criticisms, I shall nevertheless begin by considering these views, since they are widely recognised views of autonomy that have already been invoked in the debate regarding the relationship between coercion and autonomy.¹⁰ Moreover, the flaws that attend these theories do not alter the conclusions that we ought to draw about the relationship between autonomy and coercion; in fact I shall suggest that the problems with Frankfurt and Dworkin's account of why coercion can undermine autonomy are also problematic for any theorists who define autonomy solely in terms of what I have called the reflective dimension of autonomy. I shall argue that it is possible to provide an adequate account of how coercion undermines autonomy by acknowledging the importance of what I have called the practical dimension of autonomy, and its relationship to the reflective dimension.

In considering the effect of coercion on autonomy, Frankfurt claims that a coerced agent is moved to act "by a desire which is not only irresistible, but which he would overcome if he could";¹¹ as such, on his hierarchical account, the coerced agent's autonomy is undermined because he acts on the basis of a first-order desire that he does not endorse at a second order level. Similarly, Dworkin claims that coercion undermines

¹⁰ Frankfurt, *The Importance of What We Care about.*; Dworkin, "Acting Freely"; Taylor, "Autonomy, Duress, and Coercion"

¹¹ Frankfurt, *The Importance of What We Care about*, 42.

autonomy because coerced agents are compelled to act for reasons that they resent acting for.¹²

As Thalberg argues though, these hierarchical analyses of why coercion undermines autonomy are inadequate. In short, the problem that Thalberg raises is that most persons who are subjected to coercion “. . . would, at the time and later, give second-order endorsement to their cautious [and compliant] motives.”¹³ This is problematic for such analyses, since coerced agents would thereby often qualify as being autonomous on these theories when they make a coerced decision to act in some way.

It seems that the rationalist account of reflective autonomy that I developed in chapter two would be open to a similar sort of problem. The motivating desire that I form following a coercive threat could be one that I rationally endorse with a preference that coheres with my character system. For instance, even if Terry is normally an upstanding moral citizen, he may nonetheless endorse his desire to open the vault for the robber, because the content of his desire also includes saving the life of the robber's hostage, and he understands this to be a more valuable outcome than safeguarding the bank's money. It might be claimed that historical theories of autonomy are better equipped to be able to explain why coercion undermines autonomy. However, even in the absence of the problems with such theories that I highlighted in chapter one, it is not clear that they can give an adequate explanation of why coercion undermines autonomy. For instance, coerced agents are not 'compelled' to have their desires in the sense that Mele discusses, since they do not form their motivating desires in a manner that bypasses their mental control; on the contrary, coerced agents change their motivating

¹² Dworkin, "Acting Freely," 377–378.

¹³ Thalberg, "Hierarchical Analyses of Unfree Action," 126.

desires as a rational response to their changed circumstance. Furthermore, even if a historical theory were to claim that agents would not hypothetically endorse the aetiology of a coerced desire, it still seems that we are owed an account of why they would not endorse it, and why their aetiology is flawed. In lacking such an account, it seems that such theories are merely stipulating that coercion undermines autonomy by fiat.

To return to the hierarchical account of why coercion undermines autonomy, Taylor has argued that Thalberg's criticism of Dworkin and Frankfurt is misguided, because it assumes that these hierarchical theorists endorse what Taylor calls the "assumption of the transitivity of autonomy", according to which the autonomy of persons is transitive across their desires and actions.¹⁴ That is to say, according to this assumption, if an agent endorses their motivating desire to act in some way at a second order level, then this entails that they are also autonomous with respect to the *action* to which that motivating desire gives rise. Taylor suggests that Frankfurt and Dworkin can account for why coercion can undermine autonomy if they reject this assumption. According to Taylor, hierarchical theorists need not assume this, since the fact that a coerced person might endorse their first-order desire to comply with their coercer's demands in an *absolute* sense, does not entail that that they also have a preference to be moved by that desire *relative* to other possible states of affairs that they would prefer to be in. Accordingly, Taylor argues that the hierarchical theorists can claim that an agent will lack autonomy with regards to his performance of action A if:

¹⁴ Taylor, "Autonomy, Duress, and Coercion," 133.

He is motivated to perform action A by a desire that he is autonomous with respect to, but this desire is one that he would prefer not to be moved to act by, because he would prefer to be in a situation other than the one that he is actually in.¹⁵

However, even if this is what Frankfurt and Dworkin intended in their analysis, this approach to understanding why coercion undermines autonomy is still flawed. One problem with this approach that Taylor points out is that it makes the autonomy-undermining effect of coercion contingent upon the victim's attitudes towards his circumstances. As Taylor argues, this is problematic because it means that a coerced agent's autonomy will not be impaired if she simply decides not to resent the reasons for which she acts. As an example of this, Taylor considers a city-dweller who is forced at gunpoint to hand over his money to a mugger, but who does not resent his reasons for doing so, since he regards muggings as a part of living in the city.¹⁶ However, the fact that the agent in this example does not resent his reasons for acting does not alter the fact that he seems to have been coerced by the mugger in a manner that undermines his autonomy.

A second problem with this approach is that the hierarchical account seems to prove too much. There are many situations (particularly in biomedical contexts) in which an agent would prefer not to be moved to act by a desire that they are autonomous with respect to, because they would prefer it if they were in a situation different to the one that they find themselves acting in; yet we do not think that this always means that their autonomy has been violated. For example, on this approach, it

¹⁵ Ibid., 138.

¹⁶ Ibid., 154.

would appear that patients suffering from terminal illnesses could not be autonomous with respect to their decision to consent to analgesics; after all, they would surely prefer it if they were not moved by a desire to avoid severe pain, and would also prefer to be in a situation other than the one that they are actually in. Yet, to claim that such patients could thereby not make this decision autonomously is implausible.

Something that seems to underlie both of these problems is that this hierarchical approach fails to adequately acknowledge the fact that, in cases of coercion, the coerced agent's behaviour is *purposefully controlled*. Taylor himself implicitly seems to acknowledge that this is central to understanding why coercion undermines autonomy; he states that part of the reason that coercion undermines autonomy is that in order for the coerced party to satisfy their motivating desire to comply with their coercer's demands, she must relinquish control to the threatening party.¹⁷ Yet, in order to understand how coercion undermines autonomy, it seems that we need to have at least a better understanding of the nature of this control and why it is important.

As I have pointed out in earlier chapters, many theorists of autonomy claim that autonomy is simply a matter of carrying out a certain sort of reflection on one's motivating desires. On such an understanding, one way in which a third party could exert control over another's decision in a manner that undermines the latter's autonomy is by psychologically manipulating them to act in accordance with a desire that they would not reflectively endorse. To see how psychological manipulation differs from coercion, it is illustrative to reimagine the case of Terry in the following way: Prior to the robber's coercive threat, suppose that the following would have been Terry's ranking of his available options:

¹⁷ Ibid.

Best = Option *A* – (Don't hand over keys) - The money is safe.

Worst = Option *B* – (Hand over keys) - The money is stolen.

Contrary to the original case, we can imagine that the robber could have instead controlled Terry's behaviour by psychologically manipulating him (say through hypnotism) rather than threatening to kill a hostage, so that Terry came to prefer *B* over *A*. In contrast though, when the robber threatens Terry, Terry's initial preference structure is left intact; rather, the robber's coercive threat serves to take away Terry's highest ranked option, *A*, from his choice set, and replaces it with option *C* which, for Terry, is less good than *B*:

Option *C*: Don't hand over keys (money is safe but an innocent civilian is killed)

This difference between coercion and manipulation illuminates an element of autonomous agency that many theorists overlook, and which I discussed in the previous chapter. Whilst it is correct to say that autonomous agents must carry out a certain sort of reflection on their motivating desires, to say that this is all that autonomy amounts to is to neglect the point that autonomous agents decide to sustain their motivating desires in the light of their beliefs about what is *practically realisable* for them; their decision to sustain their motivation to act in some way is not made in isolation from their beliefs about what it is they are able to do. This is important, because with this in mind, it

seems that one way an agent can indirectly exert control over another's motivating desires is by limiting their freedom to pursue certain options.

This, I suggest, is crucial to beginning to understand why coercion undermines autonomy; part of the reason that the coercer can be said to control their victim's behaviour is that coercion amounts to a third party's exerting control over what is practically realisable for another. The coercer ensures that the course of action that they want their victim to carry out becomes the most preferable for their victim *by the victim's own lights*, because they make their victim believe that they do not have the freedom to pursue a more preferable option, in light of the consequences that the coercer has announced an intention to attach to courses of action which would otherwise be preferable.

It might be claimed that the approach that I am advocating here, like the hierarchical analysis, proves too much. To illustrate why, compare the case of Terry with Fred:

An unarmed robber targets Fred's bank. The robber asks Fred to open the vault; Fred refuses. However, it becomes clear to Fred that an innocent customer has trapped himself in the vault doorway, and will be killed unless Fred opens the vault. However, opening the vault will allow the robber to access the vault, and to escape with the money. Nonetheless Fred decides to open the vault.

Although Terry and Fred both do something that they had decisive reason not to do prior to having their practically realisable options reduced, it is somewhat counter-intuitive to claim that Fred was coerced into opening the vault (even if we believe that

his autonomy in doing so was diminished). How can my approach accommodate this point?

The most obvious difference between the two cases here is that Terry's freedom to pursue option A was removed by another agent, whereas Fred's freedom to do so was removed by hazard. However, it is not immediately clear why this should make any difference. Yaffe suggests that one important difference here is that in the case of coercion, the coercer will often track the compliance of their victim.¹⁸ In Fred's case, the customer getting stuck in the vault doorway makes the option of not opening the vault worse; however, Yaffe's point is that if Fred still chose not to open the vault, the bad consequences of his doing so would not normally get any worse. In contrast, if Terry chose not to open the vault, Yaffe's suggests that the robber would normally threaten to make the consequences of this choice increasingly bad (perhaps by threatening to kill more hostages) until Terry eventually agreed to comply with the robber's demands.

Yaffe's analysis seems right in part. However, whilst I agree with Yaffe that coercers will often track their victim's compliance, I do not believe that it represents the fundamental reason why coercion undermines autonomy. After all, the coercer's tracking the compliance of their victim does not seem to be a necessary element of coercion. For instance, a coercer could issue a threat and privately resolve not to threaten anything worse if the victim fails to comply with the issued threat. This would surely qualify as coercion if the initial issued threat was successful. Rather what seems to be important about a coercer's tracking compliance is that they are reducing another agent's freedom *for their own purposes*. In contrast, when our freedoms are reduced by

¹⁸ Yaffe, "Indoctrination, Coercion and Freedom of Will," 355.

hazard, there is no rhyme or reason to it; as such, even though we might say that an agent's freedom can be diminished by hazard, it does not involve a *subjugation* of the coerced agent's will. However, in the case of coercion, the intent underlying the coercer's reducing their victim's freedoms means that the victim does not simply lack autonomy, or simply fail to be self-governing; rather she will be *heteronomous* on the understanding that I have invoked in this thesis, that is, governed by another.

This observation mirrors a point that I made in my discussion of deception in chapter three. There, I argued that deception undermines autonomy, whether or not it is intentional, if it renders another agent informationally cut-off from achieving their goal. However, I also suggested that intentional deception is a greater affront to autonomy, since intentional deception (like coercion) involves the subjugation of the victim's will to another; the victims of intentional deception and coercion are both, I suggest, heteronomous. However, it should be acknowledged that I do not claim that having one's freedoms reduced by hazard necessarily renders the agent non-autonomous. Whilst having one's freedoms reduced by hazard at the point of decision will often *diminish* an agent's autonomy, I also argued in chapter three that agent's can retain be reflectively autonomy with respect to their choices to a minimum threshold level, even when they have severely restricted freedoms. They can do so as long as they adopt their motivating desires on the basis of a belief that the content of their desire is good in a reason-implying sense; and agents can still do this even if they have had their freedoms severely reduced by hazard.

In this section, I have attempted to add detail to Taylor's claim that a central element of coercion, and part of the reason that it undermines autonomy, is that it involves one party controlling the conduct of another. I have suggested that by issuing a

credible coercive threat, the coercer exerts controls over their victim's beliefs about what is practically realisable. In turn, this serves to undermine autonomy, in so far as autonomous agents choose to sustain their motivating desires in the light of what they believe is practically realisable, and the coercer intends to determine their victim's behaviour through their intrusion; the coercer subjugates the will of their victim to their own. In view of these claims, I shall now consider whether offers can be coercive in the same way as threats.

III Coercive Offers?

As Nozick points out, *prima facie*, it would appear that offers are not normally coercive; indeed, he goes so far as to claim that he is “. . . (inclined) to say that one is never coerced when one does something because of an offer”.¹⁹ This seems plausible; after all, we commonly get people to do things that they would otherwise not do by offering them inducements, and it seems clear that this does not usually involve coercion. However, whilst we intuitively believe that making someone an offer does not normally undermine their autonomy, it is not clear what the philosophically significant difference between a threat and an offer is. As Anderson points out in his introduction to the concept of coercion:

¹⁹ Nozick, “Coercion,” 452.

. . . offers may also be made with the same general intention as coercive threats: that is, to make some actions more attractive, others less so.²⁰

Anderson goes on to point out that in the case of both threats and offers, the proposing agent, *P*, claims that *she* will bring about consequences *C* if and only if the proposed-to agent, *Q*, does some action *A*, *that P wants Q to perform*.²¹ I shall suggest below that Anderson's analysis here is subtly incorrect. Prior to doing so though, I shall consider, and then reject, Nozick's seminal analysis of the difference between offers and conditional threats.²²

Intuitively, it seems that one important difference between threats and offers is the nature of the consequence that the proposer announces a conditional intention to bring about in each case. For Nozick, this point suggests that the reason that we do not believe that offers can be coercive in the same way as threats is that in the case of an offer, a rational agent would be willing to move from their pre-proposal 'baseline' situation to the post-proposal situation. They would be willing to do this because the nature of the consequence that the third party announces an intention to bring about is such that it would make the recipient better off than they would have been in comparison to the baseline of the normal expected course of events, whereas the same cannot be said in the case of threats.²³

²⁰ Anderson, "Coercion."

²¹ Ibid.

²² Nozick, "Coercion," 447–453.

²³ Ibid., 459–460. Nozick also suggest that this explains why coercion undermines autonomy, since the threatened agent is put into the post-proposal situation unwillingly. Ibid., 459.

However, Nozick himself recognizes a problem with this way of explaining the difference between threats and offers. The problem is that in some cases, the baseline situation can *itself* be coercive. To illustrate this point he asks us to consider the following example:

Suppose that usually a slave owner beats his slave each morning, for no reason connected with the slave's behaviour. Today he says to his slave, "Tomorrow I will not beat you if and only if you now do A".²⁴

Assuming that the slave owner is coercing the slave here, it seems that the difference that Nozick initially draws between threats and offer is inadequate for explaining why some threats are coercive whilst offers cannot be. Nozick responds to this case by suggesting that it is possible to view the above proposal as a threat when we consider it against the baseline of what he calls the *morally* expected course of events that can be understood as follows:

Morally Expected Course of Events: The baseline comparison course of events that incorporates what we can morally expect of others in their actions towards the recipient in the pre-proposal situation.

In Nozick's view, the latter understanding of 'what is to be expected' should take precedence in the slave case.²⁵

²⁴ Nozick, "Coercion," 450.

However, Nozick does not believe that the ‘moral baseline’ interpretation of the normal course of events should *always* take precedence when the morally expected course of events and the morally-neutral expected normal course of events diverge; for instance he claims that the following is an example of a coercive threat:

P is Q’s usual supplier of drugs, and today when he comes to Q, he says that he will not sell them to Q, as he normally does for \$20, but rather he will give them to Q if and only if Q beats up a certain person.²⁶

In order to account for this example, Nozick goes on to claim that the question of whether the morally expected course of events or the morally neutral expected course of events should serve as the relevant baseline comparison to the post proposal situation should be determined by which of the two courses of events the agent *prefers*.²⁷

However, Nozick’s analysis is inadequate even following this refinement. Here, I shall press just one particular problem,²⁸ which is that even when an agent would prefer the morally expected course of events to serve as the relevant baseline, and even if the proposer only proposes to bring about some consequence that is a part of the morally expected course of events, it still seems possible for a proposal to qualify as an offer rather than a threat. This will occur in cases in which the proposing party did not *themselves* manipulate the proposed-to party’s situation so that it failed to meet the standard of what is morally expected.

²⁵ Ibid.

²⁶ Ibid., 447.

²⁷ Ibid., 451.

²⁸ For other objections, see Carr, “Coercion and Freedom,” 62; Sachs, “Why Coercion Is Wrong When It’s Wrong,” 7.

To illustrate, compare the following three cases. First, suppose that a Mafia don gets his thugs to smash the window of a shop every week; he tells the shopkeeper that he will protect the shop from similar acts of vandalism in the future if the shopkeeper pays protection money to his mob. This is structurally similar to Nozick's slave case, which, I believe we should agree with Nozick, is an instance of coercion. Now, suppose that strong winds keep throwing up stones that are breaking the shop's windows. Suppose that a builder tells the shopkeeper that he will reinforce the shop windows for a certain fee. Here, it seems that the builder is making an offer rather than a threat; again, this is congruent with Nozick's analysis. Whilst it might be unfortunate for the shopkeeper that these winds are causing stones to break his windows, it is not as if this represents a failure to meet the morally expected course of events. Moreover, in the (morally neutral) expected course of events, the shop simply continues having its window broken by stones. Thus, the consequence that the builder announces a conditional intention to bring about (having the windows reinforced) would make the shopkeeper better off.

Nozick's theory provides an adequate analysis of these two cases; however, it cannot adequately account for the following case. Suppose that random, unorganized delinquents who had no connection to the mafia were responsible for breaking the window frames, and did so purely for their own enjoyment. Now, suppose again that the mafia don tells the shopkeeper that he will protect the shop from similar acts of vandalism in the future (say by getting his thugs to 'pay the delinquents a visit') if the shopkeeper pays protection money to the don. On Nozick's account, this proposal would count as a threat; after all, we may presume that the shopkeeper would prefer the morally expected course of events in which no-one breaks his windows to serve as the relevant baseline comparison case, and the don is only proposing to bring about a

consequence that should be understood as part of the morally expected course of events. Indeed, the nature of the proposal that the don makes to the shopkeeper is, *ex hypothesi*, exactly the same as the first case, which did constitute coercion. However, it is far from clear that this third case is an instance of a coercive threat; after all, the don was in no way responsible for the fact that the shopkeeper's status quo situation failed to meet the standard of what is morally expected. It thus seems far more plausible to claim that the don is making an offer here, not a threat.

At this point, we might observe that perhaps one of the reasons that some philosophers²⁹ agree with Nozick that we need to appeal to an agent's preferences in order to determine whether a proposal constitutes a threat or an offer, is that they endorse Anderson's claim (considered at the beginning of this section) that there is no significant difference between the structures of threats and offers that could form an adequate basis for a distinction between the two. However, there is one way in which they crucially differ that Anderson's analysis fails to acknowledge. In the case of offers, the proposer will only bring about a certain consequence *C* (which, in the case of offers, is normally, although not necessarily,³⁰ something that the proposer believes will make the recipient of the proposal better off) if and only if the recipient *complies* with the proposer's demand to do *A*. In contrast, in the case of threats, the proposer will only bring about a certain consequence *C* (which, in the case of threats, is normally, although not necessarily, something that the proposer believes will make the recipient worse off) if the recipient *refuses* to comply with the proposer's demands to do *A*.

²⁹ For example, see Zimmerman, "Coercive Wage Offers"; Stevens, "Coercive Offers". For objections to these theories that are congruous with my analysis, see Swanton, "Robert Stevens on Offers."

³⁰ This clause allows my account to accommodate what Sachs calls 'ridiculously bad offers'. See Sachs, "Why Coercion Is Wrong When It's Wrong," 7.

This difference is crucial, because it means that in making a threat, the proposer always takes away the recipient's option of maintaining their status quo situation, since the proposer's threat means that the recipient's status quo situation will now be attended by some new (usually bad) consequence. In contrast, in the case of offers, the recipient will normally retain the option of choosing to maintain what the proposer believes is the recipient's status quo situation.³¹

This observation suggests a different way of drawing the distinction between threats and offers. Recall that on my view coercion undermines autonomy because the coercer exerts control over what is practically realisable for their victim, and intends to determine their victim's behaviour through their intrusion. Rather than considering whether or not the proposed-to agent would *prefer* to be in the post-proposal situation to the pre-proposal situation, I suggest that we should instead consider what effect the proposal has on the proposed-to agent's beliefs about what is practically realisable for her. More specifically, we should consider whether the proposal would serve to increase or decrease the recipient's options in comparison to their pre-proposal status quo situation. In my view, the most salient difference between threats and offers is that threats serve to take away the option of maintaining the status quo from the recipient's choice set, whilst offers normally provide the recipient with an option in addition to maintaining the status quo.

Although this appears to be a promising way of drawing a distinction between threats and offers, Nozick's slave owner case and the first Mafia don cases seem to represent counter-examples; here, the proposals provide the recipients with an

³¹ There are some exceptions to this. For example, as Dworkin points out, making a new option available can alter the status quo itself. See Dworkin, "Is More Choice Better than Less?," 51–52. I shall assume that this is not the case here.

additional option to the status quo, but Nozick seems right to claim the recipients have been coerced. In order to explain these cases, it is prudent to return to the discrepancy between the two Mafia don cases considered above. I believe, *pace* Nozick, that it is best to interpret the don's proposal to the shop-keeper as an offer in *both* the first and third case; in both cases, the proposal can be understood as giving the shop-keeper an option that he did not have prior to the proposal in his current situation. However, the difference between the two cases that makes the first case coercive is that in the first case the proposer *himself* is responsible for manipulating the shop-keeper's pre-proposal situation by reducing the latter's freedoms. Moreover, he does so with the intention of making the new additional option that he provides in his proposal compelling in comparison. Accordingly, he has *already* taken away one of the shop-keeper's options, prior to his announcing any conditional intentions in his proposal. Strictly speaking then, the content of the proposal itself, considered in isolation, is not coercive in this case; rather, the proposal is only coercive when it is considered in conjunction with the reduction of freedoms that the proposer has engineered for the purpose of getting the recipient to accept the offer.³²

However, even if the proposing agent is not responsible for manipulating the proposed-to agent's pre-situation, this does not mean that any offer they make is a moral one. For example, consider Feinberg's lecherous millionaire case; in this case, a woman needs money to pay for expensive surgery that can save her child's life, and a millionaire indicates that he will pay for the surgery if the woman becomes his

³² This response bears some similarity to Wertheimer's rights based account of coercion. See Wertheimer, *Coercion*. However, I do not endorse Wertheimer's conclusion that coercion only occurs when the coercer threatens to violate a right. Although I cannot pursue the point here, it appears plausible to me that a coercer could threaten to bring about a consequence which is sufficiently bad to make the victim change their behaviour, but which would not necessarily violate a right.

mistress.³³ Feinberg suggests that this offer is coercive; however, in light of the fact that the millionaire is not, we assume, responsible for the woman's dire situation, I think that it is more accurate to say that what we actually find morally despicable about the millionaire's offer is that it is *exploitative*. I lack the space to offer a full account of the distinction between coercion and exploitation here; however, I believe that the following is a plausible initial proposal. I suggest that the offer here is exploitative, because although the millionaire is not himself responsible for the woman's lack of desirable options, he nonetheless uses the woman's lack of desirable options as leverage to get her to do something that he wants her to do for an unfair price, and which she believes she would otherwise have strongly decisive reasons not to do. Moreover, we should also acknowledge that it is also within the millionaire's power to provide her with a far more desirable option that does not involve getting her to do something that she would otherwise not do.

However, that an offer is exploitative does not imply that the recipient must lack autonomy with respect to their accepting it; unlike cases of coercion, the party that proposes an exploitative offer has not themselves taken away any of the recipient's desirable options and thereby subjugated the recipient's will by exerting control over what is practically realisable for her. Whilst it may be true to say that the proposer could act in ways that would do *more* to further the recipient's autonomy, in so far as it is within their power to make a more desirable option available for the recipient of the offer (for instance, they could bring about the good consequence without requesting anything in return), this does not imply that in making this offer they violate the recipient's autonomy; indeed, it is not clear that we would say that the lecherous millionaire would be *raping* the woman if she were to accept his offer. In contrast, we

³³ Feinberg, *The Moral Limits of the Criminal Law Volume 3*, 229–230.

would be prepared to say this if instead the millionaire credibly threatened to kill the woman if she did not agree to have sex with him.³⁴

To return to my account of the difference between threats and offers, one might object to my analysis by arguing that it exhibits the status quo bias. However, whilst my analysis does stress the importance of the recipient's retaining their status quo option, I do not believe that it relies on the status quo bias. The reason that it does not is that it is not an irrational bias for an agent to prefer to retain the status quo option if that option is being replaced with an option that is less valuable; in such a case, it is rational to prefer the status quo. On my account, this is what is happening in cases of coercion.

That said, on my view it seems theoretically possible for a (perhaps conceptually confused) proposer to make a very poor threat in which they threaten to bring about a consequence that they believe will *improve* the recipient's situation if they don't comply with the proposer's demands. However, the fact that this would qualify as a threat is not problematic for my view, since I do not claim that such a threat would be coercive. I agree with Nozick when he claims that in order for a threat to be coercive, the consequence that the proposer threatens to bring about must be such that part of the recipient's reasons for deciding to comply with the proposer's demands is that they want to avoid the threatened consequence of refusing (see condition 5 in section I). I simply disagree with Nozick about what distinguishes a conditional threat *simpliciter* from a conditional offer.

My analysis of threats and offers here suggests a further point regarding the relationship between coercion and autonomy. I have suggested that offers can constitute part of a coercive situation if the proposing agent was responsible for

³⁴ Ibid., 242.

manipulating the recipient's pre-proposal situation by reducing the recipient's freedoms. However, it should be acknowledged that there might be cases in which people can *legitimately* reduce the freedoms of others because they have the moral authority to do so. Thus, there may be cases in which it is legitimate for one party to take away another person's option of maintaining their status quo situation. For instance, consider a mother who tells her child that if she does not eat her vegetables then she will be sent to bed without dessert. Whilst this proposal is a coercive threat on my account, we do not believe that such a proposal is morally problematic because we believe that the nature of the relationships in these cases means that it is legitimate for the proposing party to subjugate the recipient's will to their own.³⁵ Moreover, if a proposing agent alters the recipient's pre-proposal situation by legitimately reducing their freedom (as is the case in justified state incarceration for example), prior to making them a compelling offer, this does not mean that they have engineered a coercive situation; the manipulation of a person's pre-proposal situation will only partly constitute a coercive situation if it involves an illegitimate reduction of the recipient's freedom.

Conclusion

³⁵ Carr suggests that this example shows that Nozick's account is too broad, since this proposal would count as a coercive threat on Nozick's account, and this he claims is implausible. Carr, "Coercion and Freedom," 62. However, it is not clear why we must share Carr's intuition that the mother's proposal here is not coercive. Carr is right in his implicit claim that the threats in these cases are morally acceptable. However, we can similarly accommodate this thought by claiming that this is simply an example of *legitimate* coercion.

I began this chapter by suggesting that coercion can undermine autonomy in so far as the coercer exerts controls over about what is practically realisable for their victim. I claimed that this undermines autonomy because autonomous agents choose to sustain their motivating desires in the light of what they believe is practically realisable. In cases of coercion, the coercer intends to determine their victim's behaviour through their intrusion into their victim's freedoms; they intend to subjugate their victim's will. In view of this analysis, I then suggested that threats can be distinguished from offers by virtue of the fact that they serve to take away the desirable option of maintaining the status quo from the recipient's choice set, whilst offers provide the recipient with an option in addition to maintaining the status quo. However, I claimed that if the agent proposing the offer is responsible for manipulating the proposed-to party's pre-situation by illicitly transgressing on their freedoms in order to make their offer attractive, then they can be understood as having engineered a coercive situation, of which their offer is an integral part.

This analysis suggests a framework for how we ought to consider the question of whether an agent could autonomously accept a seemingly coercive offer. Contrary to some analyses in the literature, this question cannot be settled solely by considering whether the offer could reasonably be refused, the preferences of the proposed-to party, or indeed whether the proposer intends that their offer be accepted.³⁶ Rather, in considering this question, we must investigate whether the proposing party has illicitly transgressed the proposed-to party's freedoms in order to engineer a situation in which the latter has little choice but to act in accordance with the coercer's interests.

³⁶ For examples of such arguments, see McMillan, "The Kindest Cut?".

This chapter concludes my investigation into the nature of personal autonomy. In the next chapter, I shall investigate the prudential value of autonomy, before going on to consider the some practical applications of the theory of autonomy that I have developed in this thesis.

Chapter Five - On the Prudential Value of Autonomy

In the preceding chapters, I have outlined a theory of personal autonomy; in this chapter I shall investigate the prudential value of this sort of autonomy. I shall suggest that whilst personal autonomy might be instrumentally valuable in this way, I shall defend the stronger claim that it is also valuable for its own sake. I shall go on to consider some of the prominent values that autonomy can conflict with in a bioethical context, and suggest that different sorts of autonomy may not be valuable to the same degree.

Prior to beginning this investigation, I shall mention briefly one topic concerning the value of autonomy that I shall not discuss here. It is sometimes claimed that autonomy has *moral* as well as prudential value, in so far as autonomy is sometimes understood to undergird the moral value of personhood.¹ Modern statements of this view commonly find their source in Kant's moral philosophy, and his substantive account of autonomy.² Whatever the merits of this view, I shall not discuss it here. As I have already discussed, my view of autonomy departs from Kant's substantive conception;³ accordingly, establishing that my understanding of autonomy can provide a foundation for the moral value of personhood would require lengthy argument. Whilst this would be an interesting project,⁴ here I shall focus my attention on whether autonomy bears

¹ Kant claimed that autonomy is the ground of human dignity, that is, the basis of the value that makes a human life supremely valuable. Kant, 4:435.

² For some examples, see Velleman, "A Right of Self-Termination?", Darwall, "The Value of Autonomy and Autonomy of the Will." Habermas, *The Future of Human Nature*, particularly 37–44.

³ Introduction, 2–3.

⁴ Jeff McMahan makes some remarks on this sort of project, and endorses the view that personal autonomy is a significant basis of the moral worth of persons. McMahan, *The Ethics of Killing*, 256–260.

prudential value. Finally, I shall not consider whether people have a ‘right to autonomy’.⁵ However, in so far as having a ‘right to x ’ can be understood to entail that x is of significant prudential value to the bearer of that right, my discussion of the prudential value of autonomy will not be entirely orthogonal to those interested in questions pertaining to a right to autonomy.

I The Nature of Autonomy’s Prudential Value

To begin, it is important to distinguish two ways in which something can be prudentially valuable. Consider first what Korsgaard terms ‘final value’.⁶ According to Korsgaard, we may say that something bears final value if it is valuable as an end, or for its own sake; for instance, one might argue that knowledge, happiness, and virtue, *inter alia*, can be understood as bearing final value. We can contrast final value with what Korsgaard terms ‘instrumental value’; something has instrumental value if it is only valuable for the sake of something else.⁷ For instance, we typically think that money has only instrumental value, in so far as it can be exchanged for other goods that will lead to

⁵ Beauchamp and Childress, *Principles of Biomedical Ethics*, 63.

⁶ Korsgaard, “Two Distinctions in Goodness.” Korsgaard’s aim in this paper was to separate the distinction between final and instrumental value from the distinction between intrinsic and extrinsic value. The latter distinction pertains to whether or not something bears value in virtue of its intrinsic, non-relational properties, that is, ‘in itself’. Although philosophers sometimes claim that autonomy has ‘intrinsic’ value, it seems that this is most naturally understood as the claim that autonomy has ‘final’ value. Whilst we may value autonomy as an end in-itself, it is not clear that we value it by virtue of its non-relational properties. See Wall, *Liberalism, Perfectionism and Restraint*, 145 for discussion.

⁷ Korsgaard, “Two Distinctions in Goodness,” 170.

the attainment of some further valuable end. Although I shall argue that autonomy may be instrumentally valuable, I shall also defend the claim that it has final value.

In view of the above definition of instrumental value, it seems that if we are to claim that autonomy has instrumental value, we ought to be able to give an account of what valuable end autonomy is a reliable means *to*. One plausible candidate might be well-being, broadly construed;⁸ a life lived autonomously, it might be claimed, is more likely to lead to the attainment of the goods that make for a good life. However, whilst this seems intuitively plausible, it only takes us so far, since this claim leads one to the deeper question of what constitutes the good life.

According to Parfit, there are broadly three types of theory of well-being, as schematised below:

<i>Hedonistic Theories</i> –	What would be best for someone is what would make their life happiest.
------------------------------	--

<i>Desire-Fulfilment Theories</i> -	What would be best for someone is what, throughout their life, would best fulfil their desires.
-------------------------------------	---

⁸ It might be claimed that autonomy is also instrumental to the attainment of other valuable goals such as self-development. Whilst this might be true, it seems plausible that the reason we value things like self-development is that we believe that self-development is either conducive to or incorporated in the fully good life.

Objective List Theories - Certain things are good or bad for us, whether or not we want to have the good things, or to avoid the bad things.⁹

Crisp suggests a further distinction between *enumerative* and *explanatory* theories of well-being. On this distinction, the former sort of theory seeks to answer the question ‘which things make someone's life go better for them?’ In contrast *explanatory* theories of well-being seek to explain what it is about the things listed by enumerative theories of well-being that make them good for people.¹⁰

The claim that autonomy is *only* instrumentally valuable is perhaps most congruent with explanatory hedonism; on such a theory, it might be claimed that autonomy makes a life go better just because autonomy is conducive to happiness, which is the only thing that has final value on this view.¹¹ As Young points out, this understanding of the value of autonomy is commonly, although perhaps mistakenly, attributed to Mill.¹² Such a reading of Mill might seem natural, given his insistence at

⁹ Parfit, *Reasons and Persons*, Appendix I. Although this tripartite classification is widely accepted, it has recently come under criticism, partly because it ignores Crisp's distinction between enumerative and explanatory theories. See Woodard, “Classifying Theories of Welfare”. In the interests of clarity and space, I shall follow philosophical orthodoxy in discussing the tripartite classification, but I shall supplement this discussion with considerations pertaining to Crisp's distinction.

¹⁰ Crisp, *Reasons and the Good*, 102–103.

¹¹ Happiness here is to be broadly understood in terms of the experience of pleasure (or desirable consciousness) and the absence of pain.

¹² For examples of this sort of interpretation of Mill, see Berlin, “John Stuart Mill and the Ends of Life” and Ladenson, “Mill's Conception of Individuality.” The problem with this interpretation is that it fails to acknowledge the way in which Mill departed from Bentham's monistic conception of utility. As Young points out, Mill's view actually seems to be that autonomy is incorporated into his understanding of utility. Young, “The Value of Autonomy,” 36.

the beginning of *On Liberty* that he regards utility as “ . . . the ultimate appeal on all ethical questions” (a position that he defended in his *Utilitarianism*).¹³ Moreover, this understanding might seem plausible in view of the fact that individuals seem to be in a privileged epistemic position with regards to the question of what will make them happy. As Mill himself puts the point:

With respect to his own feelings and circumstances, the most ordinary man/woman has means of knowledge immeasurably surpassing those that can be possessed by anyone else.¹⁴

However, although it seems right to claim that autonomy can be instrumentally valuable in this way, it seems problematic to claim that autonomy is valuable *only* in so far as it is a means to happiness. First, although Mill’s claim in the above quote is plausible, individuals will still often be mistaken about what will make them happy;¹⁵ thus, they may in fact achieve *less* happiness if they are autonomous than they would have done otherwise. For instance, we can imagine a young man who had autonomously decided that a career in finance would make him happy, but who comes to regret this decision in later life, when he realises that he did not enjoy his career, and his choice meant forgoing a family life that he now realises would have made him happy.

¹³ Mill, *On Liberty*, 31.

¹⁴ *Ibid*, 91.

¹⁵ For a similar point, see Hart, *Law, Liberty and Morality*, 32. Dworkin argues that Mill himself was also aware of this point. Dworkin, “Paternalism,” 73–74.

Of course, this observation alone is not an unimpeachable objection to the explanatory hedonist's claim that autonomy is only instrumentally valuable; she may reply that most people *do* know what will make them happy, and counter-examples show only that there can be exceptions to this rule. In order to provide a stronger argument against the view in question, one would need to show that a life lived in the absence of autonomy could be *worse* than a life lived autonomously, even if the former life involved more happiness.

A number of philosophers have objected to explanatory hedonism by providing examples in which this criterion is met. For example, Wall suggests that we do not believe that our lives would go better if we left our affairs to a friend who we knew to be both wise and benevolent;¹⁶ similarly, Griffin writes:

. . . even if you convince me that, as my personal despot, you would produce more desirable consciousness for me than I do myself, I shall want to go on being my own master.¹⁷

I believe that the plausible thought underlying these claims is that autonomy has a special sort of value for us; there seems to be a value in *living a life of one's own* that is of central and fundamental importance to many of us.¹⁸ Such examples, I propose, suggest that autonomy bears final value; we want to be autonomous for its own sake,

¹⁶ Wall, *Liberalism, Perfectionism and Restraint*, 146.

¹⁷ Griffin, *Well-Being*, 9.

¹⁸ See also Wall, *Liberalism, Perfectionism and Restraint*, 146–7; Glover, *Causing Death and Saving Lives*, 96.

and not just because we believe that being autonomous will lead to our attaining other prudentially valuable ends.

Whilst the above examples raise problems for explanatory hedonism, both desire-fulfilment and objective list theories of well-being can accommodate our intuitive response to these examples, and allow for the view that autonomy has final value. On a desire-fulfilment theory, it might be claimed that even if a personal despot could produce more happiness in your life, she would not be able to fulfil a non-instrumental desire to live an autonomous life. Alternatively, an objective list theory might simply claim that autonomy is an end that has final value. Indeed, many modern theorists have incorporated autonomy into their understanding of well-being in these ways.¹⁹ For instance, the desire for autonomy is a central desire in Griffin's informed desire account,²⁰ and Sumner claims that well-being consists in “ . . . authentic happiness, the happiness of an informed and autonomous subject”.²¹ In a similar vein, Finnis' description of the good of 'practical reasonableness' included in his objective list account seems to bear a close relation to autonomy as I have understood it in this thesis.²²

However, it should be acknowledged that there are important differences in how different theories of welfare account for the value of autonomy. For instance, on enumerative *actual present* desire-fulfilment theories, autonomy is only incorporated

¹⁹ Notice that a further benefit of incorporating autonomy into one's theory of welfare is that such theories are able to explain why the satisfaction of adaptive preferences may not enhance well-being. Sen raises this point in Sen, *Resources, Values and Development*, 304. See Sumner, *Welfare, Happiness, and Ethics*, 166 for discussion.

²⁰ Griffin, *Well-Being*, Part One, particularly 33-36.

²¹ Sumner, *Welfare, Happiness, and Ethics*, 172.

²² Finnis, *Natural Law and Natural Rights*, 88-90.

into the good life for a particular person if they desire it. In contrast, on enumerative objective list theories that include autonomy, the final value of autonomy is not contingent upon the subject's desires in this way. I lack the space here to defend a full view of welfare here. However, it should be acknowledged that the Parfitian view of rational desires that I have defended in this thesis is based in part on a rejection of the view that our desires *simpliciter* can provide us with reasons. As such, the view of reason, value and autonomy that I have endorsed seems to be incompatible with a purely desire-based *explanatory* account of well-being, since on such an account, the fact that something satisfies one of our desires makes that thing good for us; this sounds suspiciously like subjectivism about reasons that Parfit spends Part One of *On What Matters* rejecting.²³ Accordingly, when Parfit claims that his object-given view of reasons is compatible with a subjective desire-based account of well-being (as I discussed in chapter two),²⁴ we should understand him to be referring to an *enumerative* desire-fulfilment account theory of well-being.

In view of the failure of explanatory hedonism to adequately capture the value of autonomy, and the incompatibility of purely desire-based explanatory accounts with Parfit's objectivism about reasons, how should we understand the claim that autonomy has final value? I believe that the most plausible strategy here is to endorse an explanatory theory of well-being that appeals to objective values, and to claim that autonomy is one of the things that has such objective, final value. However, autonomy should not be understood to be the only good on this sort of theory; for instance, it seems clear that an adequate theory of well-being should allow for the possibility that

²³ See Chapter Two, 62-64.

²⁴ Chapter Two, 70-71. See also Parfit, *On What Matters*, 74.

pleasurable experiences can contribute to well-being. Moreover, in section IV, I shall point out that the realisation of some values may require the absence of autonomy.

One common objection to objective accounts of the sort that I have sketched here is that subjective experiences have an important influence on her well-being. I lack the space to adequately deal with this point here; suffice to say that if one finds this objection convincing, then I suggest that we respond to it by adopting a hybrid account, according to which an agent must have some subjective positive attitude towards the realisation of certain objective values in order for these to contribute to her well-being; such an account thus incorporates both objective and subjective elements. Indeed, this is the approach that Parfit himself, amongst others, seems to take.²⁵ The plausibility of such an account stems from the fact that although we may have reason to doubt that a theory of well-being that completely ignores individual preferences and attitudes is mistaken, it also seems plausible to claim, as Parfit does, that we can have self-

²⁵ Parfit, *Reasons and Persons*, Appendix I. See also Adams, *Finite and Infinite Goods*, 95–101; Feldman, *Pleasure and the Good Life*. Darwall also endorses an enumerative hybrid account in *Darwall, Welfare and Rational Care*, Ch. 4. However, his explanatory account in this book seems to be in tension with my claim that autonomy bears final value. On his rational care explanatory account of welfare, Darwall claims that the statement ‘x is good for y’ simply means that ‘x is something that it would be rational for someone who cared about y to want for y for y’s own sake’ Ibid., 8. However, he later distinguishes caring for someone from respecting them, claiming that care involves relating to them as a being with welfare, whilst respect requires relating to them as a being with dignity, which they have by virtue of being capable of free agency. Ibid., 14–15. Whilst Darwall’s interpretation is a novel approach, I shall not respond to it here for two reasons. First, as well as being somewhat obscure (see Feldman, “What Is the Rational Care Theory of Welfare?”), Darwall’s explanatory theory seems to be open to a number of powerful objections (see Griffin, “Darwall on Welfare as Rational Care”. See also Heathwood, “Review of Rational Care and Welfare.”) Second, Darwall does not discuss any of the literature concerning the importance of autonomy to welfare. In the absence of such discussion, I believe that his distinction between respect and care is premature.

interested reasons to want certain things that we do not believe will cause us happiness, or that we do not desire.

Although many find the claim that autonomy bears final value in the way that such a theory would claim to be intuitively plausible, Valdman has recently objected to this view by questioning the validity of basing such a claim on our intuitive reaction to examples like Griffin's personal despot case above. In the next section I shall respond to this objection.

II Valdman's Objection

Valdman argues that examples such as Griffin's personal despot case conflate the value that we may attach to making decisions for ourselves with "... the value we attach to having our decisions reflect our deepest goals and values",²⁶ that is, living what he terms an "acceptable life".²⁷ To illustrate this, he suggests a thought experiment in which you have the opportunity to cede your final decision-making authority about how to act to a Personal Expert Committee (PC); Valdman stipulates that this committee is better than you are at determining how to accomplish your goals and how to live according to your values.²⁸ Valdman asks whether we should prefer the PC to self-government, and suggests that we should.²⁹

²⁶ Valdman, "Outsourcing Self-Government," 764.

²⁷ *Ibid.*, 769.

²⁸ *Ibid.*, 770.

²⁹ *Ibid.*, Section II.

Valdman takes care to pre-emptively respond to a number of objections to his arguments.³⁰ Whilst I do not believe that all of these responses are satisfactory, I shall not pursue them here. Rather I shall raise two new objections to Valdman's PC argument. The first objection calls into question the scope of his PC example; the second objection suggests that, rather than the PC example showing that self-government has no intrinsic value, it merely indicates that different elements of autonomy can have different value.

According to Valdman, although the PC would intervene when it detected flawed practical reasoning, it would always use the agent's own goals and values as the basis for its decisions. To illustrate, suppose that David has some prudential goal *X*, and has to choose between two possible acts *A* and *B*. Suppose that out of these two acts, only *A* would serve as a reliable means to David's achieving *X*; in this case, the PC would only intervene if David believed that he prudentially ought to *B*. This model is not problematic when the value of the goal (*X*) is distinguishable from the acts that one must perform as a means to achieving that goal. However, it is problematic when this is not the case.

Unfortunately for Valdman's argument, it seems that the value of certain goals is inextricably related to the way in which we achieve that goal. Suppose that one valued playing a complex piece of music on the piano, say Rachmaninov's second concerto. In a crude sense, in order to play this piece, one would simply need to hit certain combinations of keys, in a certain order, for a certain time. In order to be able to do *this*, one would need to develop excellent motor skills and technique, normally through devoting hours to practicing the requisite movements, and to learning the

³⁰ Ibid., 780–789.

structure of the piece. Whilst it might be claimed that there is some value in the discipline and effort that this practice requires, it seems that the goal of being able to play Rachmaninov's concerto in the crude sense under consideration could retain its value for an agent, even if they achieved it via a more efficient means that did not involve effort or discipline; for example, instead of sitting through hours of lessons and practice sessions, suppose (somewhat fantastically) that one could simply 'download' the ability to play the right notes in the right order for the right amount of time.

In this crude sense of being able to play the piece, the value of the goal is distinguishable from the means that one takes to achieve it. However, consider now someone who has a more refined desire to be able to play the Rachmaninov piece; rather than valuing being able to simply 'play the right notes', this person values being able to play the piece according to their own interpretation of the music. This might involve, *inter alia*, their deciding which phrases of the piece need particular emphasis, and the strength they should exert in pressing the keys at particular points. Whilst the achievement of this goal requires the same abilities as the goal of playing the piece in the crude sense, it also requires something more, something like *creativity*; and because of this, it seems that the value of the goal is inextricably linked to the fact that the agent herself exercises *her own* creativity in its pursuit.

This is important, since in such a case, it does not make sense to say that one might be able to achieve this goal better by outsourcing to something like a PC. A PC, an expert tutor, or a futuristic downloadable music program could make you a better technical piano player; and this technical ability might be prerequisite for going on to exercise one's creativity in playing. However, completely relying on a PC to realise the

goal of playing Rachmaninov's second concerto in a sophisticated sense would defeat the value of the goal itself.

The point that this example raises is that the relationship between the value of our goals and the means that we take to achieve them is not always as simple as Valdman's argument implies. Whilst Valdman is correct to point that we often outsource decision-making authority, the examples he highlights are cases in which the value of the goal is clearly distinguishable from the manner in which the goal is achieved; for example, the value we attribute to achieving financial security is not taken away if we attain it by allowing a financial advisor to make our financial decisions for us.³¹ However, in more complex cases, the value of some goals seems to be at least partly dependent on the fact that in achieving the goal, the agent herself makes her own mark in doing so. Goodman captures a similar thought in his distinction between 'process goods' that pertain to excellence in the performance of an activity, and 'outcome goods', that pertain to the benefits that an activity creates.³² Playing Rachmaninov in the crude sense would qualify as an outcome good in my example, whilst playing the piece in the sophisticated sense would involve process goods.

Valdman might respond to this objection by claiming that cases of the sort that I am discussing here are rare, and that his claims still stand in relation to the majority of cases of outsourcing self-governance. Whilst I believe that many of the goals that agents tend to have involve process goods, let us suppose that Valdman's objection still stands in relation to a number of goals that people tend to have; I shall now argue that Valdman's arguments only show that the value of different sorts of autonomy can come into conflict.

³¹ Ibid., 772.

³² Goodman, "Cognitive Enhancement, Cheating, and Accomplishment," 146 and 152–154.

There are two central points undergirding this line of response to Valdman's argument. The first is that on Valdman's view, one may fail to be self-governing even if one is living in accordance with one's own goals and values; on his view, one will fail to be self-governing if it is the PC rather than the agent herself who ensures that they live in accordance with their goals and values. The second concerns the distinction that I have drawn upon in this thesis between global and local autonomy. As I explained in the introduction of this thesis,³³ we can understand autonomy to be a property of agents in a particular time-slice, with respect to a particular decision. When we conceive of autonomy in this way, we are considering *local* autonomy. In contrast, we can also understand autonomy as a *global* property that agents can instantiate diachronically.

When Valdman claims that one may fail to be self-governing even if one is living in accordance with one's own goals and values, the failure he is speaking of seems to be a failure of local, rather than global autonomy. After all, Valdman himself stipulates that the PC would only govern you in accordance with your own deeply held commitments and values. As such, the PC will only intervene when one's own *local* decision-making is likely to prove counter-productive to one's pursuit of the long-term goals that may be understood to undergird one's *global* autonomy.

The reason that Valdman's arguments appear so convincing is that he fails to adequately distinguish local and global autonomy. Although Valdman might be right to claim that there are cases in which we could have good reason to outsource our decision-making to experts, the strength of this reason is *itself* rooted in the value of being able to live what he calls an 'acceptable' life, in accordance with one's own freely chosen goals and values; however, it seems that this is simply what it is to be *globally*

³³ Introduction, 16-20.

autonomous. Accordingly, Valdman's arguments are only sufficient for proving that the value of local and global autonomy may sometimes be in conflict, and that we would often prioritise our global autonomy over our local autonomy. Yet, this is not a problematic conclusion for those who claim that autonomy bears final value. Indeed, as I shall explore in the next section, there are a number of real life cases in which it seems that an agent's global autonomy might best be served by over-riding their local autonomy.

III The Value of Different Sorts of Autonomy

My response to Valdman's argument turns on the claim that it is possible for local autonomy to come into conflict with global autonomy. On some views of the relationship between local and global autonomy, this claim would be implausible. As I explained in the introduction, Christman claims that global autonomy is simply an aggregate of the instances of local autonomy over time.³⁴ However, I suggested an alternative understanding of the relationship, according to which an agent's global autonomy depends on the extent to which she lives in accordance with her own diachronic plans and commitments. This understanding allows for possible conflicts between local and global autonomy.

Interestingly, Valdman himself implicitly highlights one possible explanation of why global autonomy will often have precedence over local autonomy. As Valdman suggests, the deep commitments that one must live in accordance with to live an

³⁴ Ibid, 18-19.

‘acceptable’ life are central to our personal identity on certain psychological theories of that concept.³⁵ To develop Valdman’s observation, we might point out that the goods that we attempt to pursue in our locally autonomous decision-making may sometimes be trivial, and in no way connected to any of our deep global commitments; I can, for instance, be locally autonomous with respect to my decision about what to have for lunch. As such, when the two sorts of autonomy cannot both be realised, it seems that worries concerning narrative identity may give us reason to prioritise our global autonomy over our local autonomy.

These observations are of more than purely theoretical importance, for there are cases in contemporary bioethics in which we have to choose whether to respect an individual’s local autonomy or their global autonomy. Consider the use of deceptive placebos in clinical practice. Occasionally, a patient may suffer from a condition for which there is no available medically active treatment. In such cases, it may be that an inert placebo could ameliorate the patient’s condition. However, a great deal of evidence suggests that the placebo effect is strongest when the patient is not aware that the treatment that they are receiving is a placebo.³⁶ Accordingly, in such cases, the physician faces a dilemma: On the one hand, in order to respect the patient’s local autonomy with respect to their treatment decision, it seems that they ought to inform the patient that they are being prescribed a placebo;³⁷ however, this will make it less likely that the placebo will be therapeutically efficacious. On the other hand, it seems that the patients’ most salient global commitment over the course of their illness will often be,

³⁵ Valdman, “Outsourcing Self-Government,” 768.

³⁶ Foddy, “A Duty to Deceive,” 9.

³⁷ I shall consider the nature of informed consent in chapters seven and eight.

as Kolber suggests, the goal of simply feeling better.³⁸ Thus, it might be argued that the patient's global autonomy in this instance would be best served by prescribing them a placebo deceptively, if that is most likely to help them achieve this global commitment.

There are a number of complexities to this bioethical problem that I cannot explore here.³⁹ However, it is an interesting case to consider because it illustrates the way in which local and global autonomy can conflict in bioethics. Those who defend the view that deceptive placebo use is permissible in such cases deny that having local autonomy with regards to a decision about the means one uses to pursue one's global commitment is always more important than preserving one's ability to effectively pursue that end at all.

In a similar vein, it might sometimes be the case that patients believe that the best way to achieve their global commitments in a medical context is to sacrifice their local autonomy with respect to their treatment decisions by telling their doctor to 'do what you think would be best'. On the view of autonomy that I have defended here, this may be an expression, rather than an abdication of autonomy. For instance, the patient might not trust herself to make a difficult decision that is in harmony with her evaluative judgments, or she may not feel able to weigh the complex information involved in such a decision appropriately. Crucially, in light of my above objection to Valdman, the amelioration of a patient's condition is most naturally understood as an outcome good, rather than a process good; as such, I claim that the patient may outsource her decision-making here without this undermining the value of her diachronic goal. Accordingly, I suggest that an agent may retain her global autonomy in

³⁸ Kolber, "A Limited Defense of Clinical Placebo Deception".

³⁹ I provide a fuller defence of the practice in question in Pugh, "Ravines and Sugar Pills: Defending Deceptive Placebo Use."

making this request if the patient believes that the doctor is more likely than she is to make a treatment decision that will lead to the achievement of the outcome that would best reflect her evaluative judgement about what would be good for her in a reason-implying sense.

To conclude this section, we might also consider whether the reflective and practical dimensions of autonomy may be valuable to different degrees. In my discussion of the reflective and practical dimensions of autonomy, I claimed that the reflective dimension of autonomy is theoretically prior to the practical dimension.⁴⁰ This analysis might tempt one to claim that whilst the reflective dimension of autonomy might bear final value, the practical dimension of autonomy might bear only instrumental value.⁴¹ However, I believe that this thought should be resisted, since it fails to adequately capture the point that the way in which we value autonomy for its own sake is, as I suggested above, inextricably related to our fundamental interest in *living* a life that is our own.

I have already suggested that it may be problematic to draw a discrete distinction between the reflective and practical dimensions of autonomy, because our beliefs about what we are free to do influence our reflections about the outcomes that we believe we have reason to pursue in our particular circumstances. However, even if it were possible to separate the two dimensions of autonomy into discrete categories, neither dimension *by itself* seems sufficient for the project of living a life of one's own. This point is perhaps clearest with respect to practical autonomy; the fact that an agent is able to act effectively in pursuit of an end that she does not autonomously desire does not seem to

⁴⁰ Chapter Three, 84.

⁴¹ In a similar vein, Taylor claims that increasing an agent's freedoms does not increase her autonomy, but rather increases the value of her autonomy. Taylor, *Practical Autonomy and Bioethics*, 6.

be valuable for its own sake; the freedom to pursue a goal only seems to have *final* value when the agent is reflectively autonomous with respect to their endorsement of the goal in question;⁴² only then can the agent be said to be exercising her *own* agency.

A similar point can be made with respect to reflective autonomy; however, this point is perhaps less immediately obvious. The reason for this, I submit, is that it is difficult to imagine cases in which an agent lacks *all* practical freedoms that are relevant to their practical autonomy. To illustrate, reconsider the case of Epictetus. Despite his being enslaved, and thus seemingly lacking *any* freedom, one might claim that Epictetus nonetheless represents the epitome of autonomy, in so far as he defied his lack of freedom by spending his life in the pursuit of a self-determined goal; namely the pursuit of philosophical truth. However, this example does not demonstrate that reflective autonomy *alone* is valuable for its own sake. Although Epictetus lacked many practical freedoms, he crucially retained the freedom to act effectively in pursuit of his goal of philosophizing; to this extent, he was thus *both* reflectively and practically autonomous.

In view of these reflections, I suggest that we ought to reject the claim that one dimension of autonomy is more prudentially valuable than the other. I am inclined to claim that neither dimension of autonomy in abstraction from the other is prudentially valuable for its own sake. Rather, we should understand the conjunction of the two dimensions of autonomy to form an organic whole which is prudentially valuable for its own sake, and whose value is derived from our fundamental interest in the exercise of our own agency, of living a life that is our own.

⁴² That is not to say that such freedoms cannot have instrumental value; the effective pursuit of non-autonomously endorsed goals could lead to other goods such as pleasure. Contrary to what I have claimed here, Feinberg has argued that freedom has intrinsic value. See Feinberg, *The Moral Limits of the Criminal Law Volume 3*, 211–212. For a comprehensive rebuttal of these arguments see Haworth, *Autonomy*, 139–147. See also Griffin, *Well-Being*, 237.

IV Autonomy and Conflicting Values in Bioethics

Having considered the nature of the value of autonomy, I shall in the remainder of this chapter consider the extent to which the autonomy may come into conflict with prominent values in bioethics. Prior to doing so, it is crucial to first establish that autonomy can conceptually come into conflict with other values at all. On some views autonomy cannot, strictly speaking, come into conflict with other values because autonomy itself is understood to be the fundamental value that is the source of all other values.⁴³

I mention this so-called ‘autonomism’ view, most prominently endorsed by Lawrence Haworth, only to reject it. I do so for the following reasons: First, there are some values that autonomy cannot plausibly be understood to be the source of, because their very possibility pre-supposes the absence of autonomy. For instance, Berofsky points out that values such as personal strength and dignity can be found in communities that eschew autonomy, and, moreover, that the presence of these values in such communities is “. . . in part a function of the very absence of individual autonomy”.⁴⁴ Second, it seems that we can make sense of a life having *some* value even if it lacks autonomy. Whilst we may disagree with the hedonist’s claim that we should hand over control of our lives to a benevolent personal despot if we had the chance, this does not entail that such a life would lack *any* value; *ex hypothesi*, it would contain a great deal of pleasure. It seems implausible to claim that this pleasure would count for *nothing* simply because the agent in question lacked autonomy.

⁴³ Haworth, *Autonomy*, 7 and 184.

⁴⁴ Berofsky, *Liberation from Self*, 248. Oshana also posits ‘security’ as a value that can conflict with autonomy. See Oshana, “How Much Should We Value Autonomy?”, 113-114.

The truth in autonomism is that autonomy seems to be related to a particular sort of value that we understand to be salient; I have described this as the value of ‘living a life of one’s own’. It is through pursuing our own goals that we can understand ourselves as living a life that is *ours*; and whilst this need not be understood as either a necessary foundation of *all* other values, or even as something that everyone does value, it seems plausible to claim that living a life that is one’s own has a particularly high prudential value for many people. As I intimated above, perhaps one explanation of this is that only in a life in which the agent is autonomous with respect to the sustainment of the fundamental commitments that guide her conduct is it the agent herself, in the narrative sense of identity, who realises all of the other values instantiated in that life.

With this in mind, I shall now consider how autonomy can come into conflict with other values in bioethics. As I pointed out in the introductory chapter, Beauchamp and Childress propose that biomedical ethics should be governed by four ethical principles; namely, the principle of beneficence, the principle of non-maleficence, the principle of autonomy, and the principle of justice.⁴⁵ They also claim that none of these principles takes priority over any of the others;⁴⁶ as such, it may be that principle of autonomy might come into conflict with any one of these principles, and the values that they represent. For instance, it seems that conflicts between the values of autonomy and justice will often arise in the context of health resource allocation; since the demand for many health resources (such as organs for transplantation and hospital ward space) far outstrips supply, it is not the case that the autonomous wishes of all the patients who wish to use these resources can be respected. Indeed, when societies have to make decisions about health resource allocation, considerations of justice will often trump

⁴⁵ Introduction, 1.

⁴⁶ See Beauchamp, *Principles of Biomedical Ethics*, especially 57 and 177.

considerations of individual autonomy. Despite his staunch defence of liberty, even Mill claimed that considerations of justice can trump the individual's right to liberty in this way. This thought is apparent in his 'Harm Principle', according to which:

The only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others.⁴⁷

Perhaps the most interesting theoretical value conflicts in this domain are those that involve the principle of autonomy and the principle of beneficence. The latter principle, according to Beauchamp and Childress, enjoins physicians to act in a manner that will *benefit* their patients. The extent to which the principle of beneficence and the principle of respect for autonomy can come into conflict depends on the extent to which we believe that patients can make autonomous choices that are not in their interests. This in turn, will depend on the way in which we understand the concept of well-being and its relation to autonomy.

On a view that is commonly endorsed, the concepts of autonomy and beneficence are understood to represent two distinct domains; the question of what is in a patient's interests is understood to be a conceptually different question to the question of what a patient autonomously desires. For instance, in introducing the concept of beneficence, Beauchamp and Childress point out that "(m)orality requires not only that we treat persons autonomously . . . but also that we contribute to their welfare."⁴⁸

⁴⁷ Mill, *On Liberty*, 30.

⁴⁸ Beauchamp and Childress, *Principles of Biomedical Ethics*, 165.

Whilst it would be a mistake to completely collapse the distinction between these two concepts, I believe that the commonly endorsed view is overly simplistic as it stands; it overlooks the point that I have defended in this chapter, namely that autonomy may plausibly be understood as playing an important role in a person's well-being. Treating persons autonomously and contributing to their welfare should not be understood as distinct requirements; in order to adequately contribute to a person's welfare, we must at least take into account the agent's own autonomous preferences.

The commonly endorsed view seems to rely on an overly objective conception of well-being. On purely objective accounts of well-being, it might be claimed that there are certain things that are intrinsically good or bad, that all agents have impersonal self-interested reasons to either want or avoid, regardless of their own attitudes towards these outcomes. On such a view of well-being, conflicts between the principle of autonomy and the principle of beneficence will arise whenever an agent autonomously desires to act in a manner that conflicts with what is objectively good for them. This view of well-being naturally lends support to two types of paternalism; first, what Feinberg terms 'hard paternalism', and second, what Dworkin calls 'strong paternalism'. According to hard paternalism, a third party may permissibly interfere with even an agent's voluntary choices in order to protect them from the harmful consequences of those choices; by way of contrast, soft paternalism only permits a third party to interfere with an agent's involuntary choices.⁴⁹ Dworkin also draws attention to a distinction between what he calls weak and strong paternalism.⁵⁰ According to weak paternalism it is legitimate to interfere with the *means* that agents choose to achieve

⁴⁹ Feinberg, *The Moral Limits of the Criminal Law Volume 3*, 12–13.

⁵⁰ Hard and strong paternalism are not necessarily coextensive. For example, Conly's so-called 'coercive paternalism' is hard but weak. See Conly, *Against Autonomy*. For a critique of Conly's views, see Pugh, "Coercive Paternalism and Back-Door Perfectionism."

their ends, if those means are likely to defeat their own ends. According to strong paternalism, it is legitimate to interfere to prevent people from achieving those ends that they are mistaken in believing to be good for them.⁵¹

However, the view of well-being that the commonly endorsed view seems to rely on is unattractive. This claim may seem somewhat surprising, since in section I, I endorsed an explanatory account of well-being that appeals to objective values. However, what is problematic about the view that I am considering here is not that it relies upon the claim that there are objective elements of well-being; the account of well-being that I sketched above also makes this claim. Rather, the problem with the view that I am considering here is that it seems to implicitly assume that there is an objective *ranking* of the different objective elements of well-being. This assumption is problematic for the following reason. As I pointed out in chapter two,⁵² even on theories of well-being that incorporate *only* objective elements, agents can still rationally disagree about the relative strengths of the self-interested reasons that different objective goods imply; following Parfit, I suggested that truths concerning the comparative strength of such reasons are often very imprecise. Thus, even if we accept a purely objective list theory of well-being, we need not accept the claim implicit in the view of well-being that I am rejecting here, that the goods in this list must have a set degree of goodness for everyone, or that there is a supreme value that overrides others on the list. This point is all the more powerful if we endorse a hybrid view of the sort that I sketched at the end of the section I, which incorporates both objective and subjective elements of well-being.

⁵¹ Dworkin, "Paternalism (SEP Entry)."

⁵² Chapter Two, 70-71.

On the view of the relationship between autonomy and well-being that I developed in section I of this chapter, conflicts between autonomy and beneficence will be less commonplace than on the commonly endorsed view that I have discussed here; as long as an individual's choice is autonomous, that should give us at least a *pro tanto* reason to believe that respecting that choice will benefit that person, not because the choice is likely to lead to greater happiness (as the explanatory hedonist might claim), but rather because on this view there is some value to directing the course of one's life in accordance with one's own beliefs about what is of value, and with one's own beliefs about which values should take precedence.

The view that I endorse does not entail that there *cannot* be conflicts between autonomy and beneficence. Rather, I have suggested here that our analysis of cases in which there appears to be such a conflict should be more nuanced. Contrary to the overly objectivist account that the commonly endorsed view delineated above seems to imply, I suggest that the fact that an agent has an autonomous preference for some outcome is a fact that is relevant to our assessment of what is in their best interests.

To illustrate, consider cases in which an agent autonomously refuses life-saving treatment; should we say that such cases must *always* represent conflicts between autonomy and beneficence? I do not believe that we should; to claim otherwise is to assume that continued existence should always take precedence over other goods when we consider what is in a patient's best interests. However, once we take the final value of autonomy seriously, if we can ascertain that the agent's refusal is autonomous, their refusal should be taken to indicate that the continued existence would not be a benefit to them, *all things considered*, in view of their own conception of the good life, and the

way in which they weigh the value of their continued existence against other goods.⁵³ In a more radical vein, it might even be proposed that in the absence of an objective ranking of values, we should reconceive of beneficence in bioethical contexts as pertaining to impersonal goods that are related to values that the medical profession is committed to promoting, such as health; on such a conception, we might understand conflicts between beneficence and autonomy to represent conflicts between these particular values and other goods that the patient values.

One might worry that the principle of beneficence becomes superfluous on this approach, since it seems to have been subsumed by the principle of autonomy.⁵⁴ As Buchanan points out, one response to this worry is to interpret the principle of beneficence in a purely negative sense, by understanding it as “ . . . an admonition to the physician not to allow the interests of others . . . to compromise his or her commitment to the patient”.⁵⁵ However, as Buchanan himself realise, this reading seems to give the principle too narrow an interpretation.⁵⁶

In my view, the principle of beneficence can still have substance on the understanding of well-being and autonomy that I have sketched in this chapter, because this understanding does not entail that the realisation of autonomous choices *exhausts* the concept of well-being. As I intimated above, the concept of ‘beneficence’ can

⁵³ Of course, a further matter that may complicate some of these cases is that there may be a conflict between respecting the agent’s locally autonomous decision, and facilitating her pursuit of diachronic plans central to her global autonomy by ensuring her continued existence. However, at least in some cases, the agent’s choice to refuse treatment may be motivated by a deep global commitment; I lack the space to adequately distinguish and address all such cases here.

⁵⁴ Buchanan raises this sort of concern about this approach. See Buchanan, “The Physician’s Knowledge and the Patient’s Best Interest,” 94.

⁵⁵ *Ibid.*, 95.

⁵⁶ *Ibid.*

incorporate both the patient's autonomous choices, and other goods that agents have impersonal self-interested reasons to want. On this understanding, conflicts will arise when the agent's autonomous choice is not co-extensive with what they have impersonal self-interested reason to want. To resolve the conflict, we must compare the strength of these impersonal self-interested reasons with the reasons we have to safeguard the agent's autonomy by respecting their own assessment of what is in their best interests.

That the principle of beneficence still has substance on this understanding becomes further apparent when we consider cases in which a patient has not and cannot make their choices clear to their physician. For example, considering what agents have impersonal self-interested reasons to want can give physicians guidance when they are dealing with incompetent patients who lack a surrogate decision maker or, as Buchanan points out, when asking for a competent patient's consent would be too time-consuming in an emergency situation, and in cases in which the patient delegates decision making authority to the physician. Moreover, Buchanan suggests that physicians need to have their own understanding of the patient's best interests when deciding which courses of treatment to provide as viable choices for a particular patient and when making a recommendation to patients.⁵⁷

Conclusion

I have argued that autonomy may often bear instrumental prudential value, and that it is prudentially valuable for its own sake. I have also claimed that our

⁵⁷ Ibid., 95–96.

understanding of what is in an agent's best interests should incorporate the agent's autonomous choices, and argued that this does not entail that autonomy cannot come into conflict with other values. It should be acknowledged that I have not attempted to assign a particular definite value to autonomy. I have simply suggested that autonomy represents a special type of value for us, in so far as we place value on living a life that is our own. Yet, I have left open the possibility that this value can have different weight in different people's conceptions of the good life. The extent to which autonomy will contribute to a person's welfare will depend in part on the extent to which autonomy conflicts with other outcomes that the agent has reasons to pursue.

Having considered the value of autonomy, I shall in the next chapter consider the prospect of enhancing autonomy, and the ethical issues that doing so through biotechnological means might raise.

Chapter Six - Enhancing Autonomy

In the previous chapter, I argued that autonomy is an important component of individual well-being. This conclusion naturally leads to the question of how we might seek to *enhance* autonomy. This theoretical question is also of practical interest, since advocates of human enhancement technologies have recently begun to consider ways that autonomy could be enhanced using emerging biotechnologies. In this chapter, I shall first explain how one might theoretically enhance an agent's autonomy on the theory that I have developed in this thesis. I shall then consider recent accounts from the bioethical literature of how biotechnology could be used to enhance autonomy, before arguing that these accounts overlook the importance of the relationship between the reflective and practical dimensions of autonomy, and as such fail to recognise an indirect way in which the widespread availability of enhancement technologies could enhance autonomy. I shall conclude by responding to some objections that have been raised against the claim that human enhancement technologies could be used to enhance autonomy.

I Enhancing Autonomy– Theoretical Considerations

In the first three chapters, I drew a distinction between two dimensions of autonomy. I identified a *reflective* dimension of autonomy that pertains to the reflection that agents must carry out on their motivating desires in order to be autonomous with respect to them. I also identified a *practical* dimension of autonomy pertaining to the

agent's ability to act effectively in pursuit of their ends. I identified the *minimum* threshold of autonomy on each dimension as follows:

Reflective: An agent is minimally autonomous with respect to her first order motivating desire if she has a preference to pursue the object of that desire which:

- i) . . . she sustains on the basis of a non-irrational belief that the object of that desire is good in a reason-implying sense.

And

- ii) . . . coheres with her character system.

Practical: An agent must have both the positive and negative freedom to act effectively in pursuit of the end that she is motivated to achieve.

When we consider how we might enhance an agent's autonomy, it is important to first establish whether the agent in question is already autonomous to the minimum threshold levels delineated above. If they are not, then removing impediments to their reaching this minimum threshold might appropriately be described as autonomy

enhancing. Notice here, that whether or not an agent qualifies as being autonomous is a discrete question. In contrast, if the agent is already above this minimum threshold, then their autonomy above this threshold can, I suggest, admit of degrees.

Since I have already addressed what it takes to reach a minimum threshold level of autonomy elsewhere in this thesis, in this section I shall consider how we might theoretically be able to enhance the autonomy of an agent who *already* reaches the minimum thresholds identified above.

i) *Enhancing the Practical Dimension*

Prima facie, the question of enhancing the practical dimension of autonomy seems quite straightforward. In chapter three, I claimed that an agent must be able to act effectively in pursuit of their ends in order to be practically autonomous to a minimum threshold level. In view of this, it seems plausible to advance the general claim that if we increase an agent's positive and negative freedoms so that they are able to act *more* effectively in pursuit of their ends, then this will often serve to enhance their practical autonomy. Furthermore, we might also observe that agents often come to change their preferences. As such, agents may come to require different sorts of freedoms over time in order to act effectively in pursuit of ends that they later decide they want to achieve. Accordingly, it might be claimed that having a diverse range of freedoms may promote the agent's practical autonomy, since having such freedoms accommodates for the possibility that agents may come to change their goals.

However, whilst it seems plausible to claim that increasing an agent's freedoms will often serve to enhance their practical autonomy, a number of qualifications to this claim need to be made, since giving an agent additional freedoms or options can in some cases hinder their pursuit of their goals. Most obviously, providing an agent with additional extraneous freedoms may involve replacing a freedom that they need to do what they want. For example, suppose that I have decided that I want to enjoy a particular brand of beer, but my local pub has stopped serving that brand in favour of serving fifty other beers which I do not like; here, increasing my freedom by increasing the number of beers that I can drink at this pub has failed to enhance my practical autonomy, since doing so has taken away my freedom to enjoy the beer that I actually want to have. Similarly, as I pointed out in chapter three, agents sometimes limit their own freedoms in order to enable them to effectively pursue certain goals.¹ Increasing a freedom that the agent herself has herself chosen to limit (in order to facilitate her pursuit of some goal) would then undermine, rather than enhance her practical autonomy.

Increasing an agent's general freedoms can also affect their freedom to pursue their preferred option without strictly making that option unavailable. As Dworkin points out, the addition of new freedoms may bring with it the cost of a new responsibility, such that the failure to choose the newly available option may now count against the chooser when previously it did not.² The thought is that agents may feel unable to pursue their preferred option because of this burden of responsibility. In an

¹ Chapter Three, 91.

² Dworkin, *The Theory and Practice of Autonomy*, 67.

autonomy-based objection to voluntary active euthanasia, David Velleman has argued that this sort of phenomenon might arise if voluntary active euthanasia were legalized.³

Increasing an agent's freedom to pursue one goal more effectively might also have a negative effect on their ability to pursue other goals. To illustrate, suppose that a woman valued having a successful career in business, and that she would be able to pursue this goal more effectively if she were more ruthless. In this case, whilst becoming more ruthless might enable the agent to pursue her career goal more effectively, it might also be detrimental to her pursuit of another goal that she values, like being a good parent for example. Accordingly, if increasing an agent's freedom to pursue some goal x is to enhance their practical autonomy, this enhanced freedom must either not diminish their efficacy with respect to their pursuit of another of their goals, y , or, if it would diminish their pursuit of y , then the agent must believe that they have a sufficiently strong reason to pursue x more effectively, at the cost to their efficacy in pursuing y that this might entail.

ii) *Enhancing the Reflective Dimension*

Although I shall note some exceptions to the following claim below, it seems plausible to claim that we can often increase our reflective autonomy with respect to our motivating desires by increasing the number of competing reasons that we consider in our practical deliberations. The thought underlying this claim is that it seems plausible to claim that the more alternatives we have at our disposal when we make a choice, the

³ Velleman, "Against the Right to Die."

more that our choice becomes a reflection of our own will, rather than of our restricted circumstances. Wolf makes a similar point in her theory of moral responsibility. She writes:

The more options and the more reasons for them that one is capable of seeing and understanding, the more fully one can claim one's choices to be one's own.⁴

To increase our autonomy above the minimum threshold level illustrated above then, we must not only sustain our motivating desire on the basis of a (non-irrational) belief that the outcome it concerns is good in a reason-implying sense; we must also consider the reasons we have to pursue alternative incompatible goals.

One way in which we could increase the number of alternatives that an agent considers in their practical deliberations is by increasing their cognitive capacities so that they are able to compute a greater number of the possible courses of action open to them. Another way in which we could do so becomes clear when we attend to the relationship between the practical and reflective dimensions of autonomy. Since agents form their desires in the light of what is practically realisable, and since considering more competing reasons will increase one's reflective autonomy, it seems that we might also be able to enhance an agent's autonomy simply by making more options practically realisable for them. That is, increasing an agent's freedoms can be understood to increase an agent's reflective autonomy in so far as it leads them consider more competing reasons in their deliberations.

⁴ Wolf, *Freedom within Reason*, 144.

Once again though, these claims require qualification, since increasing an agent's freedoms may not always serve to enhance their reflective autonomy, and might instead diminish it. First, if an agent does not have sufficient information, time or energy to make a choice amongst a vast number of options, then she may simply be overwhelmed by her available choices, and thus be unable to make an autonomous decision.⁵ Second, the additional choices that are made available will only serve to enhance an agent's autonomy if they are relevant to a choice domain of which the agent is part. For example, if a vegetarian is choosing between two different restaurants, the fact that one restaurant has more meat dishes than the other has no direct bearing on which restaurant will offer the vegetarian more autonomy with regards to her decision about what to order. If greater freedom is to enhance autonomy, it must make available choices that will enter into the agent's deliberation.

Finally, we might also follow Dworkin in observing that having certain choices may invite social pressure to conform in a manner that threatens the voluntariness of one's choice.⁶ To illustrate this, Dworkin provides the example of giving university students the option to live in mixed-sex dorms. Whilst it might be claimed that students who do not wish to live in mixed sex dorms could simply choose not to, Dworkin points out that to make this claim would be to "ignore the sociology of the situation";⁷ the claim fails to acknowledge the point that this new option introduces a social pressure on those who do not want to live in mixed dorms to conform to the social expectation of their peers. Accordingly, having the freedom to choose some alternative may undermine

⁵ Dworkin, *The Theory and Practice of Autonomy*, 66.

⁶ Ibid., 68.

⁷ Ibid., 69.

the voluntariness of one's choice, if having that option leaves one open to social pressure.

To take stock, I have claimed that increasing an agent's freedoms can serve to enhance their autonomy. Doing so can enhance the agent's practical autonomy in so far as it increases their effectiveness in pursuing their goals, and doing so can enhance their reflective autonomy in so far as it increases the number of competing alternatives that an agent considers in their practical deliberations. Despite this, I have also suggested that a number of qualifications need to be made to these general claims. The above discussion suggests two broad ways in which increasing an agent's freedom may *not* serve to enhance their autonomy. First, increasing an agent's freedoms can be counter-productive to their pursuit of certain goals. Second, increasing an agent's freedoms can leave them open to external forces that can affect the voluntariness of the agent's choice.

To conclude this section, we might also highlight an impediment to an agent's reaching the minimum threshold of autonomy that arises as a result of the relationship between the reflective and practical dimensions of autonomy. As I argued in chapter three, agents choose to sustain their motivating desires in the light of their beliefs about what is practically realisable for them. In view of this observation, I pointed out that adaptive preference formation can undermine autonomy. Adaptive preference formation occurs when an agent's awareness of their lack of alternative options leads them to unconsciously form preferences for available outcomes, in a manner that impedes them from critically reflecting on their motivating desires. Another way in which increasing an agent's freedoms may serve to enhance their autonomy is that it may serve as a guard

against adaptive preference formation, and the impediment to reflective autonomy that it represents.

In the next section I shall explain how some bioethicists have claimed that we might be able to use biotechnologies to enhance the reflective and practical dimensions of autonomy. I shall go on to argue in section three that by failing to recognise the relationship between these two dimensions of autonomy, they overlook the indirect effect that such technologies may have on our autonomy.

II Enhancing Autonomy With Biotechnologies

Scientific advances have led to the development of biotechnologies that could eventually be used to enhance human capabilities beyond that which is “ . . . necessary to sustain or restore good health”.⁸ Already, pharmacological agents such as Modafinil and Ritalin can be used to temporarily enhance cognitive abilities,⁹ whilst pharmacological anti-depressants might be viewed as allowing people to enhance their general mood.¹⁰ More speculatively, some theorists postulate the possibility of radical life-extension technologies,¹¹ genetic enhancements,¹² and even neuropharmacological

⁸ Juengst, “What Does Enhancement Mean?”, 29.

⁹ Sandberg, “Cognition Enhancement: Upgrading the Brain,” 73–75.

¹⁰ Berghmans et al., “Scientific, Ethical, and Social Issues in Mood Enhancement.”

¹¹ Barazzetti, “Looking for the Fountain of Youth: Scientific, Ethical, and Social Issues in the Extension of Human Lifespan.”

¹² Bostrom, “Human Genetic Enhancements.”

agents that could *morally* enhance people by reducing their aggressive tendencies and increasing empathy.¹³

There are three sorts of situation in which a person might undergo these sorts of enhancement. First, they might choose to undergo the enhancement *voluntarily*, when they themselves choose to do so without any overriding pressure from external parties. Second, and in direct contrast, they may undergo an enhancement *involuntarily*, that is, against their will. There are several ways in which an agent's decision to undergo an enhancement may be involuntary in this way. Most simply, they might be directly forced to undergo an enhancement against their will; alternatively, they might have chosen to go undergo an enhancement, but their choice may not have been voluntary because they had been coerced or manipulated into doing so. Finally, it might be argued that children who were prenatally enhanced would have been enhanced in a *non-voluntary* way, since they were enhanced prior to their capacity to choose to consent to (or refuse) the intervention.

Although the issue of non-voluntary enhancement has received attention elsewhere,¹⁴ I cannot adequately address it here.¹⁵ Rather, I shall assume that the agents who might undergo enhancements would have made a choice to undergo them, and that they are widely available to all.¹⁶ However, whilst I shall assume that the agents in question have not been directly forced to undergo an enhancement explicitly against

¹³ Douglas, "Moral Enhancement."

¹⁴ See Habermas, *The Future of Human Nature*; Davis, "The Parental Investment Factor and the Child's Right to an Open Future."

¹⁵ I consider these issues in Pugh, "Autonomy, Natality and Freedom."

¹⁶ Limitations of space mean that I must bypass an important objection to the use of enhancement technologies in making this stipulation, since a common objection to enhancement technologies is that they would create significant social inequalities between rich and poor.

their will, I leave open the possibility that the agent's choice to do so was involuntary in other ways. Thus, whilst I assume that the agents in question chose to undergo their enhancements, for now, I leave it an open question as to whether this choice was voluntary or involuntary; this will obviously have important implications for whether the enhancements can be understood as enhancing autonomy.

The primary concern in most discussions of autonomy in the context of human enhancement technologies has been with what I have called the *practical* dimension of autonomy. Niklas Juth, Neil Levy and Nick Bostrom have all suggested that biotechnologies could be used to enhance autonomy in so far as they could provide agents with better means to *realise* their own ends.¹⁷ For instance, consider the following passage from Juth:

Can enhancement technologies promote individuals' autonomy? The answer is *yes*. In general, plans require capacities in order for them to be put into effect and enhancement technologies can increase our capacities to do the things we need in order to effectuate our plans.¹⁸

To illustrate this point, suppose that John is an averagely intelligent man who has a strong desire to understand complex mathematics. Such intellectual feats are beyond the scope of most of us given our natural capacities; however, it may one day be possible for John to effectively pursue this plan by virtue of the capacities granted to

¹⁷ Levy, "Enhancing Authenticity"; Juth, "Enhancement, Autonomy, and Authenticity"; Bostrom, "In Defense of Posthuman Dignity."

¹⁸ Juth, "Enhancement, Autonomy, and Authenticity," 36.

him by a cognitive enhancement. Alternatively, we might imagine a shy person taking an enhancement that would allow her to pursue her dream to become an actress more effectively, and so on.

However, it should be acknowledged that enhancement technologies are unlikely to serve as a panacea for the effective pursuit of all of our goals. As I explained in my discussion of process goods and outcome goods in the previous chapter on the value of autonomy, the value that we attribute to certain goals is tied up with the way in which we achieve those goals. For example, in chapter five I suggested that we might describe the goal of playing a complex piece of piano music in a crude sense as simply having the goal of playing certain keys in a certain order. However, we might describe the goal in a more sophisticated sense as having the goal of exercising our own creativity in playing the piece.¹⁹

Accordingly, it is important that we understand all the facets of our goals when we consider whether undergoing an enhancement would be conducive to their achievement. To illustrate, if our goal in the above example was merely to play the right notes in the right order, then being able to download the ability to do this via a mind-machine interface would certainly increase our efficacy in pursuing that goal. However, if we wanted to play the piece in the more sophisticated sense that I discussed in chapter

¹⁹ Chapter Five, 147.

five, this particular enhancement²⁰ would not increase our ability to pursue our goal, and might instead be inimical to it.²¹

Bioethicists have only recently begun to consider whether biotechnologies could be used to enhance what I have termed the *reflective* dimension of autonomy. Schaefer *et al.* have recently argued that we might be able to enhance an agent's autonomy by enhancing their general reasoning abilities.²² In their analysis, Schaefer *et al.* do not align themselves with a particular position in the autonomy debate, but employ a two-pronged strategy in arguing that on any plausible understanding, autonomy can be enhanced through the use of biotechnologies. In their first, theory-based approach, they identify broader reasoning ability (including logical competence, comprehension and critical analysis) to be a common feature across various diverse accounts of autonomy.²³ They then argue that we would be able to increase agents' autonomy by enhancing this broader reasoning ability through the use of cognitive enhancements.

In their second, case-based approach, Schaefer *et al.* point out that

²⁰ Other enhancements might yet be conducive to the achievement of this goal; for example, a physical enhancement that enhanced one's co-ordination might allow one to exercise one's creativity more effectively.

²¹ I believe that these considerations can be used to make one of Leon Kass' objections to human enhancement more plausible. Kass objects to enhancements on the basis that we would be unable to comprehend the meaning of our achievements in human terms following an enhancement. See Kass, "Ageless Bodies, Happy Souls." Philosophers have been critical of this objection (for example, see Juth, "Enhancement, Autonomy, and Authenticity," 43). However, the objection becomes more plausible if instead of using the language of an achievement being *comprehensible* in human terms, we make the more plausible claim that the value of some goals is tied up with the way in which we go about achieving them.

²² Schaefer, Kahane, and Savulescu, "Autonomy and Enhancement."

²³ *Ibid.*, 126.

(m)uch of the debate over autonomy is centered on making sense of a few paradigmatic cases of autonomy inhibition or violation.²⁴

They cite brain-washing, psychological manipulation, and deception as examples of interventions that are generally understood to undermine autonomy. They then argue that these interventions all operate “. . . by affecting individuals’ internal psychology and ability to reason properly”.²⁵ Accordingly, they go on to argue that:

. . . preventing such inhibitions by improving people’s reasoning and deliberation should have the effect of enhancing autonomy, insofar as there would be fewer instances when one’s autonomy will be inhibited.²⁶

It should be noted here that in their theory-based approach, Schaefer *et al.* seem to be arguing that cognitive enhancements can be used to enhance autonomy above a minimum threshold, whilst in their case-based approach, they argue that cognitive enhancements can be used to enhance autonomy in so far as they remove impediments to achieving the minimum threshold of autonomy.

I broadly agree with the arguments that Schaefer *et al.* present, and endorse their conclusion that cognitive enhancements can be used to enhance reflective autonomy. However, there are two important points that the authors do not recognise in their arguments. The first is that they fail to consider the way in which enhancement

²⁴ Ibid., 127.

²⁵ Ibid.

²⁶ Ibid.

technologies could be used to overcome *internal* impediments to autonomy; the second is that they do not consider the effect that enhancement technologies could have on the agent's *practical* freedoms, and what impact this might have on autonomy.

i) *Internal Impediments to Autonomy*²⁷

In their case-based approach, the interventions that Schaefer *et al.* highlight (brain-washing, psychological manipulation, and deception) all involve an agent's autonomy being undermined *by another agent*. Whilst the interventions that Schaefer *et al.* mention undoubtedly can undermine autonomy, it seems that an agent's autonomy can also be inhibited *without* the direct intervention of another agent. One salient example of such a case that Schaefer *et al.* do not consider is an agent who acts only the basis of their impulsive desires. As I discussed in the introduction and in chapter one, we do not believe that an agent who acts on an impulsive desire is autonomous with respect to their act, because such an agent has failed to adequately reflect on that desire.

An agent who acts on the basis of an impulsive desire might do so simply because they lack the capacity for higher-order reflection. However, it also seems possible that an agent may act on the basis of an impulsive desire, not because they lack the *capacity* to reflect on their desires in the manner that is necessary for autonomy, but rather because they are unable to *exercise* the relevant capacity, even though they may have it; such an agent might feel so overwhelmed by their impulses that they act before they are able to carry out the required reflection on their motivating desires, reflection

²⁷ This section is adapted from Pugh, "Enhancing Autonomy by Reducing Impulsivity."

which they are nonetheless capable of. It seems that in order to enhance this sort of person's autonomy, we would not need to enhance their capacity for critical reflection; rather, we would need to remove impediments to their exercising that capacity.

Interestingly, it seems that we already make a pharmacological intervention that has this sort of effect when we use Methylphenidate (more commonly known as Ritalin) to treat ADHD. ADHD is typified by three symptomatic behavioural problems; inattentiveness, hyper-activity and morbid impulsivity.²⁸ One beneficial effect of taking Ritalin to combat these symptoms is suggested by recent qualitative studies by Singh,²⁹ which indicate that sufferers of ADHD who take stimulant medication “tend to feel that they have increased agency . . . in forging their life trajectories”.³⁰ It seems that one plausible explanation of this is that insofar as Ritalin suppresses their impulsive tendencies, it allows sufferers of ADHD to engage in the sort of critical reflection on their desires that is necessary for autonomy.³¹

In a similar vein, Sandberg notes that Modafinil seems to enhance adaptive response inhibition, which makes subjects taking the drug “ . . . evaluate a problem more thoroughly before responding”.³² As such, both Ritalin and Modafinil can be understood as removing impediments that some agents may face to reflecting on their

²⁸ Singh, “Beyond Polemics.”

²⁹ Singh, “Not Just Naughty: 50 Years of Stimulant Drug Advertising.”

³⁰ Singh and Kelleher, “Neuroenhancement in Young People,” 8.

³¹ In the interests of simplicity, I am assuming in my arguments that those suffering from ADHD have already developed the critical capacities that autonomy requires. However, it should be acknowledged that many young children who are treated for ADHD arguably lack these capacities. In these cases, my arguments would need to be slightly amended, since in these cases impulsivity cannot be understood as an impediment to the exercise of an existing capacity. Accordingly, in these cases, we should understand impulsivity as an impediment to the child's *future* autonomy.

³² Sandberg, “Cognition Enhancement: Upgrading the Brain,” 74.

desires prior to their moving them to act; and these impediments may plausibly be regarded as impediments to their autonomy. Furthermore, studies by Kreek et al. have suggested that destructively impulsive behaviour might also have a partly genetic basis;³³ this in turn suggests that autonomy-undermining degrees of impulsivity might also be affected by germline interventions.

Of course, those who use Ritalin in the treatment of ADHD may not regard themselves as making their patients more autonomous, but as simply restoring ‘normal’ mental functioning; in view of this, it might be argued that whilst it is permissible to use Ritalin as a treatment for a medical condition, it would not be permissible to use it to enhance a healthy individual’s autonomy. However, the sort of impulsivity that I have discussed here is not a problem faced by sufferers of ADHD alone. It seems that some ‘healthy’ individuals may not be impulsive to the pathological degree of those who suffer from ADHD, and yet feel precluded from reflecting on their motivating desires because they always act impulsively before they are able to engage in critical reflection. For example, it has recently been claimed that some obese persons eat impulsively before they are able to critically assess their reasons to refrain from doing so.³⁴

This continuity between the impulsivity of those suffering from ADHD and the impulsivity of certain healthy individuals calls into question the ethical relevance of the distinction often invoked between treatments and enhancements.³⁵ In my view, in so far as cognitive enhancements could be used to reduce impulsivity, it seems plausible to suggest that they could increase the autonomy of individuals whose autonomy is

³³ Kreek et al., “Genetic Influences on Impulsivity, Risk Taking, Stress Responsivity and Vulnerability to Drug Abuse and Addiction.”

³⁴ Mobbs et al., “Obesity and the Four Facets of Impulsivity.”

³⁵ For a discussion of the relevance of this distinction, see Savulescu, Sandberg, and Kahane, “Well-Being and Enhancement.”

impaired in this way; in the case of ADHD patients, I would claim that they already do, regardless of whether or not these interventions are intended to restore ‘normal’ mental functioning.

ii) *The Importance of the Practical Dimension of Autonomy*

It is also striking to note that Schaefer *et al.* do not consider the effect that enhancement technologies could have on the agent’s *practical* freedoms, and what impact this might have on autonomy. Indeed, they seem to reject the view that what I call the practical dimension of autonomy should be incorporated into an overall theory of autonomy. For instance, they suggest that a problematic aspect of Levy and Juth’s arguments regarding the enhancement of autonomy is that:

Juth and Levy both endorse efficacy (ability to act on and carry out one’s goals and desires) as a necessary condition of autonomy, but this condition (which has a notably external character) is not found in any of the mainstream conceptions of autonomy.³⁶

Whilst it is true that comparatively few theories of autonomy acknowledge the importance of an agent’s practical freedoms to their autonomy, I have argued in this thesis that we still have good reason to do so.

³⁶ Schaefer, Kahane, and Savulescu, “Autonomy and Enhancement,” 125.

It might be argued that in order to get a complete picture of the prospects of enhancing autonomy through the use of biotechnology, we need only consider the conclusions of Schaefer *et al.* regarding the enhancement of the reflective dimension of autonomy in conjunction with those of Juth, Bostrom and Levy regarding the enhancement of the practical dimension of autonomy. However, this would be to fail to appreciate the importance of the relationship between the two dimensions of autonomy that I have highlighted in this thesis. I shall now suggest that if we attend to the nature of this relationship, then it becomes clear that biotechnological enhancements could have other positive impacts on autonomy. However, in view of my theoretical discussion in section one, there are a number of limitations here; biotechnological enhancement might also have some potentially negative impacts on autonomy as well, some of which the above theorists overlook.

III Indirectly Enhancing Autonomy?

As I explained above, Schaefer *et al.* suggest that we could enhance an agent's autonomy by increasing their general reasoning abilities through the use of cognitive enhancements. They also acknowledge that the ability to recognise a wide range of options is important on many theories of autonomy.³⁷ Accordingly, they claim that cognitive enhancements that make an agent capable of computing more options in their

³⁷ *Ibid.*, 126.

practical deliberations would often serve to enhance autonomy;³⁸ this claim is congruous with my theoretical analysis in section I of this chapter.

We may note that the enhancements that Schaefer *et al.* argue could be used to increase the number of alternatives that an agent is able to compute all enhance this ability in a *direct* manner; that is, they increase the number of alternatives that an agent is able to compute by affecting the agent's own computational capacities. However, in failing to acknowledge the practical dimension of autonomy, Schaefer *et al.* fail to consider a further way in which enhancement technologies might *indirectly* increase the number of alternatives that agents consider in their practical deliberations. I shall now suggest that the *availability* of such means of enhancement might also have an indirect enhancing effect on the reflective dimension of autonomy, in so far as increasing the number of alternatives that an agent considers in their practical deliberations can serve to enhance their reflective autonomy.

First, the widespread availability of enhancement technologies could serve as a guard against adaptive preference formation. It would do so if the availability of such technologies made it less likely that agents would justify their adoption of long term plans on the basis that they felt 'forced' into a certain sort of life by their limited natural capacities. For instance, if cognitive enhancements were available to all, then agents could no longer justify rejecting a life plan whose fulfilment required high intelligence on the sole basis that it was not practically realisable for them given their natural capacities; cognitive enhancements might bring such a life plan into the realms of possibility. To extend the point more broadly, the availability of enhancement technologies might prompt individuals who would otherwise feel constrained into

³⁸ Ibid.

certain life plans by their limited capacities, to consider the *content* of their plans, rather than resigning themselves to that plan, and blaming their natural capacities for forcing them to live the life that they do.

Second, it seems that making a greater number of alternatives practically realisable could lead agents to reflect on their long term plans to a greater extent. I suggested above that increasing the number of alternative courses of action that one considers in one's deliberation can serve to increase one's reflective autonomy. Accordingly, if enhancement technologies could make a greater number of alternatives practically realisable, they may increase the number of alternatives that agents include in their practical deliberations, and thus increase their autonomy with respect to the motivating desires that they choose to sustain.

To some extent, it seems that making enhancement technologies widely available could have an enhancing effect on individual autonomy *regardless* of whether the particular agent in question has undergone an enhancement. The reason that making enhancement technologies generally available would *indirectly* increase autonomy is because it would force us all to consider the reasons for living the lives that we do. It would do so, because the very possibility of enhancement might make alternatives that were previously closed to us, practically realisable.

I have outlined a number of ways in which biotechnologies could be used to potentially enhance human capacities in a manner that can be understood as serving to enhance autonomy. At this point though, it is crucial to recall the qualifications I made in section I to the claim that increasing freedoms will generally serve to enhance autonomy. It seems plausible to claim that the additional choices that enhancement

technologies might give us could also feasibly diminish our autonomy in some of the ways that I described in section I.

IV Counter-productive Enhancements, and the Voluntariness of Choice

The discussion of section I suggests two broad ways in which increasing an agent's freedom may not serve to enhance their autonomy.³⁹ First, increasing an agent's freedoms can be counter-productive to their pursuit of certain goals. Second, increasing an agent's freedoms can leave them vulnerable to external forces that can affect the voluntariness of the agent's choice. I shall consider each in turn.

i) Counter-productive Enhancements

Many of the worries about increasing freedom that I considered in section I will usually not be applicable to the use of enhancements that the agent herself has chosen to undergo. For example, an informed agent would presumably not choose to replace a freedom that is necessary for the effective pursuit of her goals with other enhanced

³⁹ Some bioethicists have argued that certain enhancements would threaten the enhanced agent's authenticity. For example, see Elliott, *Better than Well*. I shall not consider this objection here for two reasons. First, the objection tends to rely on an unconvincing essentialist view of the self that the sense of authenticity that I invoke in my theory of autonomy does not rely on. Second, this objection has already received considerable attention in the literature. For a convincing rebuttal of an authenticity-based objection to enhancement technologies, see Levy, "Enhancing Authenticity."

capacities that are not. Furthermore, if the agent herself chooses to undergo an enhancement, it seems highly unlikely that she will choose to increase a freedom that she has previously chosen to limit.

However, it seems that there are some plausible scenarios in which a voluntary enhancement might be counter-productive to the enhanced agent's pursuit of certain goals. First of all, the enhancement of some capacities might have unwanted side-effects. For example, studies have suggested there is a correlation between the genetic enhancement of memory capacity in mice and susceptibility to certain sorts of pain.⁴⁰ Moreover, even if a particular enhancement works as expected and facilitates the agent's pursuit of a goal, it may undermine her ability to effectively pursue another of her goals; for example, we might imagine the businesswoman from my above example being able to take a pill to make herself more ruthless, thus precluding her pursuit of being a good mother.

However, these worries are not particularly problematic. The first worry merely suggests that enhancement technologies should be subjected to rigorous safety tests, and that subjects should be made aware of the possible side-effects of their enhancements. Similarly, the fact that an enhancement might increase an agent's effective pursuit of one goal but diminish her pursuit of another is only problematic if the agent herself has not voluntarily decided to undergo the enhancement in the knowledge that it will have this effect. For example, the business woman might be aware that being more ruthless might make her a worse mother and yet still decide that the effect that being more ruthless will have on her career gives her sufficient reason to take it.

⁴⁰ Wei et al., "Genetic Enhancement of Inflammatory Pain by Forebrain NR2B Overexpression."

However, other objections seem more powerful. For example, Michael Sandel echoes Dworkin's claim that having more choices can lead to a greater degree of unwelcome responsibility, when he suggests that the widespread availability of enhancement technologies could lead to an "explosion of responsibility".⁴¹ Sandel writes:

One of the blessings of seeing ourselves as creatures of nature, God, or fortune is that we are not wholly responsible for the way we are. The more we become masters of our genetic endowments, the greater the burden we bear for the talents we have and the way we perform.⁴²

Again though, it doesn't seem that this is a knock-down objection to making enhancement technologies widely available. First, it is illuminating to compare Sandel's concerns here with Sartre's existentialist views. Sartre claimed that those who believe that their own given and unalterable natures force them to follow a particular life-plan are living inauthentically in 'bad faith'. Living in 'bad faith' is, for Sartre, to deny the metaphysical truth of one's own freedom to transcend one's circumstances as a being-for-itself.⁴³

I am not suggesting that we should accept Sartre's dubious metaphysical claims here. However, one way of understanding the indirect impact of the availability of enhancement technologies on autonomy is one of prompting a Sartrean re-awakening in

⁴¹ Sandel, *The Case against Perfection*, 87.

⁴² Ibid.

⁴³ Sartre, *Being and Nothingness*.

response to the conservatism of the sort that Sandel espouses. The prospect of widely available enhancement technologies makes the possibility of overcoming elements of our given natures far more plausible to us than ever before. In coming to realise the possibility of overcoming certain elements of our given natures, we may come to view ourselves as beings that are able to define ourselves, rather than viewing ourselves as beings that are defined by their circumstances. Accordingly, the expansion of responsibility that enhancement technologies promise is perhaps something that we should, to some extent at least, welcome from the point of view of autonomy.

I shall also argue below that Sandel's objection seems to rely on the status quo bias. Prior to doing so, we should acknowledge that a related source of objection to making enhancement technologies widely available is that people might be overwhelmed by the degree of choice that such technologies would give us. If made widely available, enhancement technologies have the potential to increase our available options; accordingly, we might worry that we may lack the capacities to deal with such an influx of new possibilities.

The first thing to say in response to this sort of objection is that certain enhancements could be used to directly mitigate the negative effects on autonomy that having more choices could lead to. For example, it seems that certain cognitive enhancements could be used not only to increase the computational capacities of humans that would allow them to deal adequately with a large choice set, but also to improve their ability to calculate when deliberation on information is appropriate, and when it should be forgone.⁴⁴

⁴⁴ Schaefer, Kahane, and Savulescu, "Autonomy and Enhancement," 128.

Whilst Schaefer *et al.* do not directly consider the impact that having more choices can have on people's autonomy, they do recognise that a possible objection to their arguments is that increased reasoning capacity might be detrimental to autonomy because it could lead to decisional paralysis due to over-rationalization.⁴⁵ In response to this argument, Schaefer *et al.* point out that even if enhancing an agent's reasoning abilities to a very high level might undermine autonomy, it is both likely that there is an 'ideal mean' of that capacity, and that people are generally below this ideal mean. It seems that a similar argument can be made with regards to the number of choices that agents have available to them; it is likely that there is some ideal mean of available options, and that many people may not currently have an adequate range of options.

We can also invoke Bostrom and Ord's reversal test for status quo bias in support of this conclusion, and against Sandel's argument regarding the potential 'explosion of responsibility';⁴⁶ if one claims that the opportunities and freedoms that enhancements might make available would be bad for us from the point of view of autonomy, then it seems that we must then either be prepared to support a reduction of our current freedoms and opportunities, or to give an account of why we are currently at a local optimum with regards to our current freedoms and corresponding responsibilities. Since it seems safe to assume that most of us would not defend a reduction of our current freedoms, in the absence of a convincing explanation of why we are currently at a local optimum, then this sort of opposition to increasing our freedoms through the use of enhancement technologies smacks of the status quo bias.

However, it might be argued that a convincing explanation of why we are currently at a local optimum is that increasing our freedoms further might undermine

⁴⁵ Ibid.

⁴⁶ Bostrom and Ord, "The Reversal Test."

our autonomy by undermining the voluntariness of our choices. I shall conclude by considering this issue.

ii) *Voluntariness of Choice*

It seems that the most salient way in which the widespread availability of enhancement technologies could undermine the voluntariness of our choices is by leading to strong social pressures to conform. When exposed to such pressures, rather than making our choices on the basis of reason-giving facts, we might feel compelled to do something that we feel we have a reason not to do, simply because other people are doing it. Juth uses the example of memory-enhancements to illustrate this possibility; he suggests that even if an agent did not originally plan to use such technology, they might nonetheless feel compelled to use it if it became available in order to maintain their social opportunities.⁴⁷

Juth responds to the problem by pointing out that not all social pressures are bad from the point of view of autonomy; for example, he suggests that the social pressure not to resort to violence might well have positive overall consequences for what I have called our practical autonomy.⁴⁸ He then suggests that the enhancements that are “most likely”⁴⁹ to reduce overall autonomy are those that primarily confer a competitive advantage; in contrast, he suggests that enhancements that are not primarily intended to confer a competitive advantage will often have the potential to enhance autonomy for

⁴⁷ Juth, “Enhancement, Autonomy, and Authenticity,” 44.

⁴⁸ Ibid.

⁴⁹ Ibid.

the individual, regardless of the possible competitive advantages they happen to confer, in so far as they can increase the agent's ability to effectuate their plans.⁵⁰

However, Juth's response to this problem is not sufficient for dispelling the concern that the objection raises; even if we agree with Juth's theoretical claims, his arguments seem to have very limited scope. Whilst Juth may be right to claim that not all social pressures are bad from the point of view of autonomy, the likely reply to this response is that many social pressures *are* bad from the point of view of autonomy, and that the availability of enhancement technologies opens the door to at least some of them. For example, one might speculate that the availability of mood enhancements might have the effect of reinforcing the social norm that people ought to have a cheery disposition; it is not implausible to suppose that many people would not wish to change themselves into a cheerier person (perhaps they find constantly cheery people annoying), but feel that their social and economic opportunities will diminish if they maintain their current downbeat personality.

In a similar vein, whilst Juth might be right to claim that enhancements that primarily confer a competitive advantage are the ones that are *most likely* to reduce overall autonomy if there is a strong social pressure to use them, this simply means that some enhancements are more likely to undermine autonomy than others; this hardly does much to answer the objector's complaint that enhancement technologies threaten autonomy. Moreover, it seems that the enhancements that Juth claims do not primarily confer a competitive advantage might still be likely to undermine autonomy, all things considered. For instance, it seems that mood enhancements have the potential to enhance autonomy for some individuals who want to use them, even if they are not

⁵⁰ Ibid.

understood to primarily provide competitive advantages. However, this does not answer the objection that they do confer competitive advantages in societies that favour happier people, and this may lead to a social pressure to conform; and it is this that threatens the autonomy of those individuals who otherwise believe that they have reasons *not* to undergo the enhancement.

Contra Juth, rather than playing down the extent to which the availability of enhancements might lead to social pressures to conform, we should instead concede that the availability of such technologies would be likely to bring with it a variety of new social pressures. However, we should note that this is not solely a problem with enhancement technologies; it is a problem with any technological advance that radically transforms our freedoms. For example, consider the advent of online social media; whilst this has increased our ability to communicate with others, it has also introduced an array of new and powerful social pressures to conform to pre-existing norms. The point that this example raises is that even if we concede that a technology introduces social pressures to conform in a manner that could pose a threat to the voluntariness of some of our choices, we need not necessarily conclude that this technology must therefore have undermined our autonomy. The net benefits to autonomy of the technology might be sufficient to outweigh the net costs to autonomy; in view of its widespread use, it seems plausible that many of us hold this sort of attitude towards social media and the Internet more generally. In a similar vein, we might claim that the benefits to autonomy that the widespread availability of enhancement promises are sufficient to outweigh the potential threat to autonomy posed by the social pressures that they might introduce.

A further point that speaks in favour of this view as, Schaefer *et al.* point out, is that cognitive enhancements could be used to increase an agent's general reasoning capacities in a manner that could serve to increase their resistance to dogmatically conforming to social norms.⁵¹ Furthermore, we might seek to mitigate the effect of social pressures to conform by making available only those enhancements that are likely to be “. . . useful and valuable in carrying out nearly any plan of life or set of aims that humans typically have”;⁵² such enhancements have been described as General Purpose Means.⁵³ Examples of putative General Purpose Means might include good memory, epistemic rationality, and the ability to socially interact with others.

The benefit of only making available those technologies that enhance General Purpose Means is that almost all agents would have reasons to choose to undergo such enhancements; after all, they are by definition conducive to the effective pursuit of many different goals that agents might have. Accordingly, it seems that an agent would be likely to choose to undergo the enhancement of a General Purpose Means themselves, regardless of any social pressure to conform to a specific social norm that the availability of the enhancement might give rise to; furthermore, in so far as the enhancement is conducive to the pursuit of many different goals, the use of the enhancement need not involve the tacit endorsement of a particular specific set of values that could be understood as creating a specific social norm.

⁵¹ Schaefer, Kahane, and Savulescu, “Autonomy and Enhancement,” 129.

⁵² Buchanan et al., *From Chance to Choice*, 167.

⁵³ Ibid.

Conclusion

In my theoretical analysis in the first section of this chapter, I suggested a number of ways in which increasing an agent's freedoms can have detrimental effects on the agent's autonomy all things considered. These theoretical considerations have important implications for the biotechnological enhancement of autonomy; whilst enhancement technologies may allow us to significantly increase our freedoms, they will not enhance our autonomy if we fail to take heed of the effect that our beliefs concerning our freedoms have on the voluntariness of our choices, and the new social pressures that such technologies may introduce.

Yet, this observation should not make us throw the baby out with the bath water when considering the way in which enhancement technologies might affect our autonomy; we should not forget the intuitive point to which I have highlighted exceptions, namely, that increasing an agent's freedom will often serve to enhance autonomy. Moreover, many of the problems with increasing agents' freedoms that I have highlighted here are avoidable; for instance, the enhancement of General Purpose Means and resistance to conformity are examples of enhancements that do not seem to leave individuals vulnerable to strong social pressures to conform. In conjunction with my earlier defence of the view that autonomy has final and instrumental prudential value, I suggest that the use of biotechnologies to enhance autonomy is amongst the most defensible goals of these burgeoning and controversial technologies.

Chapter Seven - Informed Consent and Autonomy Part

One: Voluntariness

Informed consent requirements are ubiquitous in health care, and they are regarded as a cornerstone of ethical medical practice. Until recently, it was generally agreed that informed consent requirements were to be justified in large part (although perhaps not entirely) by appeal to the principle of respect for autonomy. However, whilst this view is still widely accepted, it has recently been brought into question, with some philosophers claiming that informed consent requirements are insufficient for safeguarding patient autonomy.

Over the course of the following two chapters, I shall argue that this objection is not convincing, though I shall claim that it suggests that we must revise our understanding of what informed consent requires. I shall begin this chapter by introducing the concept of informed consent and the debate surrounding its justification. In section two, I shall consider the ramifications that my theory of autonomy has for what we might term the ‘voluntariness elements’ of informed consent, and argue that we ought to revise the commonly endorsed account of informed consent that is based on the standard view of autonomy in bioethics, identified in the introduction of this thesis. In the third section, I shall argue that the revised view that I suggest supports a move away from the commonly endorsed shared decision making model of the doctor-patient relationship, towards a model that allows physicians to go beyond the role of a mere ‘fact-provider’, and to engage with their patients in rational discussion about their evaluative judgements. In the next chapter, I shall go on to consider the ramifications

that my theory has for what we might term the information elements of informed consent, and for the general concept of competence to consent.

I Introducing Informed Consent and its Justification

It seems that the fundamental idea that we aim to capture when we claim that ‘*A* morally ought to obtain *B*’s informed consent to *A*’s doing *x* to *B*’, is that the moral permissibility of *A*’s doing *x* to *B*, is at least partly¹ dependent on the following conditions being met:

- i) *B* must be sufficiently informed with regards to the relevant facts concerning *x* to understand what *x* is (and what consequences are likely to occur as a result of *x*).

- ii) On the basis of this information, *B herself* make the decision to allow *A* to do *x*.

The first condition highlights what we might term the ‘information element’ of informed consent, which pertains to the patient’s informational condition; the second condition highlights what we might term the ‘voluntariness element’ of informed consent, which pertains to the voluntariness of the patient’s decision. My focus in this chapter shall be the voluntariness element.

¹ Note that there may be other moral reasons for *A* to refrain from doing *x* to *B* even if *B* consents.

As I explained in the introduction, on the standard view of autonomy in contemporary bioethics, an agent is autonomous with respect to a particular act if it is carried out:

(1) Intentionally,

(2) With understanding,

And

(3) Without controlling influences that determine their action.²

Faden and Beauchamp suggest that this understanding can be used to undergird a theory of informed consent understood as a form of *autonomous authorisation*.³ On this view, to give informed consent is to perform a specific kind of autonomous action, one that “ . . . authorises a professional to initiate a medical plan for the patient”.⁴ On Faden and Beauchamp’s view, a patient gives informed consent if they provide such an authorisation whilst meeting the conditions of autonomous action outlined above.⁵

² Faden and Beauchamp, *A History and Theory of Informed Consent*, 238.

³ Beauchamp and Childress' *Principles of Biomedical Ethics* defends a similar view. I consider Faden and Beauchamp's *A History and Theory of Informed Consent* in this chapter rather than Beauchamp and Childress' view for two reasons. First, Faden and Beauchamp's work is solely on the nature of informed consent, and so represents a more focussed discussion of the concept. Second, the views on informed consent that Beauchamp and Childress have espoused in *The Principles of Biomedical Ethics* has undergone significant revisions over the numerous editions of the book. However, as Walker acknowledges, Faden and Beauchamp's account is very similar to the view that is apparent in editions of *The Principles of Biomedical Ethics* that followed it. Walker, “Respect for Rational Autonomy,” ft 3.

⁴ Faden and Beauchamp, *A History and Theory of Informed Consent*, 278.

⁵ Ibid.

Conditions (1) and (3) of Faden and Beauchamp's theory may be understood as pertaining to what I have called the voluntariness element of informed consent, whilst condition (2) may be understood as pertaining to the information elements of informed consent. As I explained in the introduction, the standard view of autonomy that undergirds this theory of informed consent endorses a thinner account of autonomy than that which I have defended in this thesis; I shall explain the reasons for this below.

Prior to beginning my consideration of the voluntariness element of informed consent, I shall conclude this section by considering the justification of informed consent requirements. As Dworkin points out, the doctrine of informed consent is a "creature of law";⁶ it has been developed in various legal domains in which one party sanctions another to perform " . . . some course of action to which the consented to party would otherwise have no moral right."⁷ Here, I am interested only in the philosophical basis of informed consent as it is invoked in the context of medical practice (when that is understood to refer only to clinical care and *not* non-therapeutic research). Whilst I shall not consider the numerous other domains in which the concept of informed consent has been invoked, I shall make some brief comments here about the justification of informed consent requirements in the context of non-therapeutic medical research, since contrasting this justification with the justification of informed consent requirements in medical practice serves to illuminate an important aspect of the latter.

It seems that a primary justification of informed consent requirements in non-therapeutic medical research is that they protect agents from being put at risk of harm

⁶ Dworkin, *The Theory and Practice of Autonomy*, 101.

⁷ Kleinig, "The Nature of Consent," 8. See also Miller and Wertheimer, *The Ethics of Consent*, for examples of the domains in which informed consent can be invoked.

(broadly construed)⁸ against their will, even if others have an interest in their participating in the research. One reason that this sort of concern is particularly salient in the context of non-therapeutic medical research is that the benefits of the intervention for which consent is being solicited here will often not accrue directly⁹ to the subject themselves. It seems that this partly explains the intuitive pull of the idea that informed consent requirements in this context are to be justified by an appeal to the principle of respect for autonomy; informed consent requirements can be understood as safeguarding autonomy in so far as they help to ensure that nobody will participate in research against their own will, even if substantial benefits to others could be accrued by carrying out the research. It is left to the participant herself to decide whether she wants to risk the harm that the research might entail.

It seems that informed consent requirements in medical practice partly play a similar role, in so far as they serve to protect patients from being forced into involuntarily receiving treatments that might serve another party's interests rather than their own. For instance, if informed consent requirements dictate that a physician must inform their patient about the likely efficacy of a proposed treatment, it will presumably be more difficult for a physician to get their patient to agree to undergo a novel treatment that the physician might have self-interested reasons to promote, but which might be less efficacious than an alternative.

⁸ I use the concept of harm broadly here to incorporate violations of individual rights that might not include 'harming' someone on a common understanding of that concept. For instance, whilst someone's giving me a pedicure without my consent does not harm me in one sense, we can still understand the act as harmful in so far as it involved a violation of a right to bodily integrity that I did not voluntarily waive.

⁹ Of course, the participants may be *indirectly* benefited by the intervention through payment.

However, it seems that informed consent requirements also play a further role in medical practice that is not applicable in the context of non-therapeutic medical research, in so far as the interventions for which consent is being solicited in the former context are normally intended to *directly* benefit the patient herself. In the context of medical practice, patients are not consenting to being put at a risk of harm in order to benefit others, (or to indirectly benefit by way of receiving financial compensation); rather, they are consenting to an intervention that is intended to bring about some beneficial physiological or psychological change *in them*. In view of this, and in view of my discussion of beneficence and autonomy at the end of chapter five,¹⁰ it seems that one purpose of obtaining an informed consent in medical practice is that it allows the physician and patient to reach an understanding of what sorts of intervention will most benefit the patient, all things considered.

The importance of this is made clear once it is observed that patients can differ from their physicians in their conclusions about what they have strongest self-interested reason to do, even in view of the same relevant descriptive facts. To illustrate this, consider this example from Savulescu:

Suppose that Joe is undergoing an operation to remove a tumour from his diaphragm. An anaesthetist consults him regarding his post-operative analgesia. The efficiency of analgesia in Joe's case is crucial, since if the analgesia is not effective, it is likely that Joe's lung will collapse and lead to the development of a potentially fatal pneumonia. Joe is given the choice between an analgesic that poses a very small risk of spinal cord

¹⁰ Chapter Five, 158-163.

damage (such as a thoracic epidural), and one that is considerably less effective but which poses no such risk (such as an intravenous narcotic infusion).¹¹

Here, the thoracic epidural is medically indicated; however, Joe might still rationally choose to receive an intravenous narcotic infusion instead, if he places sufficient weight on the value of pursuits that involve physical activity. For example, suppose that Joe is a professional athlete, and believes that his life would not be worth living if he became paralysed because he would not be able to do the things that give his life meaning. In such a case, the possibility that the more effective analgesia could paralyse Joe might give him reason to believe that the less effective analgesia that did not pose this risk of paralysis was the preferable treatment option.

This case suggests that the role of informed consent requirements in medical practice is not merely to protect the patient from competing third party interests, but also to ensure that the treatment that the patient receives is in accordance with what they want for themselves; and this may or may not coincide with what the physician believes is medically indicated. As I suggested in chapters two and five, even if rational agents agree that they have some reason to pursue an outcome, this does not entail that they will agree on the strength of that reason, relative to their reasons to pursue other good outcomes. It seems plausible to claim that this observation also goes some way towards explaining the intuitive pull of the claim that informed consent requirements in medical practice are to be justified in large part by an appeal to the principle of respect for autonomy. They can be understood as facilitating the patient's self-governance not just because they protect the patient from third party influence, but also because they give

¹¹ I take this example from Savulescu, "Rational Non-Interventional Paternalism," 327.

the patient the power to make their own treatment decisions in accordance with their assessment of the strength of the reasons that they have to pursue various outcomes.

This is not to deny that informed consent requirements may have *other* justifications. As my arguments at the end of chapter five suggest, the nature of the above justification may also be understood to appeal to an understanding of beneficence that incorporates (but is not exhausted by) the claim that autonomy has final prudential value. Furthermore, we might also point out that the informed consent procedure may be regarded as integral to establishing a relationship of trust between doctor and patient.¹² It may also serve to provide physicians with a record of what has occurred in the course of treatment that they can appeal to in cases of litigation.¹³ Accordingly, even if we believe that informed consent requirements are often to be justified by the fact that they are in place to safeguard the patient's autonomy, we should not take this to mean that this is their *only* justification.¹⁴

The view that informed consent procedures are to be justified in part by an appeal to the value of patient autonomy has previously been treated almost as a truism in bioethics. Indeed, as I illustrated above, Faden and Beauchamp simply define

¹² O'Neill, *Autonomy and Trust in Bioethics*, 145; see also Bok, *Lying*, 11, 26–27, and 63; Jackson, "Telling the Truth," 491. For recent challenges to this view, see Eyal, "Using Informed Consent to Save Trust."

¹³ Brock, *Life and Death*, 47–48.

¹⁴ Some philosophers have argued that informed consent requirements cannot be justified by an appeal to respect for autonomy because observing these requirements is not *necessary* for respecting autonomy. See Taylor, *Practical Autonomy and Bioethics*, 133; Dworkin, *The Theory and Practice of Autonomy*, 103; Archard, "Informed Consent". I do not find the objections that these philosophers press convincing, although space does not allow for a discussion of them here. However, even without considering the merits of the arguments underlying these objections, they are not problematic for the view that part of the justification for informed consent requirements is that they are in place to safeguard autonomy. This view is compatible with the claim that informed consent requirements may have other justifications, justifications that these objections might point to.

informed consent as a type of autonomous authorisation; similarly, Young simply asserts that autonomy is the “ultimate moral foundation”¹⁵ of informed consent. However, some philosophers have objected to this view by claiming that informed consent requirements are not sufficient for protecting autonomy. For instance, Manson and O’Neill claim that a decisive problem with justifying informed consent procedures by appealing to the principle of respect for autonomy is that informed consent requirements do not ensure that patients will choose autonomously; they only require that physicians respect the choices that the patient actually makes, whether or not this choice is autonomous or rational.¹⁶ Not only that, but Manson and O’Neill also claim that if informed consent requirements were reformulated so that they would protect only rational, autonomous choices, then informed consent requirements would become too demanding for the vast majority of patients.¹⁷

Manson and O’Neill invoke a broad understanding of ‘rational choice’ here. For instance, they claim that theories of rational autonomy can understand rational choice as “reflectively evaluated, or endorsed by second order desires”.¹⁸ However, it is clear that the standard view of informed consent and autonomy that they attack does not incorporate conditions pertaining to the rationality of the patient’s choice. Indeed, Faden and Beauchamp explicitly reject the suggestion that a theory of autonomy should incorporate a condition that requires that the patient’s choice be consistent with their

¹⁵ Young, “Informed Consent and Autonomy”, 441.

¹⁶ Manson and O’Neill, *Rethinking Informed Consent in Bioethics*, 21.

¹⁷ Ibid.

¹⁸ Ibid, 21.

reflectively accepted values because they agree with Manson and O'Neill that such a condition would make autonomy too demanding.¹⁹

As the preceding chapters of this thesis should make clear, I believe that Manson and O'Neill are correct to claim that the standard view of autonomy and informed consent is inadequate. In my view, if we are to claim that a primary purpose of informed consent requirements is to safeguard patient autonomy, then we should incorporate conditions pertaining to the *rationality* of the patient's choice into a theory of informed consent. However, as I explained above, Manson and O'Neill reject this solution because it would, they claim, make the standards of informed consent too demanding. In view of this, it seems that two strategies are possible. First, we could abandon the project of justifying informed consent requirements by an appeal to the principle of respect for autonomy. This is the strategy that Manson and O'Neill adopt; they argue that we ought to view an agent's provision of consent to a procedure as a waiver of an ethical and/or legal norm against performing the act in question, in limited ways in a particular context.²⁰ On the other hand, we might maintain that informed consent requirements are to be justified by an appeal to the principle of respect for autonomy, and supplement informed consent requirements with conditions that will ensure that the patient's choice will be rational, but which will not render informed consent requirements too demanding. Over the course of the next two chapters, I shall adopt the

¹⁹ Faden and Beauchamp, *A History and Theory of Informed Consent*, 262–264; Beauchamp and Childress, *Principles of Biomedical Ethics*, 59.

²⁰ Manson and O'Neill, *Rethinking Informed Consent in Bioethics*, 72. See also 69–84.

latter strategy.²¹ In the remainder of this chapter, I shall consider what ramifications my view of autonomy has for the voluntariness elements of informed consent.

II Rationality and Undue Influence

In earlier chapters of this thesis, I claimed, following Ekstrom, that an agent will be reflectively autonomous with respect to their motivating desire if they endorse that desire with a rationally warranted preference that is personally authorized by virtue of the fact that it coheres with their character system at that time. In turn, a preference will be rationally warranted if the agent sustains it on the basis of a (non-irrational) belief that the object of the motivating desire it concerns is good in a reason-implying sense. Furthermore, I suggested that an agent is practically autonomous if she has the necessary freedoms to act effectively in pursuit of the ends that she is motivated to achieve.

²¹ I cannot argue fully for why I reject Manson and O'Neill's strategy here. However, it is worth pointing out what I believe is problematic about their theory. First, it is not clear that their theory really divorces autonomy from informed consent; after all, if consent transactions are meant to signify the patient's waiving a significant legal or ethical norm, it must surely be the case that the patient should still be autonomous with respect to their decision to waive the norm. Manson and O'Neill do warn against the possibility of bogus consent, in which consent is solicited in ways that violate ethical norms; they give the examples consent following coercion, force and duress as examples of bogus consent. See *ibid*, p. 92. However, what, we might ask is the basis for *these* ethical norms? Manson and O'Neill do not offer an explanation, but one plausible explanation would be that coercion, force and duress are wrong because they violate the patient's autonomy. Elsewhere, O'Neill has claimed that these sorts of norms can be grounded in a Kantian 'principled autonomy'; see O'Neill, *Autonomy and Trust in Bioethics*, Chapter Four, especially pp. 83-86. However, this view assumes a Kantian account of obligations and rights that is contentious.

I shall claim that we should supplement the conditions set out in the standard view of autonomy and informed consent with a condition pertaining to the rationality of the patient's choice. Before defending this claim, I shall begin my analysis of the voluntariness element of informed consent by briefly considering Faden and Beauchamp's conditions through the lens of the rationalist account of autonomy that I have developed.

Consider first the condition of intentionality. Faden and Beauchamp claim that an action must be intentional in order for it to be autonomous, and that “. . . an intentional action is action willed in accordance with a plan”.²² Among the category of non-intentional acts, they include:

. . . things that persons do inadvertently, certain habitual behaviours, and instances of so called occurrent coercion in which a person is physically forced by another to do something.²³

This is congruent with the rationalist account of autonomy that I have developed in this thesis. However, the condition of intentionality is a relatively weak condition on the voluntariness of an agent's action. As such I shall not pursue this issue in detail here, and instead turn to condition (3) regarding the absence of third party controlling influences, or undue influence.

²² Faden and Beauchamp, *A History and Theory of Informed Consent*, 243.

²³ Ibid.

We may say that to exert undue influence over another is to influence an agent's decision in a manner that overrides their autonomy with regards to that decision. The problematic issue is to sort those forms of influence that serve to undermine the patient's autonomy in this way from those that do not. Faden and Beauchamp's account provides us with few clues about how to do this; it simply *stipulates* that coercion, psychological manipulation, and the manipulation of information are examples of undue influence. However, whilst the influences that they mention can plausibly be understood as undermining autonomy in most cases, it seems that Faden and Beauchamp simply rely on the intuitive plausibility of this fact in order to justify their partially defining autonomous acts as those that are carried out in the absence of such influences.

In contrast, if we understand autonomy in the manner that I have defended in this thesis, then we can offer an account of why these sorts of influence can undermine autonomy, whilst also explaining why other influences on patient decision-making are compatible with (and may even enhance) the patient's autonomy. To begin this analysis, it seems right to claim that physicians might plausibly exert undue influence by way of coercion. For example, suppose that Joe's physician told him that if he refused to comply with her recommendation of surgery, then she would not prescribe him any painkillers until he did so. The analysis of coercion that I provided in chapter four explains why this sort of influence would undermine autonomy; since I discussed this at length in chapter four, I shall not pursue the issue in detail here. Furthermore, we can follow Brock in acknowledging that instances of physicians outright coercing their patients into some treatment decision are probably rare.²⁴

²⁴ Brock, *Life and Death*, 44.

As Faden and Beauchamp point out, another way in which a physician could exert undue influence over their patient is by psychologically manipulating them. As I discussed in chapter two, one way in which an agent might be manipulated is if, as a result of third party interference, they come to adopt a motivating desire that they do not personally authorize. Furthermore, an agent, Smith, may be manipulated at a deeper level if another agent, Jones, intervenes to ensure that Smith bypasses the *cognitive* element of her decision-making process in her decision to sustain a certain preference, in so far as Jones' intervention serves to ensure that Smith will not make her decision in the light of her (non-irrational) *beliefs* about the good. As I explained in chapter two, both sorts of manipulation undermine reflective autonomy.²⁵

Faden and Beauchamp suggest ways in which physicians may exert undue influence in these sorts of ways through subliminal suggestion, and by appealing to emotional attitudes such as guilt.²⁶ To illustrate, a physician could psychologically manipulate a patient who refuses a treatment by telling them that they are just 'being awkward', or by telling the patient that all of their other patients just 'do what their doctor says'. In such cases, the physician is attempting to influence the patient, not by appealing to reason-giving facts about the nature of the treatment that could give them reasons to change their decision, but rather by appealing to a non-rational bias that the patient may have to conform to a 'norm' of 'the good patient' perpetuated by a medical authority.

Perhaps the most salient way in which physicians can exert undue influence over their patients is by deceiving them into holding false beliefs about their treatment options by manipulating the information that they receive. The conditions that obtain in

²⁵ Chapter Two, 75-76.

²⁶ Faden and Beauchamp, *A History and Theory of Informed Consent*, 366–367.

the context of medical decision-making are ripe for this sort of undue influence, since there is often a wide knowledge gap between the physician and their patient in their understanding of the relevant medical facts pertaining to the patient's condition. Moreover, there are various ways in which a physician can either intentionally or unintentionally lead a patient to develop a false belief. Most obviously, they can simply provide the patient with false information about their treatment. However, there are also more subtle means of deception. For instance, the physician may not provide the patient with any false information, but simply be selective about the information that they choose to divulge to a patient, so that the latter forms an inaccurate impression of their treatment options. To illustrate, consider the following case:

David undergoes a prostate examination, and the physician finds a tumour. Further tests reveal that the tumour is currently benign, but suggest that it has a 40% chance of developing into a malignant one in the coming years. David's physician recommends immediate surgery to remove the tumour, but fails to tell him that the operation will make him sterile.²⁷

On the theory that I have defended here, if a patient holds certain false beliefs then this can serve to undermine their practical autonomy, if they render the agent 'informationally cut-off' from achieving their ends. Notice that on this understanding a physician who *negligently* provides a patient with false information might fail to adequately respect their patient's autonomy. For example, if a physician tells their

²⁷ For a similar case regarding the disclosure of inherent risks of surgery, see *Cobbs v. Grant* (8 Cal. 3d 229) 1972.

patient (truthfully) that they believe that some treatment has a high probability²⁸ of ameliorating their condition, when in fact it has no chance of doing so, then the patient's decision to choose to receive that treatment will be disconnected from the intended end that underlies their decision to seek treatment in the first place, namely, the end of ameliorating their condition. Even though the physician here does not *intend* to influence the patient to form false beliefs, this misinformation can still serve to undermine the patient's practical autonomy.

Thus, the negligent provision of incorrect information that leads a patient to form a false belief can undermine patient autonomy. What about David's case, in which the physician negligently *omits* correct information? The reason that this undermines patient autonomy is more complex. We can begin by observing that when a patient seeks the counsel of a physician, one of the ends that they are likely to be pursuing in doing so is the amelioration of their condition. However, a patient's health is not the *only* value that informs the way in which they make their treatment decisions; their decision may have implications for their ability to pursue other goods, such as having children in David's case. In my view, the reason that the physician's negligent omission in David's case would infringe upon his autonomy is that it would take away his ability to decide which of his values should take precedence in his treatment decision. Given the wide expertise gap between physicians and their patients, it seems that patients will normally *expect* that their physicians will provide them with information about their

²⁸ Of course, doctors cannot be *certain* that a particular treatment will have its predicted effects. However, they are warranted in telling patients that certain outcomes have higher probabilities than others in light of the available evidence. Patient autonomy does not require that they know whether a certain treatment will work or not (although this knowledge would, it seem normally increase their autonomy). Rather, patient autonomy requires that they are aware of the probability of a treatment option leading to a particular outcome. Only then will they be in a position to rationally weigh competing treatment alternatives. See my discussion of related issues in Chapter Three, 100

treatment options that might reasonably be thought to bear on their treatment decision. If a physician does not tell their patient about a risk accompanying a treatment option, then the patient is unlikely to believe that the treatment option poses any such risk. This is particularly problematic when the omission is intentional; in such cases of omission, the physician is effectively deciding which of the patient's values should take priority here, and acting on the basis of that assessment without consulting the patient. I shall consider the matter of what physicians should disclose to their patients in more detail in the following chapter.

This analysis shows that much of what Faden and Beauchamp claim is congruous with my theory of autonomy. However, in view of some of the problematic cases that I highlighted in the introduction to this thesis,²⁹ I submit that we must supplement Faden and Beauchamp's conditions with one pertaining to the rationality of the agent's choice. Thus, I suggest that the following ought to be understood as a necessary condition of the voluntariness element of informed consent:

Rationality Condition: If an agent is to provide informed consent to some intervention, then they must also endorse their desire to undergo that intervention with a personally authorized preference.

With this condition in place, this view is able to account for many of the counter examples to Faden and Beauchamp's theory; for instance, it is clear that Jane, the sufferer of bulimia I considered in the introduction does not endorse her desire with a personally authorized preference. Moreover, such an account of informed consent is no

²⁹ Introduction, 9-10.

longer open to Manson and O'Neill's objection that '(i)nformed consent requirements protect actual choices, which are often not rational choices'.

As I suggested above, the main source of objection to the sort of rationality condition that I have proposed is that it might make the conditions of informed consent too demanding. This is a serious objection that this account must answer. However, it is an objection concerning the standard of competence that this condition implies rather than objection to the condition itself. As such, I shall postpone my consideration of it until I am in a position to discuss the question of competence in the following chapter.

Another objection to this condition might be that it is wholly impractical; that is, one might ask how physicians could possibly ascertain whether patients rationally endorse their motivating desires in the way that I have suggested. However, this worry is ill-founded. Whilst I have phrased the condition in philosophical terminology, all that the condition requires is that patients are able to justify their treatment decisions by appealing to reasons that they are able to recognise as corollaries of their evaluative judgments. Accordingly, all that the condition demands of physicians is that they must make efforts to find out about the evaluative judgments that undergird their patient's treatment decision. In the final section of this chapter, I shall argue that this consideration suggests that we ought to reject the widely endorsed shared decision making model of the doctor patient relationship. On the model that I defend, there may be ways in which physicians can exert influence over their patients *without* necessarily undermining their autonomy.

III The Shared Decision Making Model and Liberal Rationalism

According to the ‘shared decision-making’³⁰ (henceforth ‘SDM’) model of the doctor-patient relationship, the physician’s role in a treatment consultation is to provide the patient with the relevant descriptive facts about their condition and treatment options, whilst the patient’s role is to evaluate the different options available to them. On this account, it is usually assumed that physicians would be exerting undue influence if they were to encroach on the evaluative domain of the decision-making process. For instance, Quill and Brody suggest that on this account:

The physician should objectively answer questions but should avoid influencing the patient to take one path or another, even if the physician has strong opinions or if the patient asks for advice.³¹

In this section, I shall argue that the SDM model is flawed. I shall argue that minimizing the physician’s evaluative input into the treatment decision is not practicable, and may sometimes have a negative effect on the patient’s autonomy.

Consider first a strict version of the SDM model, according to which the physician should not encroach on the evaluative domain of the treatment decision in *any*

³⁰ For an endorsement of this position, see President’s Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research, *Making Health Care Decisions*.

³¹ Quill and Brody, “Physician Recommendations and Patient Autonomy,” 764. Note that Quill and Brody argue against this model.

way. The first thing to acknowledge against such a models is that allowing physicians to enter into a conversation regarding the values underlying a patient's treatment choice may *help* to safeguard the patient's autonomy. One reason for this is that asking the patient to explain her reasons for choosing a certain treatment might help to militate against the influence of some cognitive biases that can compromise patient autonomy.

As Levy points out, it seems that cognitive biases can compromise a patient's autonomy in two different ways. First, they might lead the patient to misapply their values by causing them to form a false belief about the world; second they might cause a patient to act in a way that does not reflect her values.³² Cognitive biases that can compromise autonomy in the first way include *inter alia*, the framing effect (which I shall discuss below), and the phenomenon of motivated reasoning, in which agents regard an argument as fallacious simply because they are already predisposed to reject its conclusion.³³ Similarly, a patient may be predisposed to reject certain sorts of treatment, and will regard their physician's arguments in favour of their receiving that treatment to be fallacious, regardless of their validity.

Cognitive biases that can compromise autonomy in the second way include the bias that agents exhibit towards the present, and their reluctance to consider the possibility of future harms when they weigh their reasons for pursuing different outcomes.³⁴ For example, a patient may reject their physician's recommendation that they stop smoking, not because they believe that the pleasure they get from smoking is more valuable than increasing the probability of a longer life-span, but rather because

³² Levy, "Forced to Be Free?", 298.

³³ *Ibid.*, 298.

³⁴ *Ibid.* See also, Brock, *Life and Death*, 84.

they fail to attend to the disvalue of the later consequences that their smoking habit will lead to.

In view of this, one way in which physicians might seek to safeguard their patient's autonomy is by attempting to disabuse their patients of the cognitive illusions that lead them to develop false beliefs about the world, or to act in a manner that does not reflect their evaluative judgements. Arguably, this might fall under the purview of the physician understood as a mere fact-provider. However, since cognitive biases can affect autonomy by causing patients to act in ways that do not reflect their values, if physicians are to disabuse their patients of cognitive biases that have this effect, then they must encroach on the evaluative domain of their patient's decision-making process in order to establish whether the patient's treatment decision is the outcome of an irrational bias, or whether it is rationally warranted, given the patient's conception of the good.³⁵

This might seem to be something of a straw man argument, since defenders of the SDM model might respond that they do not intend to claim that physicians should not encroach on the evaluative domain of the patient's decision making process in *any* way; instead they might intend to claim only that we should seek to minimize the extent to which this occurs. *Prima facie*, it seems that a physician could enter into a dialogue with their patient about their values without explicitly advocating any of their own values.

However, it may often not even be possible for physicians to relay medical information to their patient in this value-neutral way. The reason for this is that core

³⁵ It might be argued that physicians should seek to debias their patients: however, Levy is sceptical of this possibility. Levy, "Forced to Be Free?" 297.

medical concepts such as health and disease are themselves value laden, in so far as they are commonly understood (particularly in the conversational context of a treatment discussion) to imply certain value judgements.³⁶ For example, whilst telling a patient that their cancer is spreading is a factual description, both patient and physician are likely to understand that there is a judgment implicit within this statement that the cancer's spreading in this way is bad; the very term 'cancer' is understood to be implicitly laden with negative value, given what is commonly known about the disease, even though some cancers are harmless. To illustrate this point, consider Brewin's description of the following real life case:

(The patient's) great fear was that he might one day develop cancer . . . A biopsy specimen of an ulcer on his lip confirmed squamous cell carcinoma . . . After a short course of radiotherapy it would heal, and probably never cause him any further trouble. "It's not cancer, is it?", he asked, his eyes moist with tears . . . I assured him that it was not.³⁷

The example of the term 'cancer' and its implicatures in the conversational context of medical treatment discussion suggests that another problem with the SDM model is that some of information that the physician might give to their patient can include concepts that may be understood by the patient to implicitly incorporate certain value judgments.³⁸ Moreover, some of the value judgments that the patient believes to be

³⁶ For a classic account of conversational implicature, see Grice, *Studies in the Way of Words*.

³⁷ Brewin, "Telling the Truth," 1512.

³⁸ Brock, *Life and Death*, 56 and Savulescu, "Rational Non-Interventional Paternalism," 328 make a similar point.

implicit within certain terms may not be intended by the physician, and may rely on a misunderstanding on the patient's part.

We might also observe that physicians must make certain evaluative judgments in deciding upon what information to disclose to patients. For example, instead of telling a terminally ill patient about the intricate details of what will happen to them as they succumb to their condition, the physician might judge that she ought not to provide this information, and should instead simply tell the patient that she will be made 'as comfortable as possible' in her final hours. Furthermore, it seems that physicians must make certain evaluative assumptions when they are deciding what treatment options to provide a patient with. In both cases, it seems that the physician's evaluative judgements will inevitably bear upon what is (and is not) disclosed.

In a related sense, even when the facts that are communicated do not seem to incorporate a value judgment, the manner in which they are communicated might. Tversky and Kahneman have demonstrated that the way in which information is framed can affect agents' decision-making; for instance, their experiments suggest that if information is framed positively, then agents deciding on the basis of that information are more likely to be risk averse than if the information is framed negatively.³⁹ Citing a study by McNeil *et al.*,⁴⁰ Savulescu uses the following example to illustrate the importance of the framing effect in medical consultations:

. . . (l)ung cancer can be treated by surgery or radiotherapy. Surgery is associated with greater immediate mortality (10 per cent v 0 per cent mortality), but better long-term

³⁹ Tversky and Kahneman, "The Framing of Decisions and the Psychology of Choice."

⁴⁰ McNeil *et al.*, "On the Elicitation of Preferences for Alternative Therapies."

prospects (66 per cent v 78 per cent five-year mortality). The attractiveness of surgery to patients is substantially greater when the choice between surgery and radiotherapy is framed in terms of the probability of living rather than the probability of dying.⁴¹

The problem with the SDM model that these considerations highlight is that the model may not protect patient autonomy in the manner that its adherents might claim. The model might adequately safeguard patient autonomy if patients were always able to make their treatment decisions in accordance with their values on the basis of factual information provided by their physicians; however, the prevalence of cognitive biases suggests that this is often not the case. Thus, if a physician can intervene in a manner that might enable their patients to decide in accordance with their values rather than on the basis of a cognitive bias, then I submit that it is plausible to claim that such an intervention would promote autonomy, rather than representing the undue influence of the physician.

A defender of the SDM model might concede that there are some unavoidable ways in which a physician's values will bear on a patient's treatment decision. However, they might still maintain that we should not allow physicians to impose their values on the treatment decision in avoidable ways. Yet, I suggest that this too would be a mistake; it seems that physicians *ought* to be able to defend certain sorts of values.

To illustrate, suppose that a patient has been stung by a wasp; she tells her physician that she is highly allergic, and the physician tells her that a possibly fatal anaphylaxis is probable unless she receives a shot of adrenaline. Suppose that the patient understands all this, and want to live; suppose further that the patient is aware

⁴¹ Savulescu, "Rational Non-Interventional Paternalism," 328–329; See also Brock, *Life and Death*, 88 .

that the shot will be mildly painful, since she has received one before. Here, the patient's belief that the shot will be painful may give her *some* reason to not have the injection. However, this does not diminish the strength of her reason to have the injection, in view of the fact that the shot will save her life (assuming she wants to continue living).

In such a case, I suggest that the physician would be warranted in strongly recommending that the patient receive the shot, rather than simply pointing out the descriptive fact that the shot will save her life. In doing so, they would be recommending an outcome that all rational agents should agree is good in a strong reason-implying sense. As Brock suggests, part of the physician's role can be to advocate the import of certain sorts or reasons, reasons that reflect the values that shape the profession of medicine.⁴² Brock's claim here is congruous with my suggestion in chapter five, that we should reconceive of beneficence in bioethical contexts as pertaining to impersonal goods that the medical profession is committed to promoting.⁴³ My claim here is that physicians should be permitted to draw the patient's attention to the impersonal reasons that they have to pursue these particular goods. Crucially though, this does not entail that the physician would be warranted in *overriding* the patient's decision in such cases; I shall explain why in my discussion of competence in the following chapter.

With all this in mind, I suggest that physicians should not merely be restricted to providing their patients with medical facts, and leaving it solely to the patient to decide the best course of action. There are a number of ways in which physicians can go beyond this without undermining the patient's autonomy. The first is by simply

⁴² Brock, *Life and Death*, chap. 2, especially 69.

⁴³ Chapter Five, 162.

engaging with the patient in order to find out what their values are, and to tailor a recommendation on the basis of those values. Alternatively, the physician might begin the dialogue by telling the patient what is medically indicated, and asking whether the patient objects; if nothing else, this interaction can at least form the foundation of a discussion about the patient's values. To illustrate, in the example of Joe considered above, the physician might explain the relevant facts about Joe's analgesia options, recommend that Joe choose the thoracic epidural, and ask Joe whether there are any reasons why this treatment plan would not be best for him.

If the patient and physician disagree about the best course of action all things considered, it is likely that this will either be because the patient values ends that the physician does not, or because the physician believes that the patient is weighing their values irrationally. On the theory of autonomy that I have developed here, even if the patient places value on something that the physician does not, then as long as the patient's belief concerning the value of the end in question meets a minimum threshold of epistemic rationality, then the patient may be autonomous with respect to her desire to pursue that end. However, if the physician is of the opinion that the patient is weighing her values *irrationally*, although I claim that physicians would not be warranted in overriding the patient's choice, I also claim that physicians would not be exerting undue influence if they were to challenge their patients, and attempt to persuade them to their way of thinking by highlighting the nature of the reasons that they are considering. Savulescu terms this sort of approach to the doctor-patient relationship "Liberal Rationalism".⁴⁴ In support of this position, he points out that this sort of rational discussion can enhance patient autonomy. He writes:

⁴⁴ Savulescu, "Liberal Rationalism And Medical Decision-making".

Far from frustrating personal autonomy, rational discussion with those who hold different values promotes autonomy. Our evaluations become better informed, more vivid, and through challenge and defence, they become more clearly important to us.⁴⁵

In closing, it is prudent to pre-empt some objections to this model. Veatch argues that one reason that physicians ought not to make recommendations to their patients is that they themselves are subject to biases about what values their patients ought to prioritise; put simply, the worry might be that physicians will be heavily biased towards the value of health and thus base their recommendations on what is medically indicated.⁴⁶

However, this would only be a problem with the liberal rationalist model that I am endorsing here if the physician's recommendations were to be treated as sacrosanct; in this case, we would simply be replacing a decision that the patient makes whilst subject to cognitive biases with a decision that a physician makes whilst subject to other cognitive biases. However, on the liberal rationalist view, the physician's recommendation is to be understood as *a part* of a rational discussion, not the end point; in making a recommendation on the basis of what is medically indicated and their understanding of the patient's values, the physician should be understood as making a positive suggestion in order to initiate a dialogue about what ought to be done, and not making a final judgment about the matter. Moreover, on this model, whilst both parties

⁴⁵ Ibid., 126.

⁴⁶ Veatch, "Abandoning Informed Consent," 11.

may bring biases to the treatment dialogue, each party's biases may at least be mitigated by being set in opposition to the different biases of the opposing party.

Veatch also argues that doctors are not experts in making value judgments, and thus should not be encouraged to make them in their provision of treatment recommendations. In response, whilst we might not go so far as to claim that doctors are 'experts' in making value judgments, it would be to do them a disservice to claim that they have *no* relevant experience or expertise that could usefully be brought to bear on a treatment decision. First, as Savulescu points out, a physician's medical knowledge will give them knowledge about the patient's circumstances of which the patient herself is not aware;⁴⁷ and this knowledge may be crucial in determining how the patient ought to go about attempting to realise their own ends. Second, whilst physicians may be subject to biases of their own, many of them will have experience of dealing with patients who are faced with making important medical decisions, and the emotional and cognitive biases that can affect decision making in this context. Finally, I have already suggested that the profession of medicine is committed to certain values, and that physicians are warranted in promoting them.

Conclusion

In this first chapter on the nature of informed consent and autonomy, I have argued that we ought to supplement our theory of informed consent with conditions that reflect a philosophically adequate theory of autonomy. I suggested that we ought to

⁴⁷ Savulescu, "Liberal Rationalism And Medical Decision-making," 120.

supplement the voluntariness element of informed consent with a condition pertaining to the rationality of the patient's preference to pursue some treatment outcome, and considered the implications that this has for understanding undue influence and the doctor-patient relationship. In the next chapter, I shall continue my analysis of informed consent, by considering the information elements of the doctrine, and the concept of competence to consent.

Chapter Eight- Informed Consent and Autonomy Part

Two: Disclosure, Understanding and Competence

In this chapter, I shall consider the implications that my view of autonomy has for what I termed the information element of informed consent, and the question of competence. In the first part of this chapter, I shall develop an account of what it is for information to be ‘material’ to a patient’s decision, by analysing the types of information that patients must understand in order to be autonomous with respect to their treatment decisions. This will complete my analysis of the two elements of informed consent identified in the previous chapter. In view of this analysis, and my arguments in the previous chapter, I shall conclude by considering the standard of competence that informed consent demands on my account, and by arguing that incorporating a rationality condition into a theory of informed consent does not make the standard of competence too high.

I The Information Element of Informed Consent - What Should Be Disclosed?

In the legal domain, questions pertaining to the information element of consent have been framed not so much in terms of what patients need to *know* in order to be autonomous with respect to their treatment decision, but rather in terms of what doctors need to *disclose* to their patients in order to adequately inform them about their treatment options. Whilst I shall suggest below that it is misleading to claim that having

certain information disclosed to one is a necessary condition of informed consent, it is prudent to acknowledge two key positions that have been reflected in the verdicts of landmark legal cases pertaining to standards of disclosure.

According to one view, which we might term the *physician-oriented* view of disclosure, it should be solely up to the physician, in their professional capacity, to decide what to disclose to their patient about their treatment options. In turn, the physician's decision here is understood to be governed by a standard of disclosure endorsed by the professional community, according to which information ought to be disclosed if the majority of physicians within that community would customarily make such a disclosure. Such a view is apparent in *Robinson vs Bleicher 1997*, in which the court ruled that the physician's duty to disclose information is:

. . . measured by the standard of the reasonable medical practitioner under the same or similar circumstances, and must be determined by expert medical testimony establishing the prevailing standard and the defendant practitioner's departure therefrom.¹

However, other legal verdicts seem to endorse a *patient-oriented* standard of disclosure, according to which doctors should disclose to their patients any information that a hypothetical reasonable patient would want to know. For instance, in *Canterbury vs Spence (1972)*, the court claimed that:

¹ Robinson v. Bleicher (559 N.W.2d 473) 1997.

. . . the patient's right of self-decision shapes the boundaries of the duty to reveal. That right can be effectively exercised only if the patient possesses enough information to enable an intelligent choice.²

Since the court here also observed that it will often be difficult for practitioners to know what information is material to a given patient's decision, they suggest that physicians ought to disclose information that it is reasonable for them to believe that the patient would want to know.³

Below, I shall echo this patient-oriented legal standard of disclosure by arguing that physicians ought to employ a standard of disclosure that takes into account what all patients have impersonal self-interested reasons to desire, and what the particular patient in question has personal self-interested reasons to desire. Prior to making this argument, it should be noted that standards of disclosure often have a legal, rather than philosophical justification. In the previous chapter, I suggested that although respect for patient autonomy provides a large part of the justification for informed consent requirements, another justification for them is that they represent a form of defence for physicians against legal liability. This is particularly clear with regards to requirements regarding the disclosure of information to patients. Although some theorists claim that disclosure of information is a *necessary* condition of informed consent,⁴ disclosing information to a patient may not always be necessary for safeguarding their autonomy if they are *already* aware of that information. To illustrate, it is common knowledge that one needs to apply pressure to a wound in order to stop it from bleeding. It seems that I

² *Canterbury v. Spence* (464 F.2d 772), 1972.

³ *Ibid.*

⁴ For example, see Beauchamp and Childress, *Principles of Biomedical Ethics*, 2nd Ed, 67.

could surely provide a valid consent to a physician's application of a bandage to my wound, without their having disclosed to me that he needs to apply a bandage in order to stop my wound from bleeding.

Of course, given the expertise gap between physicians and their patient, and since physicians cannot be certain about their patient's prior knowledge, patients will often need to have information disclosed to them. However, from a theoretical perspective, it should be clear that the disclosure of information is not a *necessary* condition of a patient's making an autonomous decision in all cases. Accordingly, it seems that to claim that disclosure itself is a necessary condition of informed consent is to move the concept of informed consent away from an autonomy-based justification towards a justification based on avoiding legal liability. Moreover, it might be claimed that those who make this claim are invoking an institutional understanding of informed consent that differs from the understanding of informed consent understood as an autonomous authorisation.⁵

If we believe that the primary justification of informed consent requirements should be to safeguard patient autonomy, then in order to establish the standard of disclosure that physicians should employ, it seems that we should carry out a philosophical investigation into the sorts of information that a patient must be aware of in order to be autonomous with respect to their decision. In their analysis of the condition of understanding, Faden and Beauchamp point out that a patient need only have a substantial understanding of their treatment and, that in consenting, they are authorising a physician to carry out that treatment. In order to cash out what it is for a patient to have 'substantial' understanding, they write:

⁵ See Faden and Beauchamp, *A History and Theory of Informed Consent*, 276–277 for a discussion of this alternative understanding of informed consent.

. . . a person must understand those propositions about (some medical intervention) *R* and about authorizing *R* that are germane to the person's evaluation of whether *R* is an intervention the person should authorize. This criterion is entirely subjective.⁶

Faden and Beauchamp term the set of propositions that they are referring to here 'material information'.⁷

Although I agree with Faden and Beauchamp that patients need only have 'substantial' (rather than full) understanding of their treatment options (for reasons that I explain below), I shall argue that the criterion of materiality that they use to cash out the notion of 'substantial understanding' is unsatisfactory as it stands. Whilst Faden and Beauchamp are right to point out that different patients are likely to regard different information to be pertinent to their treatment decision, their account fails to explain precisely *how* patients are to subjectively assess whether or not certain information is pertinent or not; they simply point out that a person's long range goals and values can affect how individuals value various act descriptions.⁸

Furthermore, it is not clear that the standard of understanding can be *entirely* subjective as Faden and Beauchamp claim. There are three problems with this view. First, it seems that some information can be so fundamental to the nature of a decision that it will be material to that decision regardless of whether an individual deems it to be so or not. For instance, it is difficult to imagine how a patient could be autonomous with respect to their decision to undergo an anaesthetic if they failed to understand that

⁶ Ibid., 302.

⁷ Ibid.

⁸ Ibid., 302–303.

undergoing an anaesthetic will render them unconscious. It seems inappropriate to claim that the materiality of this information to a patient's decision to undergo an anaesthetic is contingent upon the patient's own assessment of the materiality of that information.

Second, in view of the fact that patients are normally not experts in medicine and may be in a vulnerable state owing to the nature of their condition, it seems that they may make mistakes about what information is and is not material to their treatment decision. Indeed, as I shall explain below, empirical evidence suggests that patients are subject to a number of cognitive biases that can distort their understanding of their condition and treatment options. Finally, it is not clear that a purely subjectivist account of materiality is practically realisable, since, as the court in *Canterbury vs Spence* pointed out, it will often be difficult for practitioners to know what information a patient believes to be relevant to their decision, and it is the physician who has to decide what information to disclose to their patient. In view of these problems, I shall, in the next section, attempt to go beyond Faden and Beauchamp's purely subjective account of materiality by appealing to the theory of autonomy that I have developed in this thesis.

II Material Information

To begin this analysis, it is prudent to reconsider the role of beliefs in the theory of autonomy that I have developed in this thesis. I have suggested that in order for an agent to be reflectively autonomous with respect to their motivating desire, they must personally authorize that desire with a rational preference that they sustain on the basis of a (non-irrational) belief that the object of their motivating desire is good in a reason-

implying sense. Furthermore, I suggested that in order for an agent to be practically autonomous, they must hold certain true beliefs about how to act effectively in pursuit of their ends; in Mele's terms, autonomous agents cannot be 'informationally cut off' from acting effectively in pursuit of their own ends.

With this analysis of the role of beliefs in autonomous agency in mind, we can begin to consider what sorts of information, and how much of it, patients must understand in order to be autonomous with respect to their treatment decisions. I suggest the following analysis of materiality:

Materiality: Information is material to a particular patient's treatment decision if it concerns facts that give the patient reasons to choose or reject a certain treatment option.

Even before elaborating on this account of materiality in more detail, it should be clear that this account departs from Faden and Beauchamp's account in two ways. First, on my account, facts that give patients impersonal self-interested reasons to choose or reject a treatment will be material to the patient's decision, regardless of the patient's own assessment. Second, on this account, information that a patient mistakenly believes to be relevant to their treatment decision will not be material if it does not concern reason-implying facts.

The first issue to address with regards to this analysis is the question of what sorts of reasons might be involved in the context of a patient's treatment decision. It seems that a patient's treatment decision is most readily understood as being motivated

by an instrumental desire, and that one of the primary valued outcomes that patients want their desired medical treatment to bring about is, broadly speaking, the amelioration, prevention, or management of their condition, or perhaps more fundamentally, the increase in welfare that this would entail. However, it should be stressed that health is not the *only* good that a patient will weigh in their assessment of their treatment options; patients will also assess the extent to which a possible treatment might impact on their future freedom to act effectively in pursuit of other ends that they value. In some cases, such as the example of Joe considered in the previous chapter,⁹ the medically indicated treatment could potentially have adverse effects on the patient's pursuit of other ends that they value. This analysis is congruous with Brock's claim that:

The goal of the requirement that consent be informed is for patients to achieve a sufficient understanding of their condition and possible treatments so that they can make a sound assessment of which treatment . . . will best serve their goals and values.¹⁰

To expand on this account, it seems that information concerning two broad aspects of a patient's treatment decision are likely to concern reason-implying facts.¹¹

⁹ Chapter Seven, 201-202.

¹⁰ Brock, *Life and Death*, 47.

¹¹ Following Grisso and Appelbaum, we might claim that patients must also understand that they *themselves* are in need of medical attention, since this fact gives them a reason to seek treatment. Grisso and Appelbaum, *Assessing Competence to Consent to Treatment*, 31. Furthermore, in order to militate against controlling influences of the sort that I identified in the previous chapter, it might also be claimed that patients must understand that they may withdraw their consent at any time.

First, it seems that information pertaining to the *nature* of the proposed intervention will often concern reason-implying facts; for instance, it seems plausible to claim that the fact that an intervention will be painful or invasive provides patients with reasons not to choose that treatment (although these reasons will often not be decisive). Second, and perhaps more importantly, it seems that facts pertaining to the probability of an intervention's bringing about some outcome will also be reason-implying. Such facts will include not only those pertaining to the probability of an intervention's ameliorating the patient's condition, but also those pertaining to the risks attending the intervention and possible side-effects. For example, if one was aware that a surgery has a high chance of leaving one paralysed, it seems that this would provide one with a (defeasible) reason not to undergo that surgery.

I have thus far identified different *types* of information that will be material to a patient's decision. Let us now consider the *extent* to which patients should be made aware of these different aspects of their treatment decision. Initially, it might be claimed that the more information that patients have disclosed to them about their treatment options, the more autonomous that they will be with respect to their treatment decision. However, this is not necessarily the case. Consider first the patient's understanding of the *nature* of their condition or a proposed intervention. Clearly, patients cannot be expected to have a *full* understanding of their condition or a proposed intervention; after all, medical conditions and procedures often admit of exceedingly complex descriptions, which, however accurate they might be, are unlikely to aid the patient in their decision-making. The reason for this is that if physicians were to aim to ensure that patients had a full understanding of their treatment options, they would be likely to completely overwhelm their patients with an excess of information that they could not

reasonably be expected to compute, especially given that many patients will be less able to deal with complex information because of their illness.

This sort of worry seems to underlie Faden and Beauchamp's appeal to 'substantial' understanding; in view of the sorts of worries delineated above, they reject the claim that a patient can only be autonomous with regards to a treatment decision if they fully understand the nature of their condition and the treatment that they have chosen.¹² My account of materiality reflects this view; on my understanding of materiality, the patient need only understand information about the nature of their treatment that concerns reason-implying facts. Whilst some information about the nature of the treatment might concern reason-implying facts (for instance, the fact that the intervention is painful), a lot of information will not. For example, information concerning the exact biological mechanism that explains why an antibiotic helps to destroy a bacterial infection will normally not be material to a patient's decision to choose to take antibiotics; such information does not itself concern facts that provide agents with self-interested reasons (although corollaries of that information, such as the fact that this mechanism means that antibiotics can relieve bacterial infections, may).

However, whilst a great deal of information about the nature of a patient's condition or treatment options is unlikely to be material to their decision, it seems plausible that information concerning the foreseeable *outcomes* of their treatment options and their attendant possibilities is always likely to be material to a patient's decision. As I argued in chapter three, one reason why the holding of true beliefs is important for autonomous agency is that we often need true beliefs to act effectively in pursuit of our ends. In lacking true beliefs or holding false ones about the probability of

¹² Faden and Beauchamp, *A History and Theory of Informed Consent*, 240–241 and 300–305.

a medical interventions leading to a certain outcome, it might be claimed that we are rendered informationally cut off from achieving our ends in the manner that Mele suggests is inimical to autonomy.¹³

Yet, disclosing full information about all the possible outcomes of a procedure would not necessarily enhance patient autonomy. The first thing to acknowledge in defence of this claim is that all medical interventions carry some risk of adverse side effects or complications; for example, even over the counter drugs such as Paracetamol can put the user at a small risk of experiencing rashes and hypotension. If patients were always able to understand the nature of small risks and incorporate them into their decision making appropriately, then it might be the case that physicians ought to disclose the risks of all foreseeable possible outcomes in order to increase the patient's autonomy with respect to their treatment decisions. However, research on cognitive biases suggests that patients are not able to compute information about risks in such an unbiased manner.¹⁴ As Cass Sunstein points out, when people have to make a decision in an emotionally charged context such as healthcare, they:

. . . tend to focus on the adverse outcome, not on its likelihood. That is, they are not closely attuned to the probability that harm will occur.¹⁵

¹³ We should recall here that this does not entail that an agent will lack autonomy simply if she fails to achieve her end. Rather, the informational condition I delineated in chapter three stipulates that the agent's poor informational condition should not thwart the possibility of her being successful in achieving her end. See Chapter Three, 100.

¹⁴ See Ingelfinger, "Informed (but Uneducated) Consent". See also Levy, "Forced to Be Free?", and Sunstein, *Risk and Reason* for analyses of relevant empirical evidence.

¹⁵ Sunstein, "Probability Neglect", 62.

As Sunstein points out, the problem with this is not simply that patients are unable to rationally compute information about risks; rather, in disclosing information about risks, physicians may simply alarm their patients, causing them psychological distress that will actually be detrimental to their ability to make a decision.¹⁶ In the terminology that I have used in this thesis, it might be claimed that disclosing information about risks might lead patients to develop epistemically irrational beliefs about the nature of the reasons that they have to choose certain treatment options; furthermore, the distress that they experience in considering the disclosed risks might even prevent them from engaging in the higher-order reflection that autonomy requires at all.

A further problem with insisting on the disclosure of all foreseeable possible outcomes is that the disclosure of this information itself can be harmful to the patient. Consider the ‘nocebo effect’ highlighted recently by Erin-Wells and Kaptchuk.¹⁷ The nocebo effect occurs when physicians induce adverse negative side-effects (via a negative placebo response) by disclosing information pertaining to the risks of those very side-effects. Here, even if informing the patient of the risk of these outcomes might increase their local autonomy with respect to their treatment decision, it can also be understood as restricting the patient’s global autonomy, if we assume (plausibly) that the patient would regard the outcome of avoiding adverse side effects as a goal that she would prefer to achieve over the course of her treatment. Here, disclosure might serve to reduce the agent’s freedom to effectively pursue this end, even though true beliefs normally enhance an agent’s ability to pursue their ends.

¹⁶Ibid, 92.

¹⁷Wells and Kaptchuk, “To Tell the Truth, the Whole Truth, May Do Patients Harm”

With all of this in mind, what conclusions should we draw about material information and the information element of informed consent? I have claimed that we ought to reject Faden and Beauchamp's purely subjective account of materiality, and suggested that the materiality of information about the nature of a treatment and its possible outcomes will depend on whether the information concerns facts that are reason-implicating for the patient. On the basis of this general conclusion, I offer the following additional conclusions that can serve as a starting point in our thinking about individual cases.

First, when material information pertains to possible treatment outcomes, the strength of the reasons implied by these facts will depend on both the goodness of the outcome in question, and the probability that it will occur. These factors are not to be considered in isolation; for instance, even if a possible outcome of some act has a very low probability, the patient might still have strong reasons not to engage in that act if they confer a sufficient amount of disvalue on the outcome in question; for instance, recall the example of Joe in chapter seven.¹⁸

In some cases, even though information can be understood as being material to the patient's decision in so far as it concerns reason-implicating facts, disclosing that information may nonetheless serve to hinder, rather than promote the patient's autonomy in view of the cognitive biases that patients are prone to exhibit. Of course, it seems that there are some measures that physicians might take to partly mitigate some of the cognitive biases that patients are subject to. For instance, it seems that we should follow Sunstein in claiming that if a disclosure about risk is to be worthwhile, then it

¹⁸ For a real life example, see *Cobbs v. Grant* (8 Cal. 3d 229) 1972.

should be “ . . . accompanied by efforts to enable people to put the risk in context”.¹⁹ Physicians should also be informed of the various cognitive biases that they and their patients are prone to exhibit. However, even where it is possible to try and mitigate the influence of these biases, it is not clear how successful such efforts will be.²⁰

In view of this, it seems that the prevalence of cognitive biases amongst patients means there may be cases in which the disclosure of material information will not facilitate the patient’s autonomy, all things considered. In view of these cases, I endorse Levy’s tentative call to introduce ‘informed consent specialists’²¹ who have received specialist training in human rationality, to act as ‘middle-men’ between the physician and their patient in complex cases. However, I also echo Levy’s warning that such specialists would not act as a panacea for the types of irrationality that I have discussed here.²²

In view of the problem of cognitive biases and the potential harmful effects of disclosure discussed above, I suggest further that one illuminating question that physicians can ask themselves when deciding what to disclose to their patients is ‘to what extent would disclosure of *x* restrict the patient’s ability to effectively pursue their autonomously chosen treatment goals?’ If part of the justification of informed consent requirements is that they safeguard patient autonomy, then there may be a theoretical moral justification for not disclosing certain material information to a patient (thereby overriding the patient’s local autonomy), if that is the *only* way in which the physician

¹⁹ Sunstein, “Probability Neglect,” 92.

²⁰ Levy is sceptical of the extent to which it is possible to disabuse patients of their cognitive biases. Levy, “Forced to Be Free?,” 297.

²¹ *Ibid.*, 299.

²² *Ibid.*, 300.

can foreseeably facilitate their patient's *global* autonomy with respect to the achievement of their autonomously chosen treatment goals.²³ However, there are, of course, practical epistemic limitations to when doctors may be justified in believing that they are facing such a case.

I have argued for a patient-oriented model of disclosure that is tailored towards facts that imply either personal or impersonal self-interested reasons for the patient. Of course, there are likely to be practical barriers to implementing this standard. Whilst there are certain facts that we may plausibly assume to imply strong impersonal self-interested reasons for any patient (for example, facts about how to avoid severe pain and how to retain key cognitive capacities say), it seems that the materiality of information about certain aspects of particular treatments will depend in part on facts about the individual in question, including facts pertaining to the weight that they assign the reasons they have to pursue different goods.

In view of this, and in view of my analysis of the doctor-patient relationship in the previous chapter, perhaps the best practical solution is for physicians to use the reasonable patient standard of disclosure defended in legal judgements as a *starting point* of the physician's disclosure, in conjunction with a duty to disclose the fact that patients can ask for any other information, and a duty to engage with the patient to find out more about *their* values. This standard of disclosure will require that physicians disclose information pertaining to what all patients have impersonal reasons to want. However, if we want informed consent procedures to safeguard patient autonomy, then the disclosure of information must go beyond the mere provision of such facts, and become a two way informational transaction, in which the physician does not merely

²³ I draw a similar conclusion with respect to the clinical use of deceptive placebos in Pugh, "Ravines and Sugar Pills: Defending Deceptive Placebo Use."

provide the patient with a list of facts, but in which the physician makes an attempt to elicit the patient's values, and tailors their disclosure in accordance with the information that will be material to the patient in view of *their* values. This is the liberal rationalist model that I defended in the previous chapter.²⁴

III (Rational) Competence

I shall conclude this chapter by considering the abilities that a patient must have in order to qualify as competent to provide valid consent on the account that I have defended. As Faden and Beauchamp explain, the core meaning underlying the concept of competence is “. . . the ability to perform a task”.²⁵ Thus, in considering whether a patient is competent to provide valid informed consent, we are asking whether they have the requisite abilities to perform the tasks that are delineated in the information and voluntariness elements of informed consent.

In the UK, the Mental Capacity Act 2005 (henceforth MCA) states that a person is competent to make a treatment decision if they have, and are capable of exercising during the decision making process, the following necessary abilities:

- a) The ability to **understand** the information relevant to the decision.
- b) The ability to **retain** the information for long enough to be able to make a decision.

²⁴ Chapter Seven, Section III

²⁵ Faden and Beauchamp, *A History and Theory of Informed Consent*, 288.

- c) The ability to **weigh up** the information as part of the process of making the decision.
- d) The ability to **communicate** the decision.²⁶

This view of competence is compatible with the analysis of informed consent that I have offered here. Consider first condition (a): As I argued above in my analysis of the information element of informed consent, patients must understand material information about their condition, their treatment options, and the possible outcomes of their treatment options in order to be autonomous with respect to their treatment decision. Conditions (b) and (c) can be understood to relate to abilities that the patient must have in order to make a decision on the basis of the reason-implicating facts conveyed in the material information. Finally, condition (d) refers to the practical element of the patient's ability to communicate her choice to the physician.

However, there are some inadequacies with the MCA's conditions. Prior to considering these though, it is important to highlight some key aspects of competence, generally conceived.²⁷ The first is that the standards of determining whether or not a patient is competent should be contextually dependent.²⁸ Different local decisions will require different degrees of aptitude in the abilities delineated above; accordingly, although a patient may not be competent to make certain sorts of decisions, this does not entail that they are incompetent to make *any* decisions for themselves. For example, whilst an agent with limited competence may be able to understand material information pertaining to a decision about relatively simple treatments, such as whether

²⁶ "Mental Capacity Act 2005", Part 1, Section 3 (1).

²⁷ Ibid.

²⁸ See also Brock, *Life and Death*, 38.

she ought to have surgery on a broken bone, she may not be able to understand material information pertaining to more complex treatment options which could lead to various possible outcomes, such as in the treatment of cancer. The level of competence required to make a certain decision will depend upon the complexity of the information about the alternative courses of action that is material to the patient.

An agent's competence may also change over time. Whilst an agent may be competent to make a decision at t , they may not be competent to make that decision at a later time $t+1$, by virtue of the fact that they have either lost one of the abilities highlighted above, or because they are temporarily unable to exercise it. Examples of patients in the first category include patients who have lost certain abilities due to the nature of their disease, or because of the cognitive decline that occurs as part of the aging process. Examples of patients in the second category include patients who are temporarily unable to exercise a capacity because they are unconscious, overwhelmed by physical pain, or under the influence of drugs. Finally, as McMahan points out, we should also acknowledge that competence is normally understood to be a range property, in so far as it is:

. . . a property that does not admit of degrees, though it is based on an individual's possession of various cognitive capacities that do admit of degree.²⁹

Although the conditions laid out in the MCA are compatible with the view of autonomy and informed consent that I have espoused here, they leave some important

²⁹ McMahan, *The Ethics of Killing*, 250.

questions unanswered. Consider first condition (a); although the MCA explains (in its comments on the above conditions) that the information relevant to a patient's decision includes information about the reasonably foreseeable *consequences* of each of the patient's treatment choices, it does not offer an explanation of *why* this information is relevant. This is important, because the MCA thereby lacks any criteria for determining the materiality of other sorts of information (such as that pertaining to the nature of the treatment itself) that might be material to the patient. In my analysis of the information element of informed consent, I hope to have offered a richer account of how to determine which information is material to a patient's treatment decision; the patient's capacity to understand and retain information should also extend to the sorts of material information explored above.

Moreover, although the MCA is right to claim that patients must be able to 'weigh up' material information as part of their decision-making process, it does not explain exactly what this entails. In failing to do so, it fails to incorporate a crucial element of autonomous decision-making, as I have understood it in this thesis. Whilst understanding material information is clearly an important element of the agent's decision-making process, the import of this information for practical decision-making is itself dependent upon the bearing that this information has on ends that agents *value* (and their pursuit thereof). However, the MCA does not stipulate that an agent must be able to make their decisions on the basis of any sort of *evaluative weighting* in order to be competent. To see why this is problematic, consider the following case:

Apathetic Andrea: Andrea suffers from clinical depression. Her physician explains to her that there are a number of treatment options available (ECT, anti-depressants,

psychiatric counselling), and provides her with extensive information about each option and their possible outcomes. Andrea understands this information, retains it, and weighs how medically effective each option is against the other. However, Andrea is pathologically apathetic, and does not care at all what happens to her. Although she considers the information about each of her treatment options, she believes that this information is simply irrelevant to anything that she cares about.

Surely, if Andrea were to make a treatment choice in this scenario, we would not want to say that she was competent to do so, despite the fact that she meets the MCA criteria. The reason for this, I suggest, is that Andrea is unable to engage in rational deliberation about what to do, because she is unable to regard herself as having self-interested reasons to pursue her own well-being, given her illness.³⁰

In the previous chapter, I suggested that we ought to supplement the standard view's account of the voluntariness element of informed consent with a rationality condition, according to which patients must endorse the motivating desire underlying their choice with a personally authorized preference. To do this, a patient must be able to recognise that they have self-interested reasons to want certain things, and they must be able to use the information provided to them to decide what course of action to pursue in the light of both descriptive facts and their own values. Yet, as the above conditions show, the MCA fails to acknowledge the role of the patient's values in the process of making a decision. As such, I claim that we ought to understand the ability of 'weighing' information that the MCA refers to in a manner that corresponds to the rationality condition defended in the previous chapter. To have the ability to 'weigh'

³⁰For discussion of similar cases and the MCA see Rudnick, "Depression and Competence to Refuse Psychiatric Treatment."

information in one's decision-making process is to have the ability to make a decision in accordance with one's personally authorized preferences.

However, this claim may raise two worries. The first is how physicians are to ascertain whether their patient is making their treatment decision in this manner, and relatedly, whether there are any limits to what a competent agent can value. The second is whether this would lead to a standard of competence that is beyond the reach of the majority of patients. I shall consider each in turn.

The first thing to acknowledge in response to the first worry is that it would be somewhat problematic if the *content* of a patient's decision were the sole basis upon which the physician assesses whether the patient is making a rationally warranted decision. As I suggested in reference to the example of Joe in the previous chapter, the values of the physician and the values of the patient may diverge.³¹ In fact, to infer the competence of a patient solely from the *content* of their decision seems to rely on a substantive account of autonomy, rather than a proceduralist account of the sort that I have defended in this thesis.

Nevertheless, although a patient's disagreement with their physician about what treatment they ought to receive is not a sufficient ground for a judgement of incompetence, it seems plausible to claim that such disagreement may give the physician grounds for suspicion, assuming that the physician has made their recommendation in light of what they believe their patient has impersonal self-interested reason to want. In view of the account of autonomy that I have defended, if a patient makes a treatment decision that is in accordance with a personally authorized preference, they will be able to justify that decision by appeal to what they understand

³¹ Chapter Seven, 201-202.

to be the reason-implicating facts about their treatment options, and its coherence with their other evaluative judgements and corresponding preferences. Accordingly, in situations where there is disagreement between the physician and their patient about the best treatment option, I submit that it is not only appropriate for the physician to ask their patient to explain the reasons underlying their decision, but in fact necessary for establishing that the decision was made in a competent fashion.

However, even if a patient is able to give reasons for her treatment decision, hard cases arise when the patient's weighting of the value of certain treatment outcomes seems to be irrational, in the sense that a patient seems to be failing to respond to what she has a strongly decisive reason to do, all things considered. Recall my example of the woman who will die from an allergic reaction to a wasp sting if she does not receive an injection that she knows will be mildly painful;³² suppose she were to refuse the injection on the basis that she is aware that it will cause her a mild pain. We might also consider Jehovah's Witnesses who refuse life-saving blood transfusions on the basis of their religious beliefs,³³ and the example of Joe that I introduced in the previous chapter, who puts his life at risk in order to avoid a negligible risk of paraplegia inherent to the medically indicated treatment.³⁴

From a third party perspective, it seems that the subjects in each of these cases have a strongly decisive reason to do something that they opt not to do. However, on the account of autonomy that I have defended, we should not infer an agent's autonomy with regards to a particular decision from the content of that decision; rather, we must

³² Chapter Seven, 219.

³³ See Bock, "Jehovah's Witnesses and Autonomy"; Savulescu and Momeyer, "Should Informed Consent Be Based on Rational Beliefs?".

³⁴ Chapter Seven, 201-202.

assess whether they sustain the motivating desire that underlies their decision in accordance with a personally authorized preference. In light of this view, and my defence of the liberal rationalist model in the previous chapter, I claim that physicians can be justified in investigating the reasons upon which the patient bases their treatment decision (and their weighting of different reasons), as well as drawing their attention to other salient reason-implying facts; furthermore, they may also seek to disabuse their patient of any irrational beliefs that might undergird their treatment decision (for example, the irrational belief that an injection will be torturously painful). However, if after all this, the physician is convinced that the patient has reached her decision on the basis of her own evaluative judgements, and not as the result of any irrational beliefs, then, I submit, the patient is competent to make her decision, and the fact that her decision might conflict with what others might regard as simply ‘good sense’ is inconsequential to whether that decision should be respected.

Part of the justification for this view is that third parties lack epistemic access to other agents’ own assessment of the comparative *strength* of certain reasons, and the truths regarding the comparative strength of our self-interested are imprecise. In the cases that I am considering here, my suggestion is that whilst physicians may advocate the value and pursuit of health and the reasons that patients have to pursue outcomes related to this good, they may not permissibly override the autonomous choices of patients, since they lack epistemic access to the comparative strength that the agent attributes to the reasons they have to promote their health, and the reasons that they have to pursue other values.

To conclude this chapter, I shall respond to the most pressing objection to the account of autonomy and informed consent that I have defended here, namely the

objection that the rationalist account of autonomy and informed consent that I have defended makes the standards of competence too high. As I pointed out in the previous chapter, Onora O'Neill, and Faden and Beauchamp have objected to the sort of account of autonomy and informed consent that I have defended here for this reason. For instance, Faden and Beauchamp claim:

If conscious, reflective identification with one's motivation were made a necessary condition of autonomous action, a great many intentional, understood, uncontrolled actions that are autonomous in our theory would be rendered non-autonomous.³⁵

The first thing to acknowledge about this objection is that Faden and Beauchamp (and Manson and O'Neill) understand it to pertain to all of the procedural theories of reflective autonomy that I surveyed in the first two chapters. This observation alone might seem to render the objection implausible. To see why, reconsider Frankfurt's view of autonomy; on Frankfurt's view, autonomy requires that one identifies with one's first order motivating desire with a second order volition. Crucially for Frankfurt, human beings can be distinguished from other creatures by virtue of the fact that they alone are able to form second order desires.³⁶ Accordingly, far from being elitist, Frankfurt might claim that the standards set in his theory of reflective autonomy are simply the standards for how we assess personhood.

However, it might be claimed that rationalist theories of autonomy of the sort that I have defended are perhaps more vulnerable to this objection. For instance,

³⁵ Faden and Beauchamp, *A History and Theory of Informed Consent*, 264.

³⁶ Frankfurt, "Freedom of the Will and the Concept of a Person", 6.

Christman and Hyun, who both endorse historical theories of reflective autonomy, have argued that rationalist theories of autonomy make autonomy too demanding.³⁷

Christman writes that:

. . . . the property of autonomy must not collapse into the property of ‘reasonable person’, where the idea of being self-governing is indistinguishable from the idea of being, simply, smart.³⁸

There are several things that should be said in response to this objection. First, I take issue with Christman’s assimilation of rationality and ‘smartness’; one need not be ‘smart’ in order to be rational. On the theory that I have developed here, agents need only be able to pursue the outcome of their desire on the basis of their belief that the outcome is something that they have reason to pursue. I do not see why this capacity is intellectually demanding; on the contrary, we might make a similar claim to Frankfurt here, and suggest that having this capacity just seems to be part of what it is to be a person. As Parfit claims in the very first sentence of *On What Matters*, humans are “ . . . the type of animal that can both understand and respond to reasons”.³⁹ Similarly, in his defence against a similar objection, Raz points out that “(t)o want to be rational is to want to be a person”.⁴⁰ As such, like Raz, I am not sure why an appeal to rational reflection should be deemed elitist in any sense; simply being rational does not imply

³⁷ Christman, “Autonomy and Personal History,” 14; Hyun, “Authentic Values and Individual Autonomy”; Hill, *Autonomy and Self-Respect*, 49.

³⁸ Christman, “Autonomy and Personal History,” 14.

³⁹ Parfit, *On What Matters*, 31.

⁴⁰ Raz, *Engaging Reason*, 18.

that one is 'smart'. To suppose otherwise is to conflate the separate concepts of rationality and intelligence.⁴¹

Perhaps part of the explanation of why rationalist theories of autonomy are deemed elitist is that critics assume that these theories are *substantively* rational rather than *procedurally* rational. This worry is particularly pertinent to the context of informed consent, since such a view of autonomy would make it all too easy for paternalism to sneak in through the back door; on such a view, physicians could readily claim that a patient lacked autonomy simply because they had failed to act in accordance with what the physician believed to be the most substantively 'rational' course of action. However, it should be clear that the theory that I have defended here is not susceptible to this charge. On the theory that I have defended, agents may act on the basis of their beliefs about which facts that provide them with either personal or impersonal self-interested reasons. This is crucial, since what an agent has *personal* self-interested reason to do can vary from person to person; moreover, rational agents may differ with regards to the weight that they place on different impersonal reasons.

The main worry undergirding the elitist objection in the context of biomedical ethics is that incorporating a rationality condition into one's theory of autonomy and informed consent will serve to drastically increase the number of patients who will lack competence; it might be claimed that patients who are facing an emotionally draining medical decision are unlikely to consider reasons for their choice. Against this view, we might first acknowledge that the MCA *already* seems to hint at incorporating some sort of rationality condition for competence, in so far as it claims that that competent patients must be able to 'weigh' material information as part of their decision making

⁴¹ See Baron, *Rationality and Intelligence* for an account of how the two differ.

process; I have already argued that we should understand this condition to pertain to an evaluative weighting. However, even if this were not the case, the worry itself is, I believe, over-stated.

First, the fact that a patient has made an irrational decision does not entail that they must be incompetent, or that their wishes should be ignored on my view. This is because physicians can take steps to help their patients make more rational decisions in the ways that I considered when defending the liberal rationalist model of the doctor-patient relationship in the previous chapter. On this account, it is part of the doctor's role to try and help the patient to exercise the rational capacities that are necessary for making a competent decision, by engaging them in a discussion about their evaluative judgments.

Second, we might also observe that far from impeding an agent's ability to consider the reasons for their practical choices, facing a serious medical condition could instead *promote* this behaviour. Since the moral stakes in medicine are often high, patients may be prompted to reconsider their own values and to decide what they want to prioritise in their own life. Consider the following:

Erica has been told that tests have found a benign tumour near her ovaries. Doctors can operate immediately, and remove the tumour; however, this will render Erica infertile. Alternatively, Erica can wait for a few years (enough time to have a child) but run the risk of the tumour developing into a malignant one.

It seems plausible to claim that Erica's predicament is likely to prompt her to reflect upon her values, and consider the extent to which she regards having children as reason-implying.

Despite these points, I cannot deny that there will be some patients who lack competence on the view of autonomy that I have defended here, but who would not be found incompetent on standard theories of informed consent. I have in mind here those patients who suffer from conditions that render them unable to make treatment decisions in accordance with what they believe they have reason to do in light of their own evaluative judgements. For instance, recall my example of Jane the bulimia sufferer from the introduction,⁴² and my example of Apathetic Andrea who is unable to make a treatment decision in accordance with her values because she simply does not have any.

Whilst such patients would lack competence on the theory of informed consent that I have developed over the course of the last two chapters, I do not believe that this is a flaw in my theory. On the contrary, I believe that it is a flaw of the standard view that it finds such patients competent to make their treatment decisions, and regards their choices as autonomous. Whilst these patients are able to express a 'choice', it is one that is unconnected to what they themselves believe they have reason to do in light of their own values.

In fact, in stark contrast to the elitist objection, consideration of certain sufferers of anorexia nervosa suggests that the theory of autonomy that I have defended here might be deemed *too* permissive. Consider Craigie's description of the regret that

⁴² Introduction, 9.

certain sufferers of anorexia nervosa feel following their recovery. Craigie argues that self-reports suggest that the source of their regret is:

. . . not that they failed to *do what they wanted to do* – on the contrary they pursued the goal of thinness very effectively. It seems more likely that the regret these people express has its source primarily in *what they valued*.⁴³

The sufferers of anorexia that Craigie describes here decide to refuse food in a manner that is rational *in light of their own values*. Certain sufferers of anorexia nervosa believe that they have good reasons to engage in morbid weight loss; not only that, they also believe that those reasons are stronger than their reasons to ensure their continued survival by living at a healthier weight.

These sorts of case are hugely complex, and I cannot hope to address them adequately here. However, I shall close by highlighting the following general point. It seems that some commentators, including Craigie herself (amongst others),⁴⁴ would want to claim that the sufferers of anorexia I am considering here are incompetent to make their own treatment decisions. The strength of the intuition underlying this view represents an important challenge for procedural theory of autonomy, since the agents in question seem to make their decisions in an autonomous manner, according to most procedural theories.

⁴³ Craigie, "Competence, Practical Rationality and What a Patient Values," 331.

⁴⁴ Ibid. ; See also Tan et al., "Competence to Make Treatment Decisions in Anorexia Nervosa."

However, I believe that the account of autonomy and informed consent that I have developed in this thesis can provide the foundation for a way in which a procedural theory of autonomy could at least offer a deeper analysis of these complex cases. In chapter two, I considered the example of a clinically depressed patient who (I argued) lacks autonomy because his desire to commit suicide is based upon an epistemically irrational evaluative belief that he seems to hold in a compulsive fashion.⁴⁵ I suggest that some of the sufferers of anorexia nervosa that Craigie considers may be in an analogous position. Whether or not these patients will be competent on my theory will depend upon whether their beliefs about the good that undergird their self-harming preferences meet the minimum threshold of epistemic rationality discussed in chapter two, and whether they are weighing their reasons here rationally.

I have already discussed the epistemic limitations that third parties face when attempting to establish the weight that other agents place on certain good, and the weight that it would be appropriate for them to place on those goods. Moreover, since the degree of rationality that sufferers of anorexia nervosa exhibit in such examples is likely to differ across cases, I cannot offer a full account of the competence of all such patients here. However, the thought that these reflections capture, and one that any adequate investigation into the matter should accommodate, is that a plausible procedural theory of autonomy in bioethics should be able to account for the fact that there is an important difference between claiming that agents whose treatment decisions do not reflect their own evaluative judgements are incompetent, and claiming that the sort of sufferers of anorexia nervosa that Craigie describes are incompetent.

⁴⁵ Chapter Two, 58-59.

Whilst I do not hope to have solved the difficulties raised by these problematic cases, they illustrate the point that the standard view of autonomy and informed consent is flawed in so far as it fails to adequately engage with the hard cases that I have considered here. The question of whether these patients should be deemed competent cannot simply be a matter of whether they are acting intentionally, understand material information, and are not subject to external controlling influences. Furthermore, contrary to the MCA, it will not do to say that such sufferers must be able to simply ‘weigh’ material information in their decision making process. In order to adequately assess competence here, we must delve deeper into the decision-making procedure of such individuals, the nature of their beliefs, and the reasons that they understand themselves to have. Whilst I do not claim to have offered a detailed investigation of this specific issue here, I believe that the theory I have developed in this thesis at least has the conceptual apparatus to engage with these problems.

In the final chapter, I shall offer some concluding remarks about the arguments that I have made in this thesis, and identify some areas in contemporary bioethics that I have not considered, but in which my views on autonomy could possibly bring fresh insight.

Conclusion

In the introduction to this thesis, I outlined my intention to provide an account of personal autonomy that can usefully be applied to issues in contemporary bioethics. I understood the concept of autonomy to denote a particular capacity to which we seem to attribute prudential value in bioethical contexts, namely, a person's capacity to both:

- (i) Make decisions about what to do in accordance with their own desires and values

And

- (ii) To act on the basis of those decisions.

I pointed out in the introduction that the standard view of autonomy in bioethics advocated by Beauchamp and Childress fails to give an adequate account of what it is for an agent to make their decisions about what to do in accordance with their own desires and values. Part of the reason that the standard view is inadequate is that it does not attend to what I termed the 'reflective dimension' of autonomy, a dimension that pertains to the reflection that agents must carry out on their motivating desires in order to be autonomous with respect to them. In the first two chapters, I analysed a range of philosophical theories that attempt to explicate this dimension of autonomy, and raised a number of objections to these theories.

In chapter two, I defended a rationalist account of the reflective dimension of autonomy, according to which an agent is autonomous when they act on a first order desire because they have a ‘personally authorized preference’ for that desire to be effective. In order to remedy flaws with Ekstrom’s version of this sort of theory, I supplemented her theory with a Parfitian account of rational desires and epistemic rationality. Following an analysis of Parfit’s claims, I suggested that we should understand a preference to be a desire that an agent sustains on the basis of a (non-irrational) belief, whose truth would make the outcome of their first-order desire good in a reason-implying sense; I also explained that this account is compatible with both subjective and objective accounts of the good. Furthermore, I suggested that we should understand the ‘acceptances’ that partly constitute the agent’s character system (in conjunction with their preferences) to be those beliefs that they sustain in a manner that is not epistemically *irrational*.

In chapter three, I turned to the second aspect of the understanding of autonomy that I invoked in the introduction, namely, the agent’s ability to act on the basis of their decisions. I argued that there are three reasons to include this practical dimension into an overall theory of autonomy in bioethics. First, I pointed out that this dimension of autonomy undergirds the negative obligation incorporated into the principle of respect for autonomy in bioethics that enjoins us to refrain from preventing the autonomous acts of others. Second, I suggested that incorporating this dimension into our overall theory of autonomy allows us to make sense of the high prudential value that we place on autonomy. Finally, I claimed that if we fail to acknowledge the practical dimension of autonomy in our overall theory of autonomy, it is not clear how we can account for the way in which our beliefs about what we are free to do can have crucial effects on our choices.

I argued that in order for an agent to be practically autonomous, they must, at the point of action, have the positive and negative freedoms that are necessary for them to act effectively in pursuit of their ends, taking care to point out that this does not entail a success condition on autonomy. I suggested that although the particular freedoms that agents require to be practically autonomous will often depend on the goals that they want to pursue, agents will lack practical autonomy if they are informationally cut off from achieving their ends. I went on to argue that the agent's beliefs about what they are free to do at the point of decision can have important ramifications for their reflective autonomy with respect to the practical decisions that they make on the basis of these beliefs. This, I suggested, represents an important relationship between the reflective and practical dimensions of autonomy. In chapter four, I argued that acknowledging this relationship between the two dimensions of autonomy is crucial for understanding how coercion can serve to undermine an agent's autonomy.

The theory of autonomy that I developed in the first four chapters of this thesis is a novel account of autonomy in bioethics, and it is, I believe, an important addition to the literature on the subject. First, I have provided a detailed defence of a rationalist approach to the reflective dimension of autonomy that enables my account to avoid the flaws of the standard account of autonomy in bioethics. It also avoids the problems in alternative philosophical accounts of this dimension of autonomy that are often invoked in bioethical contexts. Second, by explicitly incorporating a practical dimension in my account of autonomy, my understanding of autonomy is not only congruent with the way in which we use the concept of autonomy in contemporary bioethics, it is also able to offer a deeper explanation of why deception and coercion can undermine autonomy than those that are offered by other competing accounts of autonomy.

Furthermore, by incorporating this dimension of autonomy into my overall theory, my account reflects the prudential value we attribute to autonomy in bioethical contexts. In chapter five, I defended the view that we should understand autonomy as having both instrumental and final value, and that when an agent's local autonomy comes into conflict with their global autonomy, they will often have reasons to give precedence to their global autonomy. I also examined other values that autonomy may come into conflict with in contemporary bioethics. Notably, I concluded that if one incorporates the claim that autonomy bears final value into one's theory of well-being, then we should reframe our understanding of the way in which the principle of autonomy and the principle of beneficence can come into conflict; this, I believe, is a novel and important claim.

Furthermore, by recognising the relationship between the reflective and practical dimensions of autonomy, I was able to make a novel contribution to the burgeoning debate regarding the enhancement of autonomy through biotechnological means. Bioethicists have previously discussed ways in which we might be able to enhance the practical dimension of an agent's autonomy by using biotechnologies to enhance their ability to effectuate their plans. More recently, they have also begun to discuss the prospect of enhancing the reflective dimension of autonomy by enhancing various cognitive capacities. In chapter six, I argued that by overlooking the relationship between these two dimensions of autonomy, the discussion to this point has overlooked the way in which making enhancement technologies widely available could indirectly enhance autonomy.

In chapters seven and eight, I pointed out that my account of autonomy has a number of important implications for the doctrine of informed consent. In chapter

seven, I argued that we ought to supplement the standard view of informed consent with a rationality condition based on my theory of autonomy. I suggested that such a condition would enable us to explain why agents who feel alienated from their motivating desire in medical contexts (such as certain patients suffering from mental disorders such as bulimia) might lack autonomy with respect to their decisions. I also claimed that this condition is compatible with the claim that physicians might be able to influence their patient's decision-making in certain ways, without undermining autonomy. As such, I argued against the shared-decision making model of the doctor-patient relationship, and advocated a liberal rationalist model that allows physicians to go beyond the role of a mere 'fact-provider', and to engage with their patients in rational discussion about their evaluative judgements.

Finally, in chapter eight, I considered the information that patients need to understand in order to be autonomous with respect to their treatment decisions. I argued against Faden and Beauchamp's purely subjective criterion of material information, and defended the view that patients need to understand information pertaining to reasoning facts about their condition or treatment. However, I also noted some practical difficulties with implicating this standard, and concluded that the best practical solution is for physicians to use the reasonable patient standard as a starting point of the physician's disclosure, in conjunction with a duty to disclose the fact that patients can ask for any other information, and to engage with the patient to find out more about *their* values. I concluded by considering the implications of my account for the concept of competence to consent, and defended my account from the charge that it makes the conditions of competence too demanding.

In a project of this nature, it is perhaps unavoidable that there are a number of issues that I have not been able to adequately address. From a theoretical perspective, I lacked the space to fully explore the question of what epistemic rationality requires; indeed, such an exploration would require its own project. However, on my theory of autonomy, the answer to this theoretical question will have important practical ramifications, especially for our understanding of patients who seem to make medical decisions on the basis of beliefs of a questionable epistemic status.

Furthermore, in this thesis I was unable to provide a full account of how autonomy should be incorporated into an adequate theory of well-being. Whilst I suggested that a hybrid approach to well-being represents a promising way in which we could incorporate the claim that autonomy has final value into a theory of well-being, I believe that this idea could be developed further.

I also believe that the account of autonomy that I have developed here could be fruitfully applied to a number of other practical issues in contemporary bioethics that I did not fully consider here. In other published work, I have used the account of autonomy that I have developed in this thesis to address questions pertaining to the moral permissibility of genetic enhancements,¹ deceptive placebo use,² and coercive paternalism.³ As I briefly intimated at the beginning of chapter four, I also believe that the account of coercion and autonomy that I developed in that chapter could usefully be applied to topical bioethical questions concerning whether certain practices such as, *inter alia*, providing financial compensation for organ donation and offering sex offenders reduced sentences in return for undergoing chemical castration, should be

¹ Pugh, "Autonomy, Natality and Freedom."

² Pugh, "Ravines and Sugar Pills: Defending Deceptive Placebo Use."

³ Pugh, "Coercive Paternalism and Back-Door Perfectionism."

understood to be coercive. More generally, I believe that many of the arguments that I developed in chapters seven and eight could form the basis of a new way of analysing the question of competence to consent, particularly in relation to patients suffering from mental disorders such as anorexia nervosa and clinical depression. There are, of course, many other topics in contemporary bioethics in which autonomy related concerns are salient and that I have not touched upon in this thesis. However, I believe that the theory of autonomy that I have developed here may readily be applied to the many other practical issues in contemporary bioethics in which autonomy related concerns are raised.

In the preceding paragraphs, I have briefly delineated some of the main conclusions that I have drawn in individual chapters of this thesis. By way of closing, I wish to suggest that my project as a whole is an important addition to the literature on autonomy in philosophy and bioethics in so far as it represents an attempt to bridge the gap between philosophical discussions of the concept of autonomy, and the way in which the concept is invoked in bioethics. I have suggested a number of ways in which our bioethical discussions can be enriched by a philosophically informed understanding of autonomy. However, I also believe that the way in which the concept of autonomy is invoked in contemporary bioethical issues suggests some important insights for our philosophical understanding of autonomy. In particular, in view of my arguments in chapters three and four, I believe that my claims regarding the importance of acknowledging both what I have called the reflective and practical dimensions of autonomy in bioethical discussion should also be extended to our philosophical discussion of the concept of autonomy more generally. This, I believe, makes this thesis an important addition to the philosophical literature on autonomy, as well as to the literature on practical issues in contemporary bioethics.

References

- Adams, Robert. *Finite and Infinite Goods : A Framework for Ethics*. New York ; Oxford: Oxford University Press, 1999.
- Anderson, Scott. "Coercion." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, 2011 (Last accessed September 2014).
<http://plato.stanford.edu/archives/win2011/entries/coercion/>.
- Archard, David. "Informed Consent: Autonomy and Self-Ownership." *Journal of Applied Philosophy* 25, no. 1 (2008). doi:10.1111/j.1468-5930.2008.00394.x.
- Aristotle, *The Nicomachean Ethics*. Dordrecht: Reidel, 1975.
- Arpaly, Nomy. "Responsibility, Applied Ethics, and Complex Autonomy Theories." In *Personal Autonomy New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, edited by James Stacey Taylor, n.d.
- — —. *Unprincipled Virtue : An Inquiry into Moral Agency*. New York ; Oxford: Oxford University Press, 2003.
- Barazzetti, Gaia. "Looking for the Fountain of Youth: Scientific, Ethical, and Social Issues in the Extension of Human Lifespan." In *Enhancing Human Capacities*, edited by Julian Savulescu, Ruud ter Meulen, and Guy Kahane, 335–50. Wiley-Blackwell, 2011.
- Baron, Jonathan. *Rationality and Intelligence*. Cambridge Cambridgeshire: Cambridge University Press ; New York: Cambridge University Press, 1985.
- Bayles, Michael. "A Concept of Coercion." In *Nomos XIV: Coercion*, edited by James Pennock and John Chapman. Chicago: Aldine-Atherton, 1972.
- Beauchamp, Tom L., and James F. Childress. *Principles of Biomedical Ethics*. New York ; Oxford: Oxford University Press, 1979.
- — —. *Principles of Biomedical Ethics*. 5th ed.. Oxford: Oxford University Press, 2001.
- Beck, Aaron T. *Depression: Clinical, Experimental, and Theoretical Aspects*. University of Pennsylvania Press, 1967.
- Benson, Paul. "Freedom and Value." *The Journal of Philosophy* 84, no. 9 (September 1987): 465. doi:10.2307/2027060.
- Berghmans, Ron, Ruud ter Meulen, Andrea Malizia, and Rein Vos. "Scientific, Ethical, and Social Issues in Mood Enhancement." In *Enhancing Human Capacities*,

edited by Julian Savulescu, Ruud ter Meulen, and Guy Kahane. Wiley-Blackwell, 2011.

Berlin, Isaiah. "John Stuart Mill and the Ends of Life." In *On Liberty in Focus*, edited by John Gray and G. W. Smith. London: Routledge, 1991.

— — —. "Two Concepts of Liberty." In *The Liberty Reader*, edited by David Miller. Edinburgh University Press, 2006.

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

Bock, Gregory L. "Jehovah's Witnesses and Autonomy: Honouring the Refusal of Blood Transfusions." *Journal of Medical Ethics* 38, no. 11 (November 2012): 652–56.

Bok, Sissela. *Lying: Moral Choice in Public and Private Life*. New York: Vintage Books, 1989.

Bostrom, Nick. "Human Genetic Enhancements: A Transhumanist Perspective." *The Journal of Value Inquiry* 37, no. 4 (2003): 493–506.

— — —. "In Defense of Posthuman Dignity." *Bioethics* 19, no. 3 (2005). doi:10.1111/j.1467-8519.2005.00437.x.

Bostrom, Nick, and Toby Ord. "The Reversal Test: Eliminating Status Quo Bias in Applied Ethics." *Ethics* 116, no. 4 (2006): 656–79.

Bratman, Michael. "Planning Agency, Autonomous Agency." In *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, edited by James Stacey Taylor, (2005)

— — —. "Identification, Decision, and Treating as a Reason." *Philosophical Topics* 24, no. 2 (1996): 1–18.

Brewin, Thurstan. "Telling the Truth." *The Lancet* 343, no. 8911 (June 11, 1994): 1512. doi:10.1016/S0140-6736(94)92626-3.

Brock, Dan W. *Life and Death: Philosophical Essays in Biomedical Ethics*. Cambridge: Cambridge University Press, 1993.

Buchanan, Allen. "The Physician's Knowledge and the Patient's Best Interest." In *Ethics, Trust, and the Professions: Philosophical and Cultural Aspects*, edited by Edmund Pellegrino, Robert Veatch, and John Langan. Georgetown University Press, 1991.

Buchanan, Allen E., Dan W. Brock, Norman Daniels, Daniel Wikler, and Elliott Sober. *From Chance to Choice: Genetics and Justice*. Cambridge: Cambridge University Press, 2000.

- Canterbury v. Spence (464 F.2d 772), United States Court of Appeals for the District of Columbia Circuit, 1972.
<http://www.lawandbioethics.com/demo/Main/LegalResources/C5/Canterbury.htm>
 m Accessed July 16, 2014.
- Carr, Craig L. "Coercion and Freedom." *American Philosophical Quarterly* 25, no. 1 (1988): 59–67.
- Carter, Williamr. "How to Change Your Mind." *Canadian Journal of Philosophy* 19, no. 1 (1989). doi:10.1080/00455091.1989.10716464.
- Christman, John. "Autonomy and Personal History." *Canadian Journal of Philosophy* 21, no. 1 (1991). doi:10.1080/00455091.1991.10717234.
- Cobbs v. Grant (8 Cal. 3d 229), Supreme Court of California, 1972.
<http://www.lawandbioethics.com/demo/Main/LegalResources/C5/Cobbs.htm>
 Accessed July 16, 2014.
- Colburn, Ben. "Autonomy and Adaptive Preferences." *Utilitas* 23, no. 1 (2011). doi:10.1017/S0953820810000440.
- Conly, Sarah. *Against Autonomy : Justifying Coercive Paternalism*. Cambridge: Cambridge University Press, 2013.
- Craigie, Jillian. "Competence, Practical Rationality and What a Patient Values." *Bioethics* 25, no. 6 (2011): 326–33. doi:10.1111/j.1467-8519.2009.01793.x.
- Crisp, Roger. *Reasons and the Good*. Oxford University Press, 2006.
- Darwall, Stephen. "The Value of Autonomy and Autonomy of the Will." *Ethics* 116, no. 2 (January 2006): 263–84. doi:10.1086/498461.
- — —. *Welfare and Rational Care*. Princeton, NJ, USA: Princeton University Press, 2004.
- Davis, Dena S. "The Parental Investment Factor and the Child's Right to an Open Future." *Hastings Center Report* 39, no. 2 (2009).
- DeGrazia, David. *Human Identity and Bioethics*. Cambridge: Cambridge University Press, 2005.
- Dennett, D. *Consciousness Explained*. London: Allen Lane, 1992.
- Doerflinger, Richard. "Assisted Suicide: Pro-Choice or Anti-Life?" *The Hastings Center Report* 19, no. 1 (1989): 16–19.
- Doorn, Neelke. "Mental Competence or Capacity to Form a Will: An Anthropological Approach." *Philosophy, Psychiatry, & Psychology* 18, no. 2 (2011): 135–45. doi:10.1353/ppp.2011.0025.

- Douglas, Thomas. "Moral Enhancement." *Journal of Applied Philosophy* 25, no. 3 (2008). doi:10.1111/j.1468-5930.2008.00412.x.
- Dworkin, Gerald. "Acting Freely." *Noûs* 4, no. 4 (November 1970): 367. doi:10.2307/2214680.
- — —. "Autonomy and Behavior Control." *The Hastings Center Report* 6, no. 1 (1976): 23–28.
- — —. "Is More Choice Better than Less?" *Midwest Studies In Philosophy* 7, no. 1 (September 1, 1982): 47–61. doi:10.1111/j.1475-4975.1982.tb00083.x.
- — —. "Paternalism." *Monist* 56, no. 1 (1972): 64–84. doi:10.5840/monist197256119.
- — —. "Paternalism (SEP Entry)," 2010. <http://plato.stanford.edu/entries/paternalism/#ConIss>.
- — —. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press, 1988.
- Ekstrom, Laura Waddell. "A Coherence Theory of Autonomy." *Philosophy and Phenomenological Research* 53, no. 3 (1993): 599–616.
- Elliott, Carl. *Better than Well : American Medicine Meets the American Dream*. New York ; London: WWNorton, 2003.
- Elster, Jon. *Sour Grapes : Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press, 1983.
- Eyal, Nir. "Using Informed Consent to Save Trust." *Journal of Medical Ethics* 40, no. 7 (July 1, 2014): 437–44. doi:10.1136/medethics-2012-100490.
- Faden, Ruth, and Tom Beauchamp. *A History and Theory of Informed Consent*. New York: Oxford University Press, 1986.
- Feinberg, Joel. *Freedom and Fulfillment : Philosophical Essays*. Princeton: Princeton University Press, 1992.
- — —. *The Moral Limits of the Criminal Law Volume 3: Harm to Self [electronic Resource]*. New York: Oxford University Press, 1989.
- Feldman, Fred. *Pleasure and the Good Life : Concerning the Nature, Varieties and Plausibility of Hedonism*. Oxford: Clarendon Press, 2004.
- — —. "What Is the Rational Care Theory of Welfare?" *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 130, no. 3 (September 1, 2006): 585–601.
- Finnis, John. *Natural Law and Natural Rights [electronic Resource]*. 2nd ed.. Oxford ; New York: Oxford University Press, 2011.

- Fischer, John Martin. "Recent Work on Moral Responsibility." *Ethics* 110, no. 1 (1999): 93–139.
- Fischer, John Martin, and Mark Ravizza. *Responsibility and Control : A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998.
- Foddy, Bennett. "A Duty to Deceive: Placebos in Clinical Practice." *The American Journal of Bioethics* 9, no. 12 (2009). doi:10.1080/15265160903318350.
- Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66, no. 23 (1969): 829–39.
- — —. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68, no. 1 (1971): 5–20.
- — —. *Necessity, Volition, and Love*. Cambridge: Cambridge University Press, 1999.
- — —. *The Importance of What We Care about : Philosophical Essays*. Cambridge: Cambridge University Press, 1988.
- Gillon, R. "Ethics Needs Principles--Four Can Encompass the Rest--and Respect for Autonomy Should Be 'First among Equals.'" *Journal of Medical Ethics* 29, no. 5 (October 2003): 307–12. doi:10.1136/jme.29.5.307.
- Glover, Jonathan. *Causing Death and Saving Lives*. Harmondsworth: Penguin, 1977.
- Goodman, Rob. "Cognitive Enhancement, Cheating, and Accomplishment." *Kennedy Institute of Ethics Journal* 20, no. 2 (June 2010): 145–60.
- Grice, H. P. *Studies in the Way of Words*. Cambridge, Mass: Harvard University Press, 1989.
- Griffin, James. "Darwall on Welfare as Rational Care." *Utilitas* 18, no. 4 (2006). doi:10.1017/S0953820806002202.
- — —. *Well-Being : Its Meaning, Measurement and Moral Importance*. Oxford: Clarendon Press, 1986.
- Grisso, Thomas, and Paul S. Appelbaum. *Assessing Competence to Consent to Treatment : A Guide for Physicians and Other Health Professionals*. New York ; Oxford: Oxford University Press, 1998.
- Habermas, Jürgen. *The Future of Human Nature*. Cambridge: Polity, 2003.
- Hart, H. L. A. *Law, Liberty and Morality*. London: Oxford University Press, 1963.
- Haworth, Lawrence. *Autonomy : An Essay in Philosophical Psychology and Ethics*. New Haven: Yale University Press, 1986.

- Heathwood, C. "Review of Rational Care and Welfare." *Australasian Journal of Philosophy* 81, no. 4 (2003): 615–17.
- Hill, Thomas E. *Autonomy and Self-Respect*. Cambridge: Cambridge University Press, 1991.
- Hughes, P. "Ambivalence, Autonomy, and Organ Sales." *Southern Journal of Philosophy*, 44, no. 2 (2006).
- Hume, David. *A Treatise of Human Nature*. 2nd ed. / with text revised and variant readings by P.H. Nidditch.. Oxford: Clarendon Press, 1978.
- Hurka, Thomas. "Why Value Autonomy?", *Social Theory and Practice* 13, no. 3 (1987): 361–82. doi:10.5840/soctheorpract198713316.
- Huxley, Aldous. *Brave New World*. London: Vintage, 2007.
- Hyun, Insoo. "Authentic Values and Individual Autonomy." *The Journal of Value Inquiry* 35, no. 2 (June 1, 2001): 195–208. doi:10.1023/A:1010347121641.
- Ingelfinger, F J. "Informed (but Uneducated) Consent." *The New England Journal of Medicine* 287, no. 9 (August 31, 1972): 465–66. doi:10.1056/NEJM197208312870912.
- Jackson, Jennifer. "Telling the Truth." *Journal of Medical Ethics* 17, no. 1 (1991): 5–9.
- Juengst, Eric. "What Does Enhancement Mean?" In *Enhancing Human Traits: Ethical and Social Implications*, edited by Erik Parens. Washington: Georgetown University Press, 1998.
- Juth, Niklas. "Enhancement, Autonomy, and Authenticity." In *Enhancing Human Capabilities*, edited by J Savulescu, Guy Kahane, and Ruud ter Meulen, 2011.
- Kenny, Anthony. *The Self*. Milwaukee: Marquette University Press, 1988.
- Kihlbom, U. "Autonomy and Negatively Informed Consent." *Journal of Medical Ethics* 34, no. 3 (March 2008): 146–49. doi:10.1136/jme.2007.020503.
- Killmister, Suzy. "Autonomy and False Beliefs." *Philosophical Studies* 164, no. 2 (June 1, 2013): 513–31. doi:10.1007/s11098-012-9864-0.
- Kleinig, John. "The Nature of Consent." In *The Ethics of Consent*, edited by Franklin G. Miller and Alan Wertheimer. Oxford University Press, 2010.
- Kolber, Adam J. *A Limited Defense of Clinical Placebo Deception*. Yale Law Policy Review (2007), 26:75-134
- Korsgaard, Christine. "Two Distinctions in Goodness." *The Philosophical Review* 92, no. 2 (1983): 169–95.

- Kreek, Mary Jeanne, David A Nielsen, Eduardo R Butelman, and K Steven LaForge. "Genetic Influences on Impulsivity, Risk Taking, Stress Responsivity and Vulnerability to Drug Abuse and Addiction." *Nature Neuroscience* 8, no. 11 (November 2005): 1450–57. doi:10.1038/nm1583.
- Ladenson, Robert. "Mill's Conception of Individuality." *Social Theory and Practice* 4, no. 2 (1977): 167–82.
- Lamond, Graint. "Coercion, Threats, and the Puzzle of Blackmail." In *Harm and Culpability*, edited by A. P. Simester and A. T. H. Smith, 215–38. Oxford: Clarendon Press, 1996.
- Lehrer, Keith. "Reason and Autonomy." *Social Philosophy and Policy* 20, no. 2 (2003): 177–98.
- Levy, Neil. "Autonomy and Addiction." *Canadian Journal of Philosophy* 36, no. 3 (2006): 427–47.
- . "Enhancing Authenticity." *Journal of Applied Philosophy* 28, no. 3 (2011). doi:10.1111/j.1468-5930.2011.00532.x.
- . "Forced to Be Free? Increasing Patient Autonomy by Constraining It." *Journal of Medical Ethics* 40, no. 5 (May 1, 2014): 293–300. doi:10.1136/medethics-2011-100207.
- Lindley, Richard. *Autonomy*. Basingstoke: Macmillan, 1986.
- Locke, John. *An Essay on Human Understanding*. Ware: Wordsworth Editions, 1998.
- MacCallum Jr, Gerald C. "Negative and Positive Freedom." In *The Liberty Reader*. Edinburgh University Press, 2006.
- Mackenzie, Catriona, and Natalie Stoljar. *Relational Autonomy : Feminist Perspectives on Autonomy, Agency, and the Social Self*. New York ; Oxford: Oxford University Press, 1999.
- Manson, Neil C., and O'Neill, O. *Rethinking Informed Consent in Bioethics*. Cambridge: Cambridge University Press, 2007.
- Mccloskey, H. J. "Coercion: Its Nature and Significance." *Southern Journal of Philosophy* 18, no. 3 (1980). doi:10.1111/j.2041-6962.1980.tb01390.x.
- McGregor, Joan. "Bargaining Advantages and Coercion in the Market." *Philosophy Research Archives* 14 (1988): 23–50. doi:10.5840/pr1988/1989147.
- Mckenna, M. "The Relationship between Autonomous and Morally Responsible Agency." In *Personal Autonomy New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, edited by James Stacey Taylor. Cambridge University Press, n.d.

- McMahan, Jeff. *The Ethics of Killing : Problems at the Margins of Life*. Oxford: Oxford University Press, 2002.
- McMillan, John. "The Kindest Cut? Surgical Castration, Sex Offenders and Coercive Offers." *Journal of Medical Ethics*, May 11, 2013, medethics-2012-101030. doi:10.1136/medethics-2012-101030.
- McNeil, B. J., S. G. Pauker, H. C. Sox, and A. Tversky. "On the Elicitation of Preferences for Alternative Therapies." *The New England Journal of Medicine* 306, no. 21 (May 27, 1982): 1259–62. doi:10.1056/NEJM198205273062103.
- Mele, Alfred R. *Autonomous Agents : From Self-Control to Autonomy*. New York ; Oxford: Oxford University Press, 1995.
- "Mental Capacity Act 2005." Accessed July 16, 2014. <http://www.legislation.gov.uk/ukpga/2005/9>.
- Mill, John Stuart. "On Liberty." In *J.S. Mill, On Liberty, in Focus*, edited by John Gray and G. W. Smith. Routledge, 1991.
- Miller, David. "Introduction." In *The Liberty Reader*, edited by David Miller. Edinburgh University Press, 2006.
- Miller, Franklin G., and Alan Wertheimer. *The Ethics of Consent : Theory and Practice*. Oxford ; New York: Oxford University Press, 2010.
- Mobbs, Olivia, Christelle Crépin, Christelle Thiéry, Alain Golay, and Martial Van der Linden. "Obesity and the Four Facets of Impulsivity." *Patient Education and Counseling* 79, no. 3 (2010): 372–77. doi:10.1016/j.pec.2010.03.003.
- Nagel, Thomas. *The Possibility of Altruism*. Princeton: Princeton University Press, 1970.
- Noggle, Robert. "Autonomy and the Paradox of Self-Creation." In *Personal Autonomy New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, edited by James Stacey Taylor, 2005.
- Nozick, Robert. "Coercion." In *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, edited by Sidney Morgenbesser, Patrick Suppes, and Morton White. New York: St. Martin's Press, 1969.
- Nussbaum, Martha. *Women and Human Development : The Capabilities Approach*. Cambridge: Cambridge University Press, 2000.
- O'Neill, Onora. *Autonomy and Trust in Bioethics : The Gifford Lectures, University of Edinburgh, 2001*. Cambridge: Cambridge University Press, 2002.
- Olson, Eric T. *The Human Animal [electronic Resource] : Personal Identity Without Psychology*. New York: Oxford University Press, 1999.

- Oshana, Marina. "How Much Should We Value Autonomy?" *Social Philosophy and Policy* 20, no. 2 (2003). doi:10.1017/S0265052503202041.
- — — "Personal Autonomy and Society." *Journal of Social Philosophy*, 29, no. 1 (1998). doi:10.1111/j.1467-9833.1998.tb00098.x.
- — — "The Misguided Marriage of Responsibility and Autonomy." *The Journal of Ethics* 6, no. 3 (2002): 261–80.
- Parfit, Derek. *On What Matters*. Oxford: Oxford University Press, 2011.
- — —. *Reasons and Persons*. Oxford: Clarendon Press, 1984.
- Pitkin, Hanna Fenichel. "Are Freedom and Liberty Twins?" *Political Theory* 16, no. 4 (1988): 523–52.
- Plato. *Republic*. Oxford: Oxford University Press, 1993.
- Platts, Mark de Bretton. *Ways of Meaning : An Introduction to a Philosophy of Language*. 2nd ed.. Cambridge, Mass; London: MIT Press, 1997.
- Pugh, Jonathan. "Autonomy, Natality and Freedom: A Liberal Re-Examination of Habermas in the Enhancement Debate." *Bioethics*, 2014, n/a–n/a. doi:10.1111/bioe.12082.
- — —. "Coercive Paternalism and Back-Door Perfectionism." *Journal of Medical Ethics* 40, no. 5 (May 1, 2014): 350–51. doi:10.1136/medethics-2013-101556.
- — —. "Enhancing Autonomy by Reducing Impulsivity: The Case of ADHD." *Neuroethics*, February 25, 2014. doi:10.1007/s12152-014-9202-7.
- — —. "Ravines and Sugar Pills: Defending Deceptive Placebo Use." *Journal of Medicine and Philosophy*, (in press).
- Quill, T E, and H Brody. "Physician Recommendations and Patient Autonomy: Finding a Balance between Physician Power and Patient Choice." *Annals of Internal Medicine* 125, no. 9 (November 1, 1996): 763–69.
- Raz, Joseph. *Engaging Reason : On the Theory of Value and Action*. Oxford: Oxford University Press, 1999.
- — —. *The Morality of Freedom*. Oxford: Clarendon Press, 1986.
- Richards, Janet Radcliffe. *The Ethics of Transplants : Why Careless Thought Costs Lives*. Oxford: Oxford University Press, 2012.

- Robinson v. Bleicher (559 N.W.2d 473), Supreme Court of Nebraska, 1997,
<http://www.lawandbioethics.com/demo/Main/LegalResources/C5/Robinson.htm>
 . Accessed July 16, 2014.
- Rudnick, A. "Depression and Competence to Refuse Psychiatric Treatment." *Journal of Medical Ethics* 28, no. 3 (2002). doi:10.1136/jme.28.3.151.
- Sachs, Benjamin. "Why Coercion Is Wrong When It's Wrong." *Australasian Journal of Philosophy* 91, no. 1 (2013). doi:10.1080/00048402.2011.646280.
- Sandberg, Anders. "Cognition Enhancement: Upgrading the Brain." In *Enhancing Human Capacities*, edited by Julian Savulescu, Guy Kahane, and Ruud ter Meulen. Wiley-Blackwell, 2011.
- Sandel, Michael J. *The Case against Perfection : Ethics in the Age of Genetic Engineering*. Cambridge, Mass: Belknap Press of Harvard University Press, 2007.
- Sartre, Jean-Paul. *Being and Nothingness : An Essay on Phenomenological Ontology*. London: Routledge, 1989.
- Savulescu, Julian. "Rational Non-Interventional Paternalism: Why Doctors Ought to Make Judgments of What Is Best for Their Patients." *Journal of Medical Ethics* 21, no. 6 (1995). doi:10.1136/jme.21.6.327.
- — — "Liberal Rationalism And Medical Decision-making." *Bioethics* 11, no. 2 (1997). doi:10.1111/1467-8519.00049.
- — — "Rational Desires and the Limitation of Life Sustaining Treatment." *Bioethics* 8, no. 3 (1994). doi:10.1111/j.1467-8519.1994.tb00255.x.
- Savulescu, Julian, and R. W. Momeyer. "Should Informed Consent Be Based on Rational Beliefs?" *Journal of Medical Ethics* 23, no. 5 (1997). doi:10.1136/jme.23.5.282.
- Savulescu, Julian, Anders Sandberg, and Guy Kahane. "Well-Being and Enhancement." In *Enhancing Human Capacities*, edited by Julian Savulescu, Guy Kahane, and Ruud ter Meulen. Wiley-Blackwell, 2011.
- Schaefer, G. Owen, Guy Kahane, and Julian Savulescu. "Autonomy and Enhancement." *Neuroethics* 7, no. 2 (August 2014): 123–36. doi:10.1007/s12152-013-9189-5.
- Schechtman, Marya. *The Constitution of Selves*. Ithaca ; London: Cornell University Press, 1996.
- Sen, Amartya. *Development as Freedom*. Oxford: Oxford University Press, 2001.
- — — . *Resources, Values and Development*. Oxford: Basil Blackwell, 1984.

- Sher, George. "Liberal Neutrality and the Value of Autonomy." *Social Philosophy and Policy* 12, no. 1 (1995). doi:10.1017/S0265052500004593.
- Singh, I. "Beyond Polemics: Science and Ethics of ADHD." *Nature Reviews Neuroscience* 9, no. 12 (2008). doi:10.1038/nrn2514.
- — — "Not Just Naughty: 50 Years of Stimulant Drug Advertising." In *Medicating Modern America*, edited by Elizabeth Siegel Watkins. New York University Press, 2007.
- Singh, Ilina, and Kelly Kelleher. "Neuroenhancement in Young People: Proposal for Research, Policy, and Clinical Management." *AJOB Neuroscience* 1, no. 1 (2010). doi:10.1080/21507740903508591.
- Smith, Janet. "The Pre-Eminence of Autonomy in Bioethics." In *Human Lives: Critical Essays in Consequentialist Bioethics*, edited by David Oderberg and J. A. Laing, 182–95. London/New York: Macmillan/ St. Martin's Press, 1997.
- Smith, Michael. *The Moral Problem*. Oxford: Blackwell, 1994.
- Sneddon, Andrew. "What's Wrong with Selling Yourself into Slavery? Paternalism and Deep Autonomy (¿Por Qué Está Mal Moralmente Venderse Uno Mismo Como Esclavo? Paternalismo Y Autonomía Profunda)." *Crítica: Revista Hispanoamericana de Filosofía* 33, no. 98 (2001): 97–121.
- Snowdon, P. F. "Persons, Animals, and Ourselves." In *The Person and the Human Mind: Issues in Ancient and Modern Philosophy*, edited by Christopher Gill. Oxford University Press, 1990.
- Stevens, Robert. "Coercive Offers." *Australasian Journal of Philosophy* 66, no. 1 (March 1988): 83–95. doi:10.1080/00048408812350261.
- Strawson, Galen. "Hume on Himself." In *Essays in Practical Philosophy: From Action to Values*, edited by D Egonsson, J Josefsson, B Petersson, and T Rønnow-Rasmussen, 69–94. Aldershot; Ashgate Press, 2001.
- — —. "The Self." In *Models of the Self*, edited by J. Shear and Shaun Gallagher, n.d.
- Sumner, L. W. *Welfare, Happiness, and Ethics*. Oxford University Press, 1996.
- Sunstein, Cass R. "Probability Neglect: Emotions, Worst Cases, and Law." *The Yale Law Journal* 112, no. 1 (2002): 61–107.
- — —. *Risk and Reason : Safety, Law, and the Environment*. Cambridge ; Cambridge University Press, 2002.
- Swanton, Christine. "Robert Stevens on Offers." *Australasian Journal of Philosophy* 67, no. 4 (December 1989): 472–75. doi:10.1080/00048408912343991.

- Tan, Jacinta, Anne Stewart, Ray Fitzpatrick, and R. A. Hope. "Competence to Make Treatment Decisions in Anorexia Nervosa: Thinking Processes and Values." *Philosophy, Psychiatry, & Psychology* 13, no. 4 (2007).
- Taylor, James Stacey. "Autonomy, Duress, and Coercion." *Social Philosophy and Policy* 20, no. 2 (2003): 127–55.
- — —. "Introduction." In *Personal Autonomy New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, edited by James Stacey Taylor, 2005.
- — —. *Practical Autonomy and Bioethics*. New York ; London: Routledge, 2009.
- Thalberg, Irving. "Hierarchical Analyses of Unfree Action." *Canadian Journal of Philosophy* 8, no. 2 (1978). doi:10.1080/00455091.1978.10717047.
- Tversky, Amos, and Daniel Kahneman. "The Framing of Decisions and the Psychology of Choice." *Science* 211, no. 4481 (1981): 453–58.
- United States. President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. *Making Health Care Decisions : A Report on the Ethical and Legal Implications of Informed Consent in the Patient-Practitioner Relationship*. Washington, DC: President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research, 1982.
- Valdman, Mikhail. "Outsourcing Self-Government." *Ethics* 120, no. 4 (2010): 761–90.
- Van Inwagen, Peter. *Material Beings*. Ithaca ; London: Cornell University Press, 1990.
- Veatch, Robert M. "Abandoning Informed Consent." *The Hastings Center Report* 25, no. 2 (1995): 5–12.
- Velleman, J. D. "Against the Right to Die." *The Journal of Medicine and Philosophy* 17, no. 6 (1992): 665–81.
- — —. "A Right of Self Termination?" *Ethics* 109, no. 3 (1999): 606–28.
- Walker, Rebecca L. "Respect for Rational Autonomy." *Kennedy Institute of Ethics Journal* 19, no. 4 (2009): 339–66. doi:10.1353/ken.0.0301.
- Wall, Steven. *Liberalism, Perfectionism and Restraint*. Cambridge University Press, 1998.
- Watson, Gary. "Free Action and Free Will." *Mind* 96, no. 382 (1987): 145–72.
- — —. "Free Agency." *The Journal of Philosophy* 72, no. 8 (1975): 205–20.

- Wei, F, G D Wang, G A Kerchner, S J Kim, H M Xu, Z F Chen, and M Zhuo. "Genetic Enhancement of Inflammatory Pain by Forebrain NR2B Overexpression." *Nature Neuroscience* 4, no. 2 (February 2001): 164–69. doi:10.1038/83993.
- Wells, Rebecca Erwin, and Ted J Kaptchuk. "To Tell the Truth, the Whole Truth, May Do Patients Harm: The Problem of the Nocebo Effect for Informed Consent." *The American Journal of Bioethics: AJOB* 12, no. 3 (2012): 22–29. doi:10.1080/15265161.2011.652798.
- Wertheimer, Alan. *Coercion*, Princeton: Princeton University Press, 1989.
- Westlund, Andrea C. "Selflessness and Responsibility for Self: Is Deference Compatible with Autonomy?" *The Philosophical Review* 112, no. 4 (October 1, 2003): 483–523.
- Williams, Bernard. *Moral Luck : Philosophical Papers, 1973-1980*. Cambridge: Cambridge University Press, 1981.
- Wolf, Susan R. *Freedom within Reason*. New York ; Oxford: Oxford University Press, 1990.
- . "Sanity and the Metaphysics of Responsibility." In *Free Will*, edited by Gary Watson, 2nd ed. Oxford University Press, 2003.
- Woodard, Christopher. "Classifying Theories of Welfare." *Philosophical Studies* 165, no. 3 (2013). doi:10.1007/s11098-012-9978-4.
- Yaffe, Gideon. "Indoctrination, Coercion and Freedom of Will." *Philosophy and Phenomenological Research* 67, no. 2 (2003). doi:10.1111/j.1933-1592.2003.tb00293.x.
- Young, Robert. "Informed Consent and Autonomy." In *A Companion to Bioethics*, edited by Helga Kuhse and Peter Singer, 2001.
- . *Personal Autonomy : Beyond Negative and Positive Liberty*. London: Croom Helm, 1986.
- . "The Value of Autonomy." *The Philosophical Quarterly* 32, no. 126 (January 1982): 35. doi:10.2307/2218999.
- Zimmerman, David. "Coercive Wage Offers." *Philosophy & Public Affairs* 10, no. 2 (1981): 121–45.