

ASYMMETRIES AND ILL-BEING

by

Eric Mathison

A thesis submitted in conformity with the requirements  
for the degree of Doctor of Philosophy  
Department of Philosophy  
University of Toronto

© Copyright 2018 by Eric Mathison

# Abstract

Asymmetries and Ill-Being

Eric Mathison

Doctor of Philosophy

Department of Philosophy

University of Toronto

2018

Despite the significant attention that has been given to developing and defending theories of intrinsic prudential value, or what makes us well off, far less attention has been given to explaining what makes us intrinsically badly off. In Chapter 1, I argue that theories of prudential badness, or *ill-being*, do not trivially follow from theories of well-being, yet an account of ill-being is necessary for a complete theory of prudential value. In the remaining chapters, I investigate how well the major theories of well-being extend to ill-being. Considering ill-being also makes possible asymmetries between the good and the bad. I show that there are many such asymmetries, involving both structure, such as when a condition of welfare does not apply to illfare, and value, such as when the disvalue of a bad outweighs the value of its counterpart good.

Chapter 2 concerns pain as the counterpart to pleasure. I argue that, in equal intensities, pain is a greater bad than pleasure is a good. Chapter 3 is on ‘adjusted subjective theories’, which adjust the value of pleasure or happiness based on other factors. I show how one theory—L.W. Sumner’s theory of welfare as authentic happiness—fails as an account of ill-being. In Chapter 4, I argue that the most common account of ill-being for the desire theory, that frustrated desires contribute disvalue, should be replaced by an aversion account, according to which there is a negative counterpart to desires. In Chapter 5, I consider two objective goods, autonomy and knowledge. I argue that autonomy has no negative counterpart, while knowledge has many counterparts, only some of which are intrinsically bad. In Chapter 6, I argue that, insofar as achievement is a plausible objective prudential good, there is a state of anti-achievement that is intrinsically bad. Finally, in Chapter 7 I show how the hybrid theory of well-being produces an implausible account of ill-being.

## Acknowledgements

I am grateful to all the people who helped me, both professionally and personally, throughout this project. Hasko von Kriegstein and Gwen Bradford both generously looked over chapters. Julia Nefsky gave me insightful feedback and fun cases, while Wayne Sumner took breaks from his well-earned retirement to provide moral support and to let me criticize his views. Andrew Franklin-Hall supported me both as my teaching mentor and as my internal examiner. Thanks to Shelly Kagan for being my external examiner. Anyone who has spoken to me about philosophy knows, and reading this dissertation will make clear, that his work has been the starting point for many of my own ideas. Tom Hurka has been an excellent supervisor throughout this project; every part of my dissertation is better because of his insight. As an athlete and a coach for the varsity cross-country ski team, I had my doubts that I could work with someone who has criticized my sport in print (Allemang 2011)—and who likes golf, of all things—but his intellectual generosity makes collaboration possible despite even the largest disagreements.

Jeremy Davis has been both a fruitful collaborator and a great friend. His individual contributions to this project would require dozens of footnotes, and that is even if I only gave him credit for the ideas he has given me while we watched baseball. Steve Coyne and I have run thousands of kilometers together while talking philosophy, during which time he has pushed me both physically and intellectually. I am grateful to my parents for never trying to talk me out of becoming a philosopher, and for their support in all kinds of ways I probably will not recognize unless I become a parent. Chris and Jill went above and beyond by putting me up this past year and up with me for far longer. Thanks to Charlotte for being a great roommate; sorry I told you about collective action problems. Finally, Anneke has been the best LP I could ever ask for. By writing on ill-being, I have chosen to ignore the advice to write what you know. My life has been as good as it gets, and this is in large part due to her.

# Contents

<b>Contents</b>	<b>iv</b>
<b>List of Figures</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.1.1 The Concept of Well-Being . . . . .	2
1.1.2 Two Tests . . . . .	2
1.1.3 Structural Asymmetries . . . . .	3
1.1.4 Prudential Asymmetries . . . . .	3
1.1.5 What Is the Default? . . . . .	4
1.1.6 An Intuitive Asymmetry . . . . .	4
1.1.7 Parfit's List . . . . .	5
<b>2 Pain</b>	<b>6</b>
2.1 Introduction . . . . .	6
2.2 Hedonism . . . . .	6
2.3 What Makes a Pair? . . . . .	7
2.4 The Pairwise Asymmetry . . . . .	9
2.5 Two Asymmetries about Value . . . . .	10
2.5.1 Deontic Asymmetry . . . . .	10
2.5.2 Agent-Neutral Asymmetry . . . . .	13
2.6 Prudential Asymmetries . . . . .	14
2.6.1 Anti-Natalism . . . . .	15
2.6.2 Mayerfeld . . . . .	16
2.6.3 Marginal Utility Test . . . . .	19
2.7 Conclusion . . . . .	22
<b>3 Adjusted Subjective Theories</b>	<b>24</b>
3.1 Introduction . . . . .	24
3.2 Authentic Happiness . . . . .	24
3.2.1 Happiness . . . . .	25
3.2.2 Information . . . . .	26
3.2.3 Autonomy . . . . .	28

3.3	Authentic Unhappiness . . . . .	29
3.3.1	Subjective Adjustments . . . . .	29
3.3.2	Non-Subjective Adjustments for Truth . . . . .	30
3.3.3	Indoctrination . . . . .	32
3.3.4	The ‘Undefined’ Dilemma . . . . .	34
3.4	Conclusion . . . . .	35
<b>4</b>	<b>Desires</b>	<b>36</b>
4.1	Introduction . . . . .	36
4.2	The Privation View . . . . .	37
4.3	The Frustration View . . . . .	39
4.3.1	The Polar Problem . . . . .	40
4.3.2	The Risk Aversion Accounting Problem . . . . .	41
4.3.3	Conditional Desires . . . . .	42
4.4	The Aversion View . . . . .	44
4.4.1	Extensional Equivalence . . . . .	49
4.5	Conclusion . . . . .	50
<b>5</b>	<b>Objective Theories</b>	<b>51</b>
5.1	Introduction . . . . .	51
5.1.1	The List . . . . .	52
5.2	Autonomy . . . . .	52
5.2.1	The Conditional View . . . . .	53
5.2.2	The Non-Conditional View . . . . .	54
5.2.3	Wrongness . . . . .	56
5.2.4	Me and Your Life . . . . .	57
5.3	Knowledge . . . . .	57
5.4	Kraut’s Developmentalism . . . . .	60
5.4.1	Un-Flourishing . . . . .	61
5.5	Narrow Perfectionism’s Problem . . . . .	63
5.6	Conclusion . . . . .	64
<b>6</b>	<b>Anti-Achievement</b>	<b>65</b>
6.1	Introduction . . . . .	65
6.2	Achievement . . . . .	66
6.3	Failure . . . . .	68
6.3.1	Goals and Attempts . . . . .	68
6.3.2	The Value Jump . . . . .	70
6.3.3	The Outcome Gap View . . . . .	72
6.3.4	The Proximity View . . . . .	74
6.3.5	Missed Value View . . . . .	77
6.4	Competence . . . . .	78
6.5	Difficulty . . . . .	79
6.6	Bungling . . . . .	80

6.6.1	Combining the Accounts . . . . .	82
6.7	Conclusion . . . . .	83
<b>7</b>	<b>Hybrid Theories</b>	<b>84</b>
7.1	Introduction . . . . .	84
7.2	The Structure of Well-Being . . . . .	84
7.2.1	Attractiveness . . . . .	88
7.3	The Structure of Ill-Being . . . . .	92
7.3.1	Restrictivism . . . . .	92
7.3.2	Permissivism . . . . .	94
	<b>Bibliography</b>	<b>97</b>

# List of Figures

2.1	Symmetry. . . . .	19
2.2	Linear Decrease. . . . .	20
2.3	Increasing Marginal Disutility of Suffering. . . . .	21
2.4	Symmetrical Curves. . . . .	21
2.5	Linear Increase, Curved Decrease. . . . .	22
2.6	Moorean Curves. . . . .	23
6.1	Election Value Jump. . . . .	72
6.2	Decreasing Marginal Disutility of Failure. . . . .	73
6.3	Outcome and Proximity (Outcome-Weighted). . . . .	77
6.4	Outcome and Proximity (Equal Weight). . . . .	77

# Chapter 1

## Introduction

*A theory about what is good is not, all by itself, a theory about what is bad. It can easily seem otherwise, because we might suppose, unreflectively, that what is bad for someone is simply the absence of what is good for him (Kraut 2007, pp. 148–149).*

### 1.1 Introduction

Well-being matters. We care about what it means to be well off, for our lives to be going well, or to be living the good life. Some moral theories endorse welfarism, according to which well-being is the *only* intrinsic good. Others deny welfarism, but most still accept that well-being is one relevant good among others. Despite the importance and appropriate attention that has been given to explaining what it means to be *well* off, far less attention has been given to well-being's negative counterpart: ill-being.<sup>1</sup> It is not clear why. Perhaps philosophers have assumed that a theory of ill-being will trivially follow from a theory of well-being, or perhaps they have assumed that what constitutes the bad life is just more obvious. In this chapter, I argue that a complete theory of prudential value, or well-being, requires an account of intrinsic badness. In the remaining chapters, I show that such an account does not trivially follow from theories of welfare by exploring how well each theory of well-being can be extended to ill-being. Some do better than others. This means that, if a theory claims to provide a full account of prudential value—i.e., it explains the goods *and* the bads—then we cannot judge it only based on what it says about the goods.

Investigating ill-being opens another area of inquiry. In the coming chapters, I will show that most theories of ill-being are in some way asymmetrical to their well-being counterpart. In some cases, features that make the account of welfare plausible become implausible when applied to illfare. In other cases, ill-being has a greater effect on the value of a life than well-being does. These asymmetries show, once again, that ill-being does not trivially follow from well-being. They also show that well-being is at most half the story.

The goal of this project is not to identify the best account of ill-being. Rather, I assess each major theory of well-being to identify its most plausible version of ill-being. In that way, this is an attempt to correct the discrepancy in attention given to welfare over illfare.

---

<sup>1</sup>There are exceptions. For an examination of ill-being in the same style I am undertaking here, see Kagan (2014).

### 1.1.1 The Concept of Well-Being

To speak of well-being is to speak of what is good for someone. To denote this concept, I have already used the terms ‘well-being’, ‘welfare’, and ‘prudential value’, but there are many other ways of describing the same thing. We can speak of self-interest. This is what the egoist solely pursues and what the altruist sacrifices. Self-sacrifice involves making oneself less well off for the sake of others. Similarly, when one acts paternalistically, one interferes with someone else for the other person’s sake. We can also speak of quality of life, the good life, benefit, and flourishing. Well-being involves a kind of subject-relativity in that it concerns how the agent is affected (whether or not he is aware of it). Following Stephen Campbell, “to say that  $x$  is good for Sam implies that Sam stands in a special relationship to  $x$ : it is something that benefits *him* and improves *his* well-being” (2015, p. 403). While on some accounts of well-being someone’s welfare can be affected without that person experiencing the change, the change still happens to the person. Sam might not *realize* that his welfare has increased, but he is still the recipient of that increase.

A prudentially good life is distinct from other types of value (Sumner 1996, pp. 20–25; Campbell 2015, p. 403). For instance, one can be living a morally good life, but that is not, on its own, any guarantee that such an individual will also be living a prudentially good life. A morally good person might be engaging in self-sacrifice, giving up her prudential interests for the sake of others. Similarly, one can live an aesthetically good life without being well off prudentially. The trope of the tortured artist demonstrates this distinction. Such a person makes beautiful works of art while being miserable. Finally, an admirable life need not be high in welfare. We admire those who unwaveringly pursue projects despite hardship and low odds of success—television coverage of Olympic athletes often captures admiration of this form—but such people can be unhappy, anxious, stressed, and might end up failing.

### 1.1.2 Two Tests

L.W. Sumner describes two tests any total theory of welfare must pass. The first is a completeness condition. For a theory to be complete, it needs to describe all of the ways that one can be made better or worse off, as well as an individual’s welfare level in any situation (Sumner 1996, p. 13). In other words, a complete theory must be able to list and justify all the relevant factors involved in well-being, both positive and negative. We need to be able to situate someone’s welfare level at any given time, and we need to track welfare as it moves up or down. This condition is obvious. If we were developing a way to measure wealth, an account that could only explain how much wealth someone possesses in certain circumstances would, obviously, be a bad account. For any given person, the account should tell us how much money she has at any given time, as well as provide a way to track gains and losses. The same is true of welfare.

There is a second structural test all complete theories must pass, which is that the theory must account for intrinsic goods, intrinsic bads, and a neutral middle. That intrinsic bads exist is intuitively obvious. Sumner notes that this structure justifies, among other things, physician-assisted death, for we recognize that the value of a life can be positive (worth living) or negative (worth not living). Though not universally acknowledged, it is clear that our lives can become so bad that we would be better off dead, a point which occurs when the disvalue of a life outweighs its value. This is only possible if there are intrinsic bads.

The existence of intrinsic bads is intuitively obvious for other reasons that align with the good–

neutral–bad structure. Hedonism gets this much right. Suppose that, as I am on my way to buy an ice cream cone, I experience excruciating pain in my back. The pain is so bad that I am unable to make it to the ice cream stand. In this case, I have quite obviously lost out on the pleasure of eating the ice cream, but that does not exhaust the source of my problems, for I have also suffered the pain in my back. It is difficult to see how one could explain my problems purely in terms of the absence of pleasure, so the most obvious explanation is that there exist both intrinsic goods and bads. Such pain is obviously bad for the one who experiences it, so intrinsic bads exist. A theory that denies the existence of any sort of intrinsic badness—such theories are called privation theories—is implausible enough to reject it.

### 1.1.3 Structural Asymmetries

In the following chapters, I will show how two types of asymmetries can arise between the good and the bad. A *structural asymmetry* arises when a feature, condition, or element of one pole does not apply to the other. A broad type of this asymmetry occurs when an intrinsic good has no negative counterpart. For example, I will argue in Chapter 5 that autonomy has this form. There is no intrinsically bad counterpart to autonomy, so even the full absence of autonomy is prudentially neutral, not bad. Privation views are structurally asymmetrical in the same way, for they claim that an element (either positive or negative) has no counterpart, and because of this, there is no intrinsic badness.<sup>2</sup>

A narrower structural asymmetry involves conditions. In Chapter 3, I discuss Sumner’s theory of welfare as authentic happiness, according to which welfare consists of being happy while being informed and autonomous. While this might be a plausible view of welfare, I argue that such conditions are implausible for illfare, which means that the complete theory is structurally asymmetrical.

There is another possibility, which I will mention here but set aside for the remainder of this project. The broadest structural asymmetry holds that a theory of the good does not extend to ill-being, so there is a completely different account of intrinsic prudential badness than there is for intrinsic prudential goodness. Suppose the desire theory is the correct account of well-being. Instead of extending the desire theory to incorporate ill-being, it is possible to argue that intrinsic bads exist, but that they are justified in some non-desire way. Such a theory would be structurally asymmetrical, for it would claim that the justification for intrinsic goodness is different from the justification for intrinsic badness. How this theory is instantiated can vary: it might claim, for instance, that the good consists in satisfied desires but the bad consists in suffering. This sort of combination is possible, but it strikes me as unlikely. Being badly off, it seems to me, has *something* to do with being well off. They are connected in some way.

### 1.1.4 Prudential Asymmetries

Whereas structural asymmetries concern which elements are included in the theory, prudential asymmetries concern the weight those elements receive. A prudentially symmetrical view of hedonism, for instance, holds that pleasure is the good, pain is the bad, and that pleasures and pains occurring in equal intensities are worth the same (they cancel each other out). Suppose we can assign values to the intensities of pleasures and pains, so that we can say, for example, that listening to Cardi B, the hippest rapper in the game, is worth +10. (Suppose also that intensity is the only dimension of evaluating pleasures and pains—i.e., there are only quantitative differences, not qualitative ones. A value of 0 occurs

---

<sup>2</sup>What makes privation views worthy of rejection is their claim that there are no intrinsic bads. A pluralist theory can remain plausible while claiming that some goods lack negative counterparts. It just cannot claim that all goods lack counterparts.

when there is neither pleasure nor pain.) Meanwhile, the pain of stubbing one's toe is  $-10$ . Prudentially symmetrical hedonism holds, therefore, that listening to Cardi B is as good as stubbing one's toe is bad because the intensities are the same.

In contrast, a theory is prudentially asymmetrical if it assigns different values to the positive and negative components of a pair when they occur in the same amount or intensity. For example, in Chapter 2, I will defend the view that there is a prudential asymmetry between pleasure and pain, such that, in equal intensities, the badness of pain outweighs the value of the pleasure. Thus, even though the pleasure from Cardi B and the pain of a stubbed toe are equivalent in terms of intensity, the pain of the toe makes one worse off than the pleasure of Cardi B makes one well off.

Two further points are worth mentioning. First, a theory with multiple goods and bads need not be prudentially asymmetrical to the same degree for each pair of elements. The weighting of pleasure and pain might be different than, say, achievement and failure. Second, all asymmetries need not be in the same direction. It is possible for the good to outweigh the bad.

Finally, notice that prudential asymmetries rely on the concept of counterparts, which holds that goods and bads can be paired (or in some sense related) in a way that others are not. When we speak of pleasures and pains being prudentially symmetrical, we are claiming that they exist on the same spectrum or dimension, such that it makes sense to claim that the positive element can be compared with the negative one, and that they are related in some meaningful way.

### 1.1.5 What Is the Default?

Theories can be structurally or prudentially asymmetrical, but one might wonder—now that I have described these possibilities—if the burden of proof falls on the simpler symmetrical view or the more complex asymmetrical one. Asymmetrical views are much less commonly defended than symmetrical ones in theories of well-being. Ill-being is rarely discussed, but when it is, it is often assumed that the theory of ill-being will match the theory of well-being. As a matter of practice, therefore, it seems that symmetry is the default. However, symmetry is often assumed but rarely defended.

If we ask what grounds we have for assuming that symmetry is true, the answer is that we have no such grounds. Minimally, symmetry is a substantive claim in need of justification. G.E. Moore puts the point this way:

We have no title whatever to assume that the truth on any subject-matter will display such symmetry as we desire to see [...] The study of Ethics would, no doubt, be far more simple, and its results far more 'systematic,' if, for instance, pain were an evil of exactly the same magnitude as pleasure is a good; but we have no reason whatever to assume that the Universe is such that ethical truths must display this kind of symmetry (1903a, p. 222).

Moore is correct in claiming that ethics would be far simpler were symmetry the norm, and it is likely this quest for simplicity that has led philosophers to assume symmetry. However, I will show throughout this project that symmetry is frequently unjustified. Although it might be nicer—more parsimonious—if the ethical world displayed symmetry, that does not constitute an argument.

### 1.1.6 An Intuitive Asymmetry

Before I develop and assess theories of ill-being for asymmetries in the coming chapters, I will describe here what I take to be an intuitive asymmetry. My intuition is that it is easier for life to go badly than

it is for it to go well. This point can be brought out in different ways. One is that, while there are many ways to be well off, there are even more to be badly off. This is the welfare version of Tolstoy’s opening line of *Anna Karenina*, according to which “Happy families are all alike; every unhappy family is unhappy in its own way.” Many things have to go right to be happy, but only one thing has to go wrong to be unhappy. That there are more ways to be badly off strikes me as obviously true. There are more sources of pain, for instance, than there are sources of pleasure. As David Benatar points out, many people suffer from chronic pain, but there is no such thing as chronic pleasure (2017).

Second, entropy is the default. It usually takes *effort* to maintain or increase our well-being, yet ill-being can result from no effort at all. The achievement of running a successful business takes more effort than the failure of running it into the ground. Obtaining knowledge takes more work than being ignorant, and for that reason knowledge is rarer than having unjustified false beliefs. Maintaining health requires exercise and healthy eating, whereas being unhealthy—a source of suffering—is easy to achieve. Even for those who stay healthy for long periods, the body is breaking down over time.

Although this project will find no universal rules for asymmetries in all theories, I will argue that the majority of plausible theories of ill-being align with the intuitive asymmetry. Sometimes this is for structural reasons, as when a condition that applies to welfare does not apply to illfare, making ill-being easier to obtain. And sometimes it is for reasons to do with value, as when the disvalue of the bad is greater than the value of the good. To be clear, affirming the intuitive asymmetry is not a test all theories of ill-being must pass. This asymmetry can be captured outside the theory of ill-being by holding, for instance, that it is a purely empirical point that there are more sources of pain than pleasure, and not something the prudential theory needs to describe. But passing the test does make the theory more plausible.

### 1.1.7 Parfit’s List

So far I have considered some of the relevant ways that theories can be distinguished, as well as the various possible asymmetries that might exist. Going forward, it will be helpful to use a list of theories that is both familiar and plausible. For this reason, the remaining chapters roughly follow Derek Parfit’s list in Appendix I of *Reasons and Persons*, according to which there are mental state accounts, with hedonism being the most prominent member; satisfaction accounts, which includes both desires and preferences, as well as the numerous restrictions to both; objective list accounts, which have perfectionist and non-perfectionist versions; and hybrid theories, which Parfit calls ‘composite theories’ (1984, pp. 493–502). Parfit’s list is the standard one, although I, like most others, will leave out some of the options he considers, such as the success theory. For organizational reasons, I will also break up families of theories into more than one chapter. For example, I have two chapters, 5 and 6, on objective theories.

Using this list is not to say that other options are impossible, or that Parfit’s approach captures the nuances of other proposed theories. His list remains attractive for its familiarity. While some have argued for other approaches, his remains the most common. And while people have argued that their theory does not fit neatly into Parfit’s categorization, most theories can be accommodated adequately, if not perfectly.

My goal for the remaining chapters is to determine the most plausible theory of ill-being for each family of well-being that Parfit describes. A different list might lead to a discussion with different emphases, but I do not think the final result would be significantly different. Even with a different approach, much of what I argue in the following chapters should still apply.

# Chapter 2

## Pain

### 2.1 Introduction

Hedonism is unique among theories of well-being in that it specifies ill-being in its simplest version. It holds that the only intrinsic good is pleasure and the only intrinsic bad is pain. We are therefore best off when we have the greatest amount of pleasure and the smallest amount of pain. Despite the *prima facie* simplicity of hedonism, there are many issues surrounding a complete account of hedonistic ill-being, many of which are due to the recent resurgence of philosophers defending the theory.<sup>1</sup> Roger Crisp remarks that the historical focus on the good as opposed to the bad is part of the reason for hedonism's previous decline. The resurgence is partly due to an increased focus on ill-being. He says that, "To many people, the hedonistic account of what is bad for people seems on the face of it more plausible than the hedonistic account of what is good" (2006, p. 102, fn. 21). Although hedonism is the most well-known theory that incorporates pain, my discussion in this chapter applies to any theory that includes pain as an intrinsic bad.

This chapter proceeds in the following way. First, I take up the issue of whether, and in what sense, we can say that pleasure is the proper counterpart of pain. I then argue for the structural asymmetry that there are more intense forms of suffering than there are types of happiness. Finally, I turn to the prudential asymmetry, where I describe problems with one method of defending the asymmetry before endorsing a second method.

### 2.2 Hedonism

According to hedonism, pleasures are intrinsically good and pains are intrinsically bad.<sup>2</sup> Hedonism in its standard version is prudentially symmetrical. It holds that pleasures and pains of the same intensity cancel out, such that the intrinsic value of a pleasure of intensity  $x$  is equal to the intrinsic disvalue of a pain of intensity  $x$ . Different types of hedonism posit different factors that influence the value of the

---

<sup>1</sup>They include Fred Feldman (2004, 2010) and Roger Crisp (2006). Peter Singer has also become more sympathetic to hedonism (2014 with de Lazari-Radek) after spending most of his career as a preference theorist.

<sup>2</sup>One of the major hedonism themes in recent years is a debate regarding what pleasures and pains actually are. Historically, hedonists claimed that pleasure and pain are sensations. In recent years, Fred Feldman (2004) has defended a view according to which they are attitudes. This debate is important for hedonism generally, but not for ill-being, so I will put it aside by speaking only of sensations. Everything I say here applies equally to attitudes.

pleasure or pain. All types agree that pleasures and pains vary in terms of intensity and duration, while some views also add quality.<sup>3</sup>

## 2.3 What Makes a Pair?

Many of the issues under discussion rely on it being the case that pain and pleasure are a pair. To say, e.g., that some amount of pain is more bad than an equivalent amount of pleasure is good is to hold that pain and pleasure are comparable in a way that, say, the mass of a rock and the height of a giraffe are not. Claims of the latter sort are incommensurable, while pleasure and pain (I shall argue) are not.

Most of us have no problem ranking intensities of pain. Patients in hospital emergency rooms are often asked to rank their pain on some sort of scale. Triage nurses ask questions such as “On a scale of one to ten, with ten being the worst pain imaginable, how bad is your pain?” and we think such questions make sense. Jamie Mayerfeld calls the result a *vague proportional measurement* (1996, p. 319).<sup>4</sup> It might not be possible to say precisely how much worse one type of pain is than another, but we can say that one is worse and give a vague sense of the difference. (Presumably emergency rooms employ the pain scale question because it is useful. This can be so even if comparing one’s own pain to the worst pain imaginable would elicit numbers clumped at the low end most of the time, and even if a two out of ten is still very painful indeed.) We can also make intuitive sense of comparisons about pain even if we grant that pains can have different qualities. A headache feels qualitatively different to plantar fasciitis, yet we can give a vague proportional measurement regarding the difference in intensities between them.

Such comparisons are also possible for pleasure, although the level of heterogeneity appears greater. It is also more difficult to imagine extraordinary levels of pleasure in the same way we can imagine extreme pain. It is not only that the depth of possible pain seems greater than the heights of pleasure—i.e., we can imagine intensities of pain that have no rival amount of pleasure—but also that it is more difficult to ascertain what the ten on the scale would be. (I return to this point in more detail in the next section.) Nevertheless, we can make sense of claims of the form ‘pleasure  $x$  is more intense than pleasure  $y$ ’, even when  $x$  and  $y$  are of different kinds. The pleasure of a back rub feels nothing like reading Wodehouse, but it is possible to make a rough comparison between them.

In order for comparisons to be possible, it needs to be the case not just that we can make proportional measurements of pleasures with other pleasures and pains with other pains, but also pleasures with pains. Just as with the other cases, comparisons in intensity between pleasure and pain are intuitively plausible. The enjoyment one experiences from a flow state is more intense than a mild headache, the pain of torture is more intense than the pleasure of listening to The Beatles, and so on. As Mayerfeld points out, we might never be able to confirm that our comparisons are accurate—even granting that they are vague in the first place—but there is no problem in principle with making comparisons, both within our own lives and in the lives of others (1996, p. 322).

Comparisons get more difficult when they are made between people. The difficulty largely stems from the subjective nature of judgements about pleasure and pain, and more so as we extend from pure physical sensations to mental ones, which we might call happiness and suffering instead. Both types are internal to the subject, and not easily measured for interpersonal comparability. The effect of some type of pain can be greater for one person than for another, just as what one person would characterize as the

<sup>3</sup>I take up some of these adjustments in Chapter 3.

<sup>4</sup>My case for the comparability of happiness and suffering tracks his argument on pp. 318–322.

happiest moment of her life due to some source of happiness might have less of an impact on someone else. The intensities themselves can also vary. While communicating using a pain scale can be helpful, we cannot know that my three is the same as your three. These challenges aside, vague comparisons are still possible. This is, of course, the point behind the pain scale used in hospitals. People might have different conceptions of the worst pain possible, or other interpretive differences—e.g., someone in a delirium might wonder if the scale is linear or logarithmic—but the scale is still useful because it allows vague comparisons, using patients' reports and other factors, such as their behaviour, most of the time.

An objection to the sort of pair view I am arguing for is that there is no non-evaluative sense in which we can compare intensities of pleasure and pain within the same person. Rather, any claim of the form “ $x$  is more bad than  $y$  is good” is already building in the asymmetry. If this is true, asking what the graph should look like for the value of pleasure and pain is pointless. We can draw it linearly and have all the details built in. To get around this problem, we need a non-normative way to establish relative intensities before we can go on to discuss their relative values.<sup>5</sup> To see why, consider an example of a good that is purely evaluative. It would be redundant or nonsensical to ask “As the beauty of an object increases, what happens to the value of beauty?” The reason this question makes no sense is that beauty is purely evaluative. However, pleasure and pain are different. There is a non-evaluative measure, intensity, that can be established separately from the evaluative measure. This distinction makes the following sentence well formed: “As the intensity of the pleasure increases, does the prudential value of the pleasure increase at the same rate?” This distinction addresses the criticism that the prudential asymmetry is simply built into the scale of value, even though the units do not reflect it.

Suppose that a psychologist is measuring pain and pleasure in a laboratory. If a neutral, non-evaluative comparison is possible, the psychologist will be able to say that pain and pleasure are on the same spectrum, and that a certain pleasure can be more intense than a certain pain. These claims are non-evaluative because they are not referring to the relative worth of the pleasure and pain. It is a descriptive claim that pleasure  $x$  and pain  $y$  have the same intensity, but a normative claim to hold that the badness of pain  $y$  is greater than the goodness of pleasure  $x$ . If the objection just considered is correct, to make the normative claim is just to make the descriptive one, and vice versa: there is no non-normative sense in which one can make comparisons of that sort. If we say that a pleasure and pain are of equal intensities, then we are claiming that the goodness of the pleasure cancels out the badness of the pain if they are prudentially symmetrical, and that intensity cannot be distinguished from goodness and badness.

But we can compare intensities independent of their values. One datum for a non-evaluative comparison is that pleasure and pain are contrary to each other. Hurka gives the example of comparing the intensities of the pain of being tortured to the pleasure of eating a jelly bean (2010, p. 203). That we can make sense of this comparison is evidence that pleasure and pain are different from the stone–giraffe example. Given the possibility of these sorts of comparisons, we could in principle develop a rough list of pairs for each intensity. Perhaps the pleasure of a cup of coffee in the morning is as intense as the pain of drinking cough syrup, while the pleasure of an orgasm is as intense as a broken leg, and so on.

A different metric Hurka mentions is that pleasures and pains of comparable intensities take up a similar amount of our attention (2010, p. 203). As a pain gets more intense, it takes up more of one's attention until it is impossible to focus on anything else. Happiness and suffering work against each other, such that the happier one is, the less suffering one is feeling, and vice versa (Mayerfeld 1996, p.

---

<sup>5</sup>Thomas Hurka discusses this issue (2010, pp. 201–203).

318). It is possible to be both happy and suffering at the same time, but the presence of one precludes the presence of the other, except perhaps when both occur in low amounts. (I can enjoy a meal while being bothered by a comment a friend made to me earlier. The meal is less pleasurable due to the comment.) The same is true of physical sensations. I can feel the pleasure of a backrub while my plantar fasciitis is acting up, and I can identify that I am experiencing pleasure and pain at the same time. However, in higher amounts, they are also mutually exclusive due to the attention they take up as their intensities increase.

Collectively, these points constitute good evidence for the claim that pleasure and pain (and happiness and suffering) constitute a pair.

## 2.4 The Pairwise Asymmetry

One question is whether, as an empirical point, pleasure and pain are symmetrical in the sense that any intensity of pleasure has a corresponding possible pain of the same intensity. To explain this concept in graphical terms, we can say that the length of the  $y$ -axis in positive terms is equal to the length of the  $y$ -axis in the negative, such that there is no positive point on the  $y$ -axis for which there is not a corresponding negative point of the same value. This view comes in different forms. Perhaps additional pleasures and pains are always possible, such that there is no limit to what one can experience. This is not the only type, however, for it could be that there is some maximum amount in both directions—say, 100. In this case, there is a limit to the amount of pleasure someone can feel (+100) and there is an equal limit to the amount of possible pain (−100). Call this the pairwise structural symmetry, which we can describe as follows:

*Pairwise symmetry:* For every possible pleasure of any intensity, there is a corresponding pain of the same intensity. The same is true of pains to pleasures.

There is something to be said for this view. If you take a person feeling any amount of happiness, it seems possible to increase her happiness by, e.g., giving her some ice cream or putting on her favourite song or adding some other source of pleasure. The same point follows for pain.

Despite these initial appearances, I believe that pairwise symmetry is mistaken. In particular, there are intensities of pain that have no corresponding intensity of pleasure, and that levels of suffering one can possibly feel are significantly more intense than the possible corresponding levels of happiness. If this is correct, then there is an asymmetry between happiness and suffering, such that there are types of suffering that are much more intense than any possible corresponding amount of happiness. (This is a claim about the way the world is for humans as a matter of empirical fact, not logical necessity.)

First, consider physical pleasures and pains. Imagine the most intense pleasure possible. Perhaps it comes from an orgasm or some type of drug. Consider next the most intense physical pain possible: the pain of torture, for example. My intuition is that the types of pain possible due to torture are much more intense than any amount of pleasure. The point works equally well going the other way. Imagine the worst pain you have felt in your life or witnessed someone else experiencing (either real or in fiction). Now try to imagine a corresponding amount of pleasure. When the pain becomes great enough, there is simply no corresponding pleasure possible. A similar asymmetry holds for happiness and suffering more generally. The levels of suffering one can experience have no corresponding level of happiness. The best forms of happiness are overwhelmed by the depths of suffering it is possible to feel.

It follows that happiness does in fact have a limit, and upon reflection this is the correct result. For each source of happiness, there is an amount past which more will not increase one's level of happiness. As one adds more of some source of happiness, at some point that source will no longer contribute value to one's life. One can have too much wine, mountain climbing, sex, and Bach. (Which is, again, an unfortunate empirical fact about us as humans. We can imagine other beings having no such limitation.) There is also a finite number of sources of happiness, and although they can be combined in myriad ways, it is plausible to think that there is a limit to the amount of happiness one can experience at any moment. The same is not true of pain, however. At a certain point, having an additional gin and tonic will no longer contribute happiness, but whereas a source of happiness plateaus, more gin and tonics will contribute additional suffering. (If there is a limit, it is greater than the limit for pleasures.) This is true of other experiences. An hour-long massage would be nice, but at some point the massage would no longer be pleasurable. The same does not hold with sources of pain: being burned with a match is painful. More matches just contribute more pain.<sup>6</sup> Cases of this sort are evidence that suffering can occur in far stronger intensities than happiness can, although it is possible that, while these intuitions are common, they merely reflect a limitation of our imagination. Although the pairwise asymmetry plays no role in determining the success of hedonism as a theory of welfare, it will be relevant when I defend the prudential asymmetry below.

## 2.5 Two Asymmetries about Value

Two pleasure–pain asymmetries have received attention. In this section I describe some of their proponents. There are also those who argue for symmetry.<sup>7</sup> Both sides are guilty of assuming their positions, such that they rarely defend their views besides claiming that they are self-evident. In order to clarify a different asymmetry, which I discuss below, it will be useful to discuss these two other forms here. In contrast to the third asymmetry, which I will argue exists, I will not argue for or against these two other forms.

I will use the term *deontic asymmetry* for the claim that we have a stronger obligation to prevent suffering than we do to cause happiness. This form makes no claim about the value of the happiness and suffering, either from an agent-neutral or prudential perspective. The second asymmetry concerns not obligation, but agent-neutral value. At the same intensity, it holds that suffering is morally worse than happiness is good impersonally, or, as Sidgwick puts it, from the point of view of the universe. This type of asymmetry involves a claim about the good, for it holds that happiness and suffering are good and bad on the whole. I will call this form the *agent-neutral asymmetry*.

### 2.5.1 Deontic Asymmetry

A clear example of the deontic asymmetry comes from W.D. Ross, who is not a hedonist, but who discusses our obligations regarding pleasure and pain. He says that “the infliction of pain on any person is justified [...] by the conferment not of an equal but of a substantially greater amount of pleasure on someone else. [...] We think the principle ‘do evil to no one’ more pressing than the principle ‘do good to every one’, except when the evil is very substantially outweighed by the good” (Ross 1939, p.

<sup>6</sup>Though perhaps not every source of suffering is like this. Victims of extreme trauma can become inured as a coping mechanism, but this on its own is a source of other problems.

<sup>7</sup>They include Henry Sidgwick (1907, p. 413), J.J.C. Smart (1961, pp. 19–20), Richard Brandt (1979), James Griffin (1989), Russell Hardin (1988), Peter Singer (1993), and Torbjörn Tännsjö (1996).

75).<sup>8</sup> Here, Ross is discussing what it is right for us to do, not what is good. Later he adds that “We do consider the state of pleasure [...] to be in some sense a better state of affairs than the state of pain; and we feel ourselves under a certain obligation to produce it for other people [...] and still more to prevent or minimize pain” (1939, p. 275).

A well-known defense of the deontic asymmetry comes from Karl Popper, whose view was later called ‘negative utilitarianism’ by R.N. Smart (1958).<sup>9</sup> According to Popper, “Human suffering makes a direct moral appeal, namely, the appeal for help, while there is no similar call to increase the happiness of a man who is doing well anyway” (1971, pp. 284–285). Whereas standard utilitarianism says that our duty is to maximize the good and minimize the bad, where these are equal in weight, negative utilitarianism holds that we have a stronger duty to minimize the bad. Unfortunately, Popper says little by way of justification for this asymmetry. He seems to take it as intuitively obvious, and perhaps it is, but he does little to motivate these intuitions. The negative utilitarians are claiming that pleasure and pain are worth different amounts when they occur at the same intensity, but in a deontic rather than prudential way. This is the ‘direct moral appeal’ mentioned by Popper.

This does not mean that suffering always outweighs happiness though. H.B. Acton endorses negative utilitarianism, but writes that sometimes causing happiness can be more important than ameliorating suffering:

It might be held, indeed, that there could be a greater moral urgency to help some happy man to become happier than to lessen some other person’s self-induced suffering. If I am right about this, there is no principle according to which the lessening of suffering is always morally preferable to the promotion of happiness (1963, p. 87).

As Acton makes clear, his version of negative utilitarianism holds that pleasure is still morally relevant, but that pain morally outweighs pleasure of the same intensity. In other words, he endorses the deontic asymmetry, and not the version of negative utilitarianism rejected by R.N. Smart (1958, fn. 17), according to which there is no obligation to promote pleasure. Popper and Acton agree on this point, so Smart’s criticism is directed at a view Popper never actually held. (If Popper and Acton were making the claim that Smart accuses Popper of making, then it would be worth rejecting, but few actually endorse Smart’s version of negative utilitarianism.)<sup>10</sup> Note that Acton is adding a further deontic claim by mentioning ‘self-induced suffering’, which suggests that our duties might change when the agent’s suffering was caused by the agent.

There are also deontic claims made about the comparative deontic value of different amounts of pain.<sup>11</sup> One area where these comparisons occur is in arguments for and against prioritarianism. According to prioritarianism, the worse off someone is, the stronger our obligation is to help her. This is

<sup>8</sup>Ross holds that there are other morally relevant factors, and thus is not a utilitarian. But others hold that only this deontic asymmetry counts, which produces negative utilitarianism. For example, see Walker (1974, pp. 424–428).

<sup>9</sup>Smart’s criticism is only directed at what is sometimes called absolute negative utilitarianism. This form holds that there is only the duty to relieve and prevent suffering, from which it follows that someone in possession of a sufficiently large nuclear arsenal would be doing the best act possible by killing everyone, given that doing so would prevent all future suffering. So-called weak negative utilitarianism—the sort defended by H.B. Acton—is not prey to this objection.

<sup>10</sup>Peter Singer did endorse this version of the preference theory in the first edition of *Practical Ethics* (1979, pp. 101–103), but abandoned it in later editions.

<sup>11</sup>For this reason, those who defend it are not directly arguing for an asymmetry. However, given that less suffering requires less obligation, it follows that the steepness of the obligation line continues to decrease once it crosses from suffering to happiness. Therefore, prioritarians (and Otsuka and Voorhoeve, whom I discuss shortly) do think that happiness and suffering are asymmetrical. This discussion is relevant here because it still involves duties. It will also be relevant as a contrast to Mayerfeld’s argument for prudential marginal utility, which I defend below.

not for efficiency reasons—i.e., that the worst off are *easier* to help than the better off—but rather a claim about duty (at least according to some versions). The prioritarian claims that even if we could bring about a greater increase in welfare to someone who is better off, up to some point we still ought to help the worse-off person first. So, for example, if we can raise A from +20 to +25 or B from -10 to -8, where each of these units represents an equal change in intensity of the pleasure or pain, we have a stronger obligation to help B. It is also possible to defend prioritarianism by appealing to prudence, in which case we have a stronger obligation to help the worse off because it will make a bigger prudential difference to the individual we help. Dennis McKerlie, for example, defends this approach (2001).

Others explicitly reject the prudential claim in favour of a deontic asymmetry. Michael Otsuka and Alex Voorhoeve (2009) give the following case to show how our intuitions change when we switch from intrapersonal cases to interpersonal ones.<sup>12</sup> A healthy young adult is given the distressing news that she will soon develop one of the following conditions. She has a fifty percent chance of getting either of them. Either she will become slightly impaired, which will make it difficult for her to walk more than two kilometers, or she will experience very severe impairment, which will leave her bedridden most of the time, except for the ability to sit in a wheelchair for part of the day with the assistance of others. Luckily, a treatment is available for each of these conditions, but she cannot take both of them, and the treatment must be taken before she knows which impairment she will suffer. If she takes the treatment for the mild impairment, it will completely eliminate the impairment. If she takes the treatment for the very severe impairment, it will have no effect on the mild impairment, but it will mean that she is now less severely impaired. She will be able to sit up on her own for the entire day, although she will still require assistance to move around.

Surveys show that people are generally indifferent between the two treatments (Nord et al. 1999), which Otsuka and Voorhoeve use to conclude that the treatments would yield the same expected utility. My own intuition is similarly split, perhaps leaning slightly to taking the treatment for the very severe impairment. Otsuka and Voorhoeve then argue as follows. If you were a morally motivated stranger tasked with making this decision for someone, you would be justified in choosing either option (2009, p. 173). But consider instead a case involving not just one person, but a group of individuals, all of whom are presently healthy. Half of them will get the very severe impairment while the other half would get the mild impairment. Each member of the group has the same indifference for themselves regarding which treatment they prefer. Which group should you help in this case? According to Otsuka and Voorhoeve, the only reasonable option is to help the worse-off group (2009, p. 174). In the survey I cited above, participants strongly agree, claiming that they would prefer to treat one person with the severe impairment over a *few hundred* with the mild impairment.

Suppose this is correct. If so, it supports the purely deontic claim that there is a stronger obligation to help the worst off in interpersonal cases but not in intrapersonal cases. Therefore, this supports the claim that we have a stronger obligation to alleviate greater forms of suffering even when we could make an equal difference in terms of the change in intensity. This view is defended without any appeal to prudential value, so claiming that we should help the worst off need not entail a claim that there is a prudential asymmetry between pleasure and pain. (The details of the explanation Otsuka and Voorhoeve offer, which appeals to comparative equality, is unimportant here.) Whether or not their argument is correct, it shows that obligations need not track a difference in prudential value.

<sup>12</sup>They are arguing against prioritarianism in this paper by showing that there is a difference between intra- and interpersonal cases, which pure prioritarianism according to Parfit's description (1997) does not permit.

In contrast, others have rejected the distinction between intra- and interpersonal cases. Dennis McKerlie argues that periods of time within a life are open to priority-based judgements. For example, he thinks that a benefit will be more important when it is experienced when a person is worse off compared to better off (2001, p. 284). Similarly, he thinks it is more important to relieve pain by a smaller amount when one is suffering intensely compared to a greater reduction in milder suffering (2001, p. 287). This is for prudential reasons.

## 2.5.2 Agent-Neutral Asymmetry

Jamie Mayerfeld provides a lengthy defence of the agent-neutral asymmetry in *Suffering and Moral Responsibility* (1999). Mayerfeld endorses what he calls the *intrinsic property view*, according to which “suffering is more bad than happiness is good” (1999, p. 136). He claims that this view is not prudential: it does not hold that suffering is bad because it makes our lives go worse from our own points of view, but rather that suffering is impersonally bad (a claim he defends in chapter 4). The majority of Mayerfeld’s book is taken up with defending the agent-neutral asymmetry. However, he says that suffering is also personally bad (1999, pp. 84–85). Mayerfeld goes on to endorse what he calls the strong asymmetry, according to which “the bliss of millions cannot justify the lifelong torture of one” (1999, p. 148). Pain is morally more important than pleasure, and as the pain becomes more intense, its moral importance increases faster than the intensity increases. Although Mayerfeld is principally interested in defending the deontic asymmetry, prudence plays a foundational role in his account, so I will have reason to return to his view.

Others either reject the agent-neutral asymmetry or do not discuss it while discussing hedonism more generally. Henry Sidgwick, for example, rejects the view.<sup>13</sup> Others disagree with Sidgwick. C.D. Broad does not directly endorse the agent-neutral asymmetry, but he says that “I am more inclined to think that pain is intrinsically evil than that pleasure is intrinsically good” (1930, p. 134). G.E. Moore, at least in *Principia Ethica*, claims that the agent-neutral asymmetry is correct:

The case of pain thus seems to differ from that of pleasure: for the mere consciousness of pleasure, however intense, does not, by itself, appear to be a great good, even if it has some slight intrinsic value. In short, pain [...] appears to be a far worse evil than pleasure is a good (1903a, p. 212).

There are certainly many who disagree with Moore. Sidgwick, as I noted, thinks that pleasure and pain are symmetrical (he also gives no justification for his position), while others such as Bentham and Mill simply do not consider the possibility of any type of asymmetry. Many are perplexed at the idea of any sort of asymmetry, thinking that it is obvious that a unit of pleasure is analytically equivalent to a unit of pain. In other words, when we use units to compare things on the same spectrum we are claiming that the symmetry is true.

However, I have already discussed why this criticism fails. First, it begs the question because it assumes by definition that intensity tracks value. The asymmetry claim is that this is not the case, and so, while the burden of proof is not solely on the symmetry, it is not obviously on the asymmetry either. Another way of putting the symmetry assumption is that, while it might be the case that the axiology scale matches the linear intensity scale, this does not follow trivially.

<sup>13</sup>Sidgwick (1972, p. 413). Hurka (2014, pp. 197–198) discusses the views of Sidgwick, Broad, Carritt, and Moore on the issue of symmetry.

Defenses of the deontic or agent-neutral asymmetries are frequently silent on the issue of prudence. (Mayerfeld is an exception.) As we have seen, the non-prudential claim is either that we have a greater obligation to relieve suffering than we do to promote happiness, or that suffering is a worse bad than happiness is a good from the point of view of the universe. It is possible to hold these views without also endorsing the prudential asymmetry. In fact, most defenders of negative utilitarianism—where the agent-neutral asymmetry is clear—are silent on the prudential issue. Another way of putting this point is that negative utilitarianism can be defended in two ways: (1) By defending the agent-neutral asymmetry, or (2) by defending the prudential asymmetry. Most opt for (1) without exploring (2).

I am interested in the prudential question. A deontic asymmetry can be justified prudentially—as I will show, this is Mayerfeld’s strategy—or justified on other grounds. The presence of a prudential asymmetry provides justification, all else being equal, to reduce suffering instead of promoting happiness. But we can explore the well-being implications entirely apart from our moral obligations.

Moore’s asymmetry claim is agent-neutral—pain is a far worse evil than pleasure is a good—but he offers no defense of his position. If any asymmetry exists, we need some defense for it, and the defense cannot rely on obligations between individuals, for that could also support the agent-neutral or deontic claim, not the prudential one. We also must be careful to avoid mistaking axiology for structure. As I argued above, the pairwise symmetry claim that every intensity of pleasure has a corresponding intensity of pain is false. Torture has no positive analogue, for instance. The depths of suffering are deeper than the heights of happiness, and this structural point might mislead us into thinking that it is due to axiology.

For the negative utilitarians, pain is a worse evil than pleasure is a good when they occur at the same intensity. This is a claim about agent-neutral value. However, it is an open question whether or not the prudential asymmetry also holds. As the case of prioritarianism shows, there is no logical contradiction in claiming that there is an asymmetry that applies agent-neutrally or deontically but not prudentially. Perhaps many defenders of the former will also want to make the latter point, but it does not follow trivially.

Despite a tradition of defending some form of asymmetry between happiness and suffering, it has done little to help determine whether or not there is a prudential asymmetry. Now that we have clarified the various types of asymmetries, we are in a position to consider the prudential claim head on.

## 2.6 Prudential Asymmetries

I have argued that two proposed asymmetries—the deontic and the agent-neutral—do not directly lead to the prudential asymmetry. Our obligations or the agent-neutral value of an outcome are different than what is good for the individual experiencing the happiness or suffering. To figure out if a prudential asymmetry exists, we have seen that we cannot appeal only to what we ought to do or what has agent-neutral value. We have also seen that it is possible to mistake a prudential or moral asymmetry for a structural one.

There are additional obstacles. For instance, the following method will not work: We cannot ask someone, “on a scale in which 10 is the best pleasure imaginable, 0 is no feeling, and  $-10$  is the worst pain imaginable, find a pleasure and pain of the same intensity.” The reason this strategy will fail is that, due to the pairwise asymmetry, 10 for pain is likely much greater in intensity than 10 for pleasure, the result being that the ranking gets stretched (as it were) for pain. At some point in the scale the

numbers will not line up.

It is difficult to come up with clear ways of testing prudential symmetry versus asymmetry. One way of sorting out this issue is to ask what an individual would do when faced with a decision about decreasing some amount of suffering or increasing her happiness by the same amount. A problem with this approach, as we have just seen, is that it requires comparisons of intensity, which at best will be vague. In this section, I describe two arguments for the prudential asymmetry. The first is unhelpful and the second is open to a criticism some might think is serious.

### 2.6.1 Anti-Natalism

David Benatar defends a type of prudential asymmetry in his defence of anti-natalism, the view that it is morally wrong to have children. His asymmetry goes as follows. Everyone agrees that the presence of pleasure is good and the presence of pain is bad, but the same is not true of their absence. For Benatar, the absence of pain is good “even if that good is not enjoyed by anyone” while the absence of pleasure is merely not bad “unless there is somebody for whom this absence is a deprivation” (2006, p. 30).<sup>14</sup>

For Benatar, to be brought into existence and suffer pain is bad for the one who suffers it, whereas not being brought into existence means that, while that individual will miss out on pleasure she otherwise would have experienced, the absence of pleasure is not bad, because there is no one who is deprived of that pleasure (because that person will not exist).

With this in mind, it might seem that we have a defense of a prudential asymmetry to evaluate. Unfortunately, Benatar’s discussion is not useful as a more general argument for the prudential asymmetry for two reasons. First, his argument is solely focused on the non-identity problem, for which the standard tools to evaluate ethical claims are unhelpful. Benatar recognizes this point by claiming that deprivations of pleasure are not bad unless an individual is deprived of it. Of course, in all identity cases (i.e., everything besides non-identity ones) there is such a person who is deprived of pleasure. Benatar’s stipulation that the asymmetry only applies to non-identity implies that he thinks the asymmetry does not apply in standard cases: i.e., the cases with which I am concerned here.

Being deprived of pleasure in the standard cases do seem to constitute being made worse off in some sense. If I prevent a blind person from having his vision restored, I have harmed him because, counterfactually, his well-being would have increased had I done nothing (Hanser 2008, p. 427). This, however, reveals the second problem with applying Benatar’s argument toward more general claims about well-being, which is that being harmed or benefited need not always produce a change in well-being. Deprivation cases of the sort just mentioned demonstrate this. The harm I caused is that someone whose well-being would have increased did not get an increase in well-being. In other words, his well-being stayed the same when it otherwise would have increased, so the harm is that his well-being did not change at all.

Considering harms and benefits points to a closely related problem to the one I just described. It might be that there is, in fact, an asymmetry between harms and benefits that applies not just to non-identity cases. But given that preventative harms—i.e., a harm that prevents someone from getting a benefit she otherwise would have received<sup>15</sup>—need not involve changes in well-being, this is yet another reason why thought experiments might track the wrong type of asymmetry. It might turn out that

<sup>14</sup>Elsewhere he gives the same argument but directly refers to harms and benefits, saying “we need to make the comparison with reference to the interests of  $x$ , because we want to know whether it is better for  $x$  to come into existence or never to come into existence” (Benatar 2015, p. 23).

<sup>15</sup>The term is Hanser’s (2008, p. 427).

an asymmetry does not involve well-being *per se*, but rather harms and benefits. (And perhaps not standard harms and benefits, but only preventative ones.)

## 2.6.2 Mayerfeld

Benatar offers a type of prudential asymmetry, but its scope is limited. Mayerfeld's approach is more promising. He motivates his case for the deontic asymmetry by extrapolating from arguments for an intrapersonal prudential asymmetry, which he defends by appealing to intuitions he thinks many of us will share. (In this regard my goal here is narrower than Mayerfeld's. I am interested in establishing the prudential claim, not the additional claim that the prudential asymmetry is connected to the moral asymmetry.) He is explicit that he relies on a prudential asymmetry, which leads to the agent-neutral moral asymmetry.<sup>16</sup> In this section and the next I assess and ultimately endorse Mayerfeld's conclusion that, prudentially, pain is a greater bad than pleasure is a good. In this section I discuss arguments that I find convincing, but which to varying degrees face a common objection.

Mayerfeld provides three cases.<sup>17</sup> He first appeals to our beliefs about suffering by asking us to imagine an episode of very intense suffering, perhaps extreme physical pain. Now imagine happiness of the same intensity—it would be very intense happiness “of the most glorious kind” (1999, p. 133). According to Mayerfeld, the intense suffering would not be compensated for by the intense happiness. Instead, in terms of how well off such a person is, the happiness episode would need to be significantly longer to balance off the suffering.

The second case begins in the same way: most people will agree that an episode of intense pain would not be balanced off by an episode of intense happiness of the same length. According to Mayerfeld, the common intuition is that the happiness episode would have to be much longer or much more intense (1999, p. 133). He then offers a more specific case involving anesthesia. Suppose that a drug were invented that produced pleasure as intense as the pain anesthesia averts, and that this drug has no side effects. (It would not replace anesthesia, but rather would be given to different people to make them happy in situations where anesthesia is not required.) If happiness and suffering are prudentially symmetrical, the invention of the pleasure drug would be hailed as a discovery on par with anesthesia, for it would mean that the pain of the health problem could be balanced off by the pleasure of the drug. However, most people do not have the intuition that the drug would be as important as anesthesia, and the most plausible explanation for this intuition is that suffering is more bad than happiness is good. In order for this intuition to work, we need to consider which drug is more important for the same person, not as a tradeoff involving anesthesia for some and euphoria for others. In other words, we are to imagine the comparative importance of providing anesthesia at one time or the euphoria drug at another time to the same person. Mayerfeld thinks the choice is easy: anesthesia is more important.

Mayerfeld's final example concerns the attitudes we can have towards our own lives. He claims that we can become indifferent towards future happiness, but not towards future suffering. He says that “The

---

<sup>16</sup>The section where he gives the prudential cases is called “The Moral Asymmetry of Happiness and Suffering within Lives,” and he sometimes frames the asymmetry in terms of our duty to prevent suffering being stronger than our duty to promote happiness (1999, e.g., p. 131, p. 135 fn. 7). But at the end of the section he makes explicit that “the claimed asymmetry emerges as the view that there exists a prudential asymmetry between happiness and suffering” (1999, p. 136). In other words, the moral asymmetry that we have a stronger duty to relieve suffering than to promote happiness rests on the prudential asymmetry.

<sup>17</sup>There is an important difference between Mayerfeld's argument here and the case from Otsuka and Voorhoeve I discussed above, which is that Mayerfeld is comparing happiness and suffering while Otsuka and Voorhoeve are only looking at different degrees of suffering (or, more specifically, degrees of disability which, we can assume, will cause suffering). Below I assess Mayerfeld's claim that the badness of suffering increases faster than its intensity.

prospect of future pain disturbs us more than the forfeit of future pleasure” (1999, p. 133). Mayerfeld uses the Epicurean argument that death is not a harm as evidence for this claim. According to the Epicureans, death is not a harm, because the person who dies is not around to experience it. An upshot of this argument is that death is nothing to be frightened of, given that there is nothing bad about it for the one who dies. Epicureans, who are hedonists, could “look with equanimity on the deprivation of future happiness through death” (1999, pp. 133–134). As Mayerfeld notes, this argument has problems. But he is interested in whether the Epicureans could have held the corresponding claim about death not being a benefit. Clearly it is a benefit to die when life entails only agony. (It is one justification for physician-assisted death.) But just as the Epicureans claim that death is not a harm for the one who dies, they likewise must claim that death is not a *benefit* for the one who dies. Regardless of the plausibility of their actual view, there is an asymmetry between it and the corresponding claim about the deprivation of harm, which is evidence that, while we can be indifferent toward future happiness, we are less likely to become indifferent toward future pain. Samuel Scheffler makes the same point as Mayerfeld but in a more forceful way:

Imagine, however, a torture victim who is undergoing such horrible agonies at the hands of a sadistic Epicurean that he begs his tormenter to kill him. And imagine that the Epicurean torturer replies: “So death, the thing you fervently desire, is nothing to you, since so long as you exist, death is not with you; but when death comes, then you will not exist. It does not then concern you either when you are living or when you are dead, since in the first case it is not, and in the second case you are no more”. If the Epicurean torturer’s response seems preposterous, then it is unclear why Epicurus’s own response to those who fear death should be any less so (2014, p. 84).

Are the intuitions generated by these cases sufficient to show that there is a prudential asymmetry? I shall address each one in turn. One might reject the first case—that extreme suffering will not be compensated for by extreme happiness—by pointing to some other factor to explain our intuition. (Assuming, that is, that intuitions are consistently in agreement with him on this point.) One explanation for the asymmetry is structural. I argued above that there is an empirical structural asymmetry between pain and pleasure, such that the types of possible pain are unmatched—have no counterpart—in terms of pleasure. Therefore, when we consider extreme pain and pleasure, we might be considering more intense pain than the intensity of the pleasure. But when we try to account for this, as Mayerfeld does, by stipulating that the intensities are the same, it is difficult to know whether or not we are succeeding at comparing equal amounts. In fact, based on my argument for the pairwise asymmetry above, I believe that we are unable to imagine happiness comparable to extreme suffering.<sup>18</sup> As we make the pain less severe, at some point we will be able to imagine comparable happiness, but it is difficult to generate these sorts of cases.

The second example faces the same problem, though to a smaller degree. I am horrified when I imagine an extreme amount of suffering caused by having surgery without anesthesia. Hollywood is replete with examples, such as when Tom Hanks’s character has to knock out his own infected tooth with a rock and a figure skate in *Castaway*; or when the surgeon, played by Paul Bettany, has to operate on himself to remove a bullet in *Master and Commander*. People are similarly horrified when they hear stories of anesthesia awareness—cases in which patients are aware, and sometimes can even feel

<sup>18</sup>We could just stipulate that the happiness is unimaginably good, but the usefulness of such a move is limited.

pain, during surgery. Anesthesia is an amazing medical accomplishment without which many modern medical procedures would be impossible. In contrast, no one seriously laments the lack of a euphoria drug. While such a drug would be nice to have, it has none of the importance of anesthesia. I would be extremely distraught to know that I had to undergo surgery without sedation, but I have no similar response knowing that I have to get through my day without the euphoria drug. This sort of case does a better job of motivating the prudential asymmetry, but they might also be caused by the pairwise asymmetry that extreme forms of suffering have no positive counterpart. Just as in financial investing, where it would be unwise to choose an investment that has more room to go down than up, we might similarly choose anesthesia because it blocks off the possibility of going down so far. Anesthesia is a hedge against pain.

The final point to consider is Mayerfeld's claim that we can become indifferent toward future happiness but not suffering. Consider a variant on Parfit's amnesia case (1984, pp. 165–166), in which a patient wakes up in a hospital unaware if he has had a painful surgery or not. The procedure requires that the patient remain conscious—another case where the lack of anesthesia is deeply troubling—but he will be given a drug at the end that permanently erases his memory of the procedure, including the pain. In Parfit's example, we are to imagine ourselves as the patient, and ask which we would prefer: finding out that the surgery has already taken place, or that it has not. Everyone will prefer the former, even though the amount of pain that we would experience overall is the same either way.

But now consider how we would feel if the procedure were the same in all respects except that it caused extreme happiness instead of pain (at the same intensity).<sup>19</sup> We would, of course, prefer that we had not yet had the surgery. But if we found out that we had already experienced it, we would only be mildly disappointed. (This is not quite the same as Mayerfeld's claim that we would be indifferent, although perhaps some people would be, knowing that they would still have to take the amnesia pill afterward.) In contrast, in the original example, finding out that we had yet to have the painful procedure would be extremely upsetting. It would fill us with dread. Although the structural worry could be employed here also, this case is the least threatened by it. To completely parry it, we can start with an imaginable amount of happiness and then consider the comparable intensity of pain.

Regarding Epicurus's argument, perhaps loss aversion poses a threat to Mayerfeld's approach. When we look into the future and imagine a life of pain, we lament the loss of the goods we currently possess in a way that does not occur when we imagine continued pleasure. Looking closer, it is unclear how loss aversion affects Mayerfeld's point here. He claims that "The prospect of future pain disturbs us more than the forfeit of future pleasure" (1999, p. 133). We experience loss in both cases. It is only the size of the loss that changes. In other words, Mayerfeld is comparing the badness of death when our lives are worth living to the value of death when our lives are worth not living. In the first case, he grants the Epicurean conclusion that the loss of goods is not bad. (This is just for the sake of argument, not because Mayerfeld agrees with the conclusion.) But going from the current point of having a life worth living to a state of agony in which one's life is now worth not living fills us with dread, Mayerfeld claims, because suffering and happiness are asymmetrical.

Of the problems with Mayerfeld's approach that I have described, the alternative explanation of the pairwise asymmetry is the most pervasive. It offers an alternative explanation for why we cannot imagine happiness equal to suffering—although with varying degrees of success—and it does so without appealing to a difference in the prudential value of the two sides. When faced with the same set of cases,

<sup>19</sup>This variant on Parfit's case is used for a different purpose than mine by Brueckner and Fischer (1986, pp. 218–219).

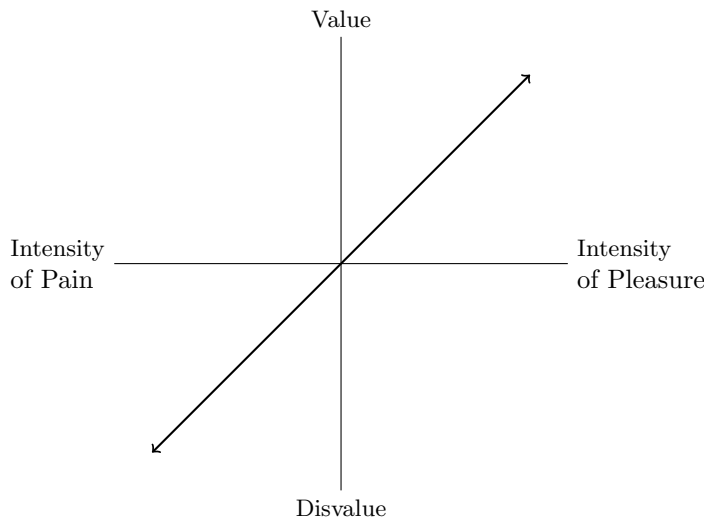


Figure 2.1: Symmetry.

someone defending prudential symmetry could appeal to the pairwise asymmetry that there are pains for which humans have no corresponding pleasure. I agree with Mayerfeld's conclusion that there is a prudential asymmetry, but I will argue in the next section that there is a better way to argue for it than the strategy he uses here. In fact, he himself uses a better strategy.

### 2.6.3 Marginal Utility Test

Mayerfeld has given us arguments that suffering is a greater bad than happiness is a good, but even if it is successful, all we can conclude from this is that the graph would look unlike Figure 2.1 and more like Figure 2.2. The problem with his approach is that it asks us to compare extreme suffering with happiness, a task which is impossible if there is no corresponding pleasure for the most extreme types of pain. A better strategy is to look at happiness on its own, then suffering on its own, then see if they produce different shapes when graphed. If they do, much of Mayerfeld's previous claims will follow.

In addition to his prudential asymmetry claim, Mayerfeld adds another: the badness of pain rises faster than its intensity. For example, a pain of intensity 20 is more than twice as bad as a pain of intensity 10. Regarding suffering, Mayerfeld has in mind something like the view depicted in Figure 2.3. Comparing suffering of a "serious, fairly intense kind" with suffering four times as intense, Mayerfeld thinks that it would be better to experience far more than four hours of the less intense suffering than experience the more intense one. This description is unaffected by the structural asymmetry for the reason that it is only comparing different levels of suffering. He expands on this point as follows:

If I think about the unimaginable prospect of twelve hours of severe torture, it seems to me that almost an indefinite extension of low-intensity suffering would be preferable. Think of someone who lives a life of persistent bleakness and deprivation: without close companionship, without diversions to lift up the spirit and gratify the senses, and without either an inward sense of accomplishment or the reception of admiration and recognition from other people to raise her feelings of self-worth. At the same time, let us imagine that this person has a certain bare level of material security, stoicism, and underlying self-esteem that together

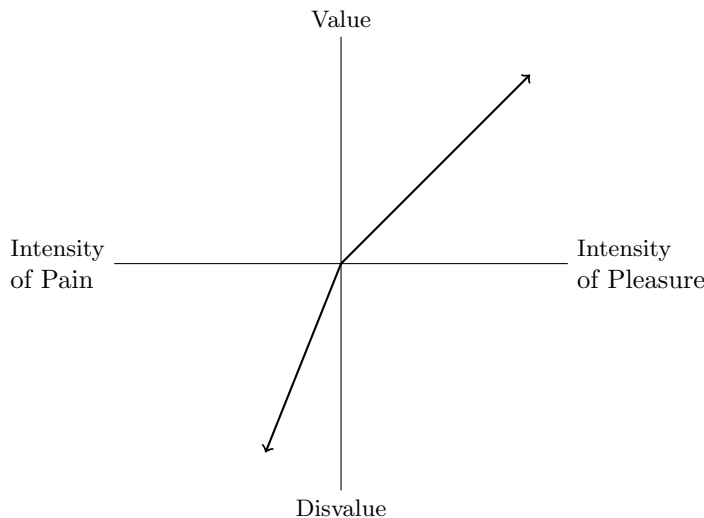


Figure 2.2: Linear Decrease.

limit the misery she feels. It seems to me that, as far as suffering is concerned, an entire lifetime spent like this would be less bad than the alternative of twelve hours of extreme torture as experienced by the same person (1999, p. 135).

This is a point about marginal utility, not just that happiness and suffering are asymmetrical. One potential problem with this type of approach is that most people will have no ability to imagine twelve hours of extreme torture. (Mayerfeld notes that humans go to great lengths to shield themselves from even glimpsing the amount of suffering in the world.) This is a difficulty because it makes precision more difficult. But this problem is not very serious. We are aware—or can become aware—that such suffering is possible. Therefore, this argument further bolsters the conclusions of the previous section, but without the potential objection that the intuition can be explained by the pairwise asymmetry.

What about happiness? If we compare a bout of happiness with happiness four times as intense, do we conclude that it would take significantly more than four hours of the former to equal one hour of the latter? The answer seems to be no. I have a much clearer intuition regarding the analogue of the torture case. (Of course, due to the structural asymmetry, there is no analogue in terms of equivalent intensity, but only in terms of a very intense amount of pleasure with a less intense form.) Consider, then, the most intense pleasure you can imagine, and imagine having that pleasure for twelve hours. Perhaps it is the pleasure of an orgasm or some euphoric drug. Now compare that to some low-intensity pleasure, such as a backrub, watching a movie, or eating a good (but not great) meal. If happiness is analogous to suffering, it should be the case that the euphoria would be preferable to a near-indefinite amount of the less intense pleasure. Is it that preferable? No. Exactly what the trade off should be is unclear, but twelve hours of euphoria would not be preferable to a lifetime of mild pleasure.

We know from the marginal utility point about suffering that the badness of suffering increases faster than its intensity. Graphically, the line for badness gets steeper as intensity increases. Based on the argument of happiness, we can rule out a symmetrical shape, where the line increases upward, forming the graph in Figure 2.4. Given that the value of happiness does not increase faster than its intensity, the main contender—and indeed the one most hedonists endorse—is that it increases linearly, as in Figure 2.5. However, this is not the only option. Moore, for instance, says that for equal increases of intensity,

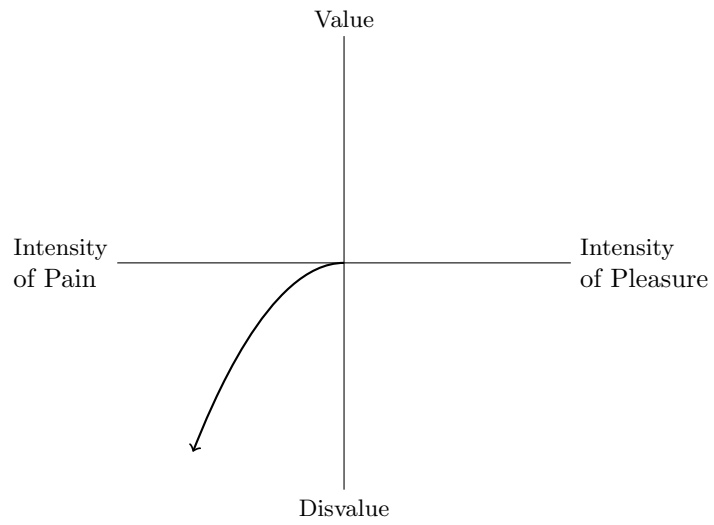


Figure 2.3: Increasing Marginal Disutility of Suffering.

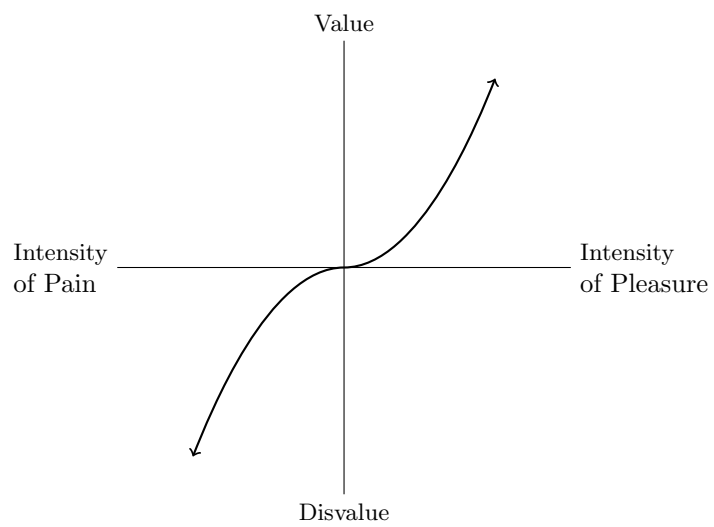


Figure 2.4: Symmetrical Curves.

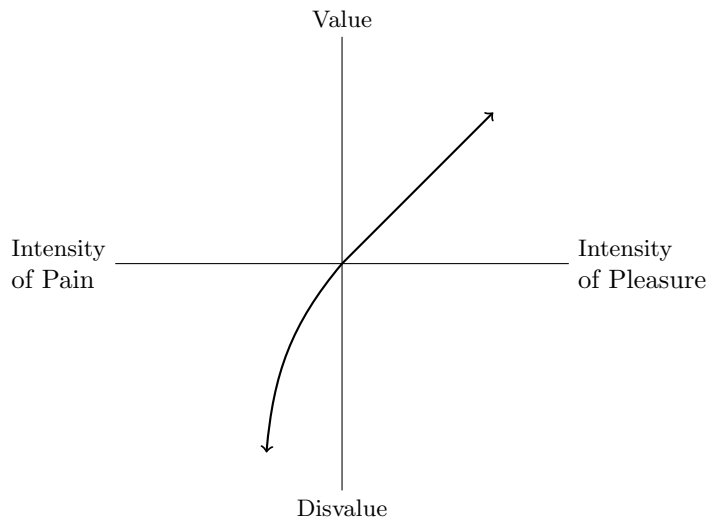


Figure 2.5: Linear Increase, Curved Decrease.

the value of pleasure gets smaller, as in Figure 2.6 (1903b, p. 358). For my purposes, it is sufficient that I have shown that happiness and suffering are asymmetrical, so I will not assess the linear claim for happiness in more detail.

## 2.7 Conclusion

In this chapter I have argued for the following claims. First, that pleasure and pain are a pair. Next, I argued against the pairwise structural symmetry, which holds that for every pleasure there is a corresponding pain. Instead, I defended the pairwise asymmetry, according to which there are pains that are more intense than any pleasure. I then demonstrated that there are problems with Mayerfeld's main argument that there is a prudential asymmetry between happiness and suffering. Most of his claims, I showed, can be explained by appealing to the pairwise asymmetry. In place of this approach, I defended the marginal utility test, which assesses happiness and suffering individually. This test supports Mayerfeld's prudential asymmetry by showing that the badness of suffering increases faster than its intensity, which is not the case for happiness.

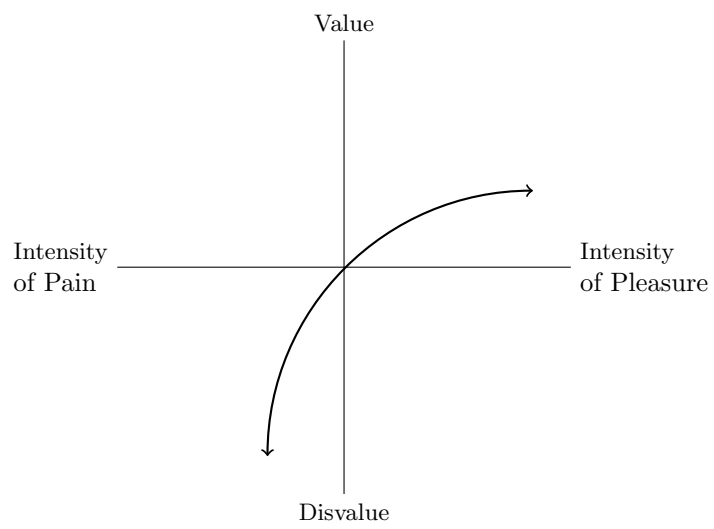


Figure 2.6: Moorean Curves.

## Chapter 3

# Adjusted Subjective Theories

### 3.1 Introduction

This chapter concerns what I call adjusted subjective theories. AS theories are subjective in that mental states play a necessary role in determining well-being. However, well-being levels can be affected by non-subjective factors.

The main example I discuss is L.W. Sumner's theory of authentic happiness, according to which someone is well off just in case she is happy and certain conditions of authenticity, namely that her happiness is informed and autonomous, are met. Ill-being on Sumner's account is autonomous dissatisfaction or unfulfillment (1996, p. 177). In the process of evaluating Sumner's theory, I discuss some related AS examples from Fred Feldman (2004), who argues that hedonism can be adjusted based on factors such as autonomy and truth.<sup>1</sup> Feldman and Sumner both offer these theories as improvements on traditional forms of hedonism, which are open to well-known objections. (Feldman characterizes his theories as types of hedonism, whereas Sumner positions his between mental state and desire theories. Pleasure need not be the only good, however. One could have an adjusted view of pleasure or happiness while allowing for other goods.) I describe how AS theories are meant to avoid some of these objections, but, as with the other theories I consider, my goal is not to fully determine if they succeed. Rather, my goal is to consider what theories of this sort say about ill-being, and whether or not considering ill-being introduces any plausible asymmetries.

### 3.2 Authentic Happiness

According to Sumner's theory of authentic happiness, the components of well-being are as follows. First, one must have an affirmative attitude towards the conditions of one's life. Second, this affirmative attitude must be authentic by being informed and autonomous. One is informed so long as one has sufficient relevant information that one's life is going well (1996, p. 160), something is relevant if the information would make a difference to the individual's subjective response, and people are autonomous insofar as they are able to critically assess and select their own values. According to Sumner, "Welfare

---

<sup>1</sup>Feldman's use of 'adjusted' is my motivation for calling this class of theories 'adjusted subjective'. Alexander Sarch calls them 'objective adjustment theories' (2012), which puts me at risk of being maximally confusing by using the opposite term. However, because only the subjective element contributes intrinsic value, Sarch's use of 'objective' is misleading. It is the subjectivity that is being adjusted.

therefore consists in authentic happiness, the happiness of an informed and autonomous subject” (1996, p. 172).

### 3.2.1 Happiness

For Sumner, happiness and suffering are better metrics for welfare than pleasure and pain. Pleasure on its own is no guarantee that our welfare will increase. If pleasures are sensations, then according to hedonism we should always prefer more of this sensation to less. However, Sumner thinks this is false. Following James Griffin, he gives the example of Sigmund Freud, who during his final days suffering from cancer chose to take only aspirin instead of stronger drugs that made him unable to think clearly (1996, p. 92). For Freud, his ability to think was more valuable to him than the absence of pain, so he chose the former, but the sensation view is committed to saying that Freud made a mistake, which will strike most of us as too heavy-handed. Freud’s decision gives priority to his preference, and therefore his subjective assessment of his life, which Sumner thinks is the correct approach. When hedonism becomes divorced from the subject’s own beliefs about the quality of his life it loses the subjectivity that makes it appealing.

The hedonist might attempt to avoid the problems with the sensation account by claiming instead that pleasures are positive attitudes (Feldman 2004, 2010), but this approach falls prey to the experience machine objection that illusory experiences are as valuable as veridical ones, even though in many cases we think veridical experiences lead to more welfare (Sumner 1996, p. 98). This is a problem for all mental state theories. While attitudinal theories can correct the problem with the sensation approach, both types mistakenly conclude that the mental state is all that matters.

Sumner notes that the point about illusion does not apply to pain—physical pains on the experience machine are as bad as pains off it (1996, p. 100)—but physical pains on their own face the problem with the sensation account, namely that there is no necessary connection between the sensation of the pain and our psychological assessment of it. Instead of focusing on pains, for which there is no necessary connection to our attitudes, we should instead focus on suffering, which “by its very nature [...] is an experience we dislike or find disagreeable” (1996, p. 105). The same is true of positive attitudes: it is happiness, not pleasure, that should be the focus of a subjective theory of welfare, for happiness is necessarily something we like or find agreeable. Importantly, the sources of happiness are broader than what we typically think of as pleasure. To say that one is happy is not the same as saying that one has pleasure. But as the criticism of the attitudinal view shows, even with this switch to happiness and suffering, it does not follow that one’s happiness maps perfectly onto one’s welfare, for some types of happiness can be more valuable than others. Life on the experience machine might still be worse than life off it.

Sumner later adds that, while attitudinal hedonists sometimes characterize pleasure as enjoyment, enjoyment is also too narrow. Happiness on his account involves satisfaction or fulfillment, so he speaks of life satisfaction.<sup>2</sup> Life satisfaction is both cognitive, involving a positive evaluation of the conditions of one’s life, and affective, involving a sense or feeling of well-being. To be satisfied with one’s life one must judge one’s life to be going well and feel good.

---

<sup>2</sup>This has led some to interpret Sumner’s view as a whole-life satisfaction account, but this is a mistake. He thinks that being happy with portions of one’s life matters too.

### 3.2.2 Information

So far I have described why Sumner thinks that the life satisfaction account is best, provided that certain restrictions are added to deal with objections that apply to all mental state theories. I turn now to the information and autonomy conditions he places on happiness.

Given that the life satisfaction account holds that someone is well off to the extent that she endorses how her life is going, it matters how she would adjust her welfare with relevant information (as judged by her). In other words, if she says that her life is going well, for her to be genuinely well off this means that either she is not being deceived about her life or deception would not bother her. And if she were to specify the features of her life that are making her happy, it is relevant to her subjective endorsement of her life that those features are genuine, that they are actually the way she believes they are.

How much does information matter? For Sumner, it depends on the individual. Once we realize that some beliefs about our lives are actually false, we might heavily discount the value of the well-being during the period of deception, or we might not. Qualifying the importance of information in this way keeps Sumner's theory subjective. When people describe the deceived businessman case, in order to motivate it they often need to stress that there is no chance the businessman will ever find out. For Sumner, finding out is irrelevant. What matters is how the businessman *would* respond were he to find out. If someone mentioned the possibility of deception before the businessman suspected anything and he said 'I would be totally devastated to discover that', then his welfare during the deception is discounted, whether he eventually finds out or not. Discussing the experience machine, Sumner says that "The extent to which the illusoriness of the experiences *matters* for an individual's well-being therefore depends on the extent to which she decides (or would decide) to *make* it matter" (1996, p. 161).

The discount rate is not uniformly applied to everyone who is deceived, but only to those who would be bothered by the deception at the rate of how bothered they would be were they to find out. The information requirement therefore parries deception criticisms of pure mental state theories. It is also different from a truth condition, which says that well-being is discounted if your beliefs are not based on the truth, regardless of how much you actually care about the truth. The information requirement keeps the determination of value firmly in the hands of the subject.

(Note that, for the reasons I have just described, the information requirement is not actually a requirement in the sense that it applies to everyone. After all, someone who does not care about being informed is not forced to discount his welfare if he is in any way deceived. He is free to say 'I understand why some people might care about information, but personally it does not bother me at all to think that I am being deceived right now, or even that I am attached to the experience machine.' In that sort of case, we cannot insist that he is worse off than he thinks he is. The information requirement is conditional. It says that well-being is affected if some piece of information would or does change your subjective assessment of your life.)

Subjective discounting has issues. It produces the strange (indeed, implausible) result that the businessman's welfare will not change when he actually finds out, even though his happiness clearly does. Imagine the businessman telling his tale of woe to a bartender, who responds that his patron is actually no worse off now that he has discovered the truth. Especially given Sumner's emphasis on subjectivity, this is a bad result.

The deceived businessman is a case of adjusting from good to bad. The same problem arises in the other direction—i.e., someone who presently believes herself to be well off but, due to misinformation, puts her welfare lower than she would with the information. Suppose Tina has just got fourth in an

Olympic race, but unbeknownst to her, the three athletes who beat her are all about to be disqualified for doping. Suppose further that fourth is the best Tina has ever placed, so she is happy with her result, but when she finds out her updated place she will be overjoyed. Although Sumner does not discuss this sort of case, we might presume that, according to his view, welfare can be adjusted to reflect the misinformation.<sup>3</sup> Tina is doing better than she thinks she is. But this leads to a problem similar to the businessman case, for on Sumner's view, it makes no difference to her welfare to find out, which is as implausible as the businessman case.

It is not that adjustment *per se* is implausible. Rather, it is the *rate* of adjustment that leads to these issues. Sumner can avoid this problem by changing the rate, which does not even require abandoning subjectivity. For example, the adjustment rate could be set at one-half of whatever the total change would be. This would mean that, until he discovers the truth, the businessman's welfare is halfway between where it would be if his family actually loved him (i.e., his current welfare level) and where he would be were he to discover the deception. This approach might be unattractive for other reasons, but it avoids the full force of the objection I have been discussing.

Adjusting in both directions makes for a type of symmetry. As I have shown, both cases have implausible implications, but adjusting up has a more serious implication, which is that, on Sumner's theory of authentic happiness, it is possible to have welfare without authentic happiness! This is the result of the Olympic case. Tina's welfare should be adjusted to the level she would be at were she in possession of the information that she has won. We can assume this level is very high. So she is well off even though she is neither informed nor as happy as she would be with the knowledge that she had won. Notice that the amendment I introduced does not solve this problem. Even if Tina's welfare is only halfway between where it is before she finds out she won and where it will be once she finds out, the features of the case mean that she will still be quite well off. In this case, adjusting up when the agent is already happy will always be a problem, regardless of the size of the adjustment.

Another problem with adjusting for information is that it erodes the subjectivity of the theory. In his discussion of the experience requirement in the context of the desire theory, Sumner uses death to explain an implication of the requirement, and in so doing argues against versions of the desire theory that reject the requirement: "Whatever may befall the living after our demise, nothing can ever again go badly (or, alas, well) for us" (1996, p. 127). If nothing can ever go badly or well for us in death, this is because we cannot have experiences. But this raises a new problem, which is that Sumner's view claims that welfare discounting can occur even if the agent never finds out about the source of the adjustment (as in the deceived businessman case). So there is an inconsistency. If discounting is possible, then his view must reject the experience requirement because he wants discounting even if the agent is unaware of the deception. But then his criticism of the desire theory—that the unrestricted view implausibly accepts posthumous harms and benefits—fails.<sup>4</sup> The experience requirement is important for Sumner. He says that "A theory of welfare can be descriptively adequate only if it incorporates some form of experience requirement; this was the important insight in classical hedonism" (1996, p. 128). But given the way adjustment works, his own view does not meet the experience requirement. His view holds that welfare can change without any experiential change, which classical hedonism denies.

<sup>3</sup>In personal correspondence he has said that this is his view.

<sup>4</sup>Note, however, that even if he rejects the experience condition, this does not entail posthumous harms and benefits on his view, for dead people can be neither informed nor misinformed.

### 3.2.3 Autonomy

The other condition of authenticity is autonomy. Suppose a woman has been indoctrinated from childhood to believe that religious worship is the sole source of genuine happiness, and enjoyment of any other sort is wrong, and therefore worthy of guilt. Or perhaps someone's life is going so badly that he lowers his expectations of his happiness, which leads him to believe that he is less badly off than someone viewing his life from the outside would conclude. Amartya Sen describes this worry as follows:

The hopeless beggar, the precarious landless labourer, the dominated housewife, the hardened unemployed or the over-exhausted coolie may all take pleasures in small mercies, and manage to suppress intense suffering for the necessity of continuing survival, but it would be ethically deeply mistaken to attach a correspondingly small value to the loss of their well-being because of this survival strategy (1987, pp. 45–46. Quoted by Sumner 1996, p. 162).

In each of these cases, the issue is a lack of autonomy, so Sumner thinks that the happiness in question is not properly the agent's own. The religious woman was indoctrinated into her faith and the man was forced to lower his expectations to make his life tolerable. A subject cannot endorse her own life or decide how satisfied she is with it when her values are not her own, so Sumner concludes that "(self-assessed) happiness or life satisfaction counts as well-being only when it is autonomous" (1996, p. 167).

How does the lack of autonomy affect welfare? To address Sen's worry, we need to do something with happiness due to lack of autonomy. Sumner describes the issue this way:

Why are we reluctant to take at face value the life satisfaction reported by 'the hopeless beggar, the precarious landless labourer, the dominated housewife, the hardened unemployed or the over-exhausted coolie'? Presumably because we suspect that the standards which their self-assessments reflect have been artificially lowered or distorted by processes of indoctrination or exploitation. In that case, the obvious remedy is to correct for the conditions under which their expectations about themselves came to be formed (1996, p. 166).

If we suspect that one's standards are artificially low—so one's welfare is artificially high—then we can adjust. But by how much? The problem with discounting for autonomy is that there is no agent-relative assessment we can appeal to, because, by definition, the agent cannot give us an authentic answer. We could instead use an objective standard regarding the badness of lacking autonomy, but Sumner rejects such objective standards because they are "unacceptably patronizing and puritanical" (1996, p. 166). There is no fact of the matter of what the values are. Additionally, in contrast to the information requirement, the autonomy requirement cannot be rejected by the individual assessing her own well-being. Someone cannot claim that autonomy is unimportant to her, or that it does not matter to her whether she was indoctrinated or not. This is because, first, such a claim would lack authenticity, and second, for Sumner, autonomy is necessary for evaluating the agent's life. If we evaluate the welfare of someone without autonomy we are, in a very real way, not evaluating *her* welfare, because she has not endorsed it.

On Sumner's account, pursuing projects that we did not autonomously choose means that those projects do not impact our well-being in the same way because we are not assessing *our* life, but rather the life someone else chose for us. If we lack autonomy, the happiness or unhappiness that results is inauthentic, and therefore does not count. The woman who was indoctrinated into her religion believes

that her life is going okay, when in fact, due to the indoctrination, it is going worse than she thought. The same is true of Sen's landless labourer, who "take[s] pleasure in small mercies". The labourer is able to suppress his intense suffering, which means that his assessment of his life will be inflated. The problem here is not that we cannot guess at what the appropriate welfare level is. After all, that is how we make sense of Sen's case. We are skeptical of self-assessments of life satisfaction by people in such circumstances "Presumably because we suspect that the standards which their self-assessments reflect have been artificially lowered or distorted by processes of indoctrination or exploitation" (1996, p. 166). If someone is not autonomous then the theory says the individual has no authentic happiness level. It produces a result of 'undefined'.<sup>5</sup> The reason for this is that, at least regarding the part of the life in question, the person has not endorsed it authentically.

### 3.3 Authentic Unhappiness

Welfare as authentic happiness states that one is well off to the extent that someone subjectively evaluates her life as going well, so long as the evaluation is neither misinformed nor coerced. Sumner describes what his theory says about ill-being in the following passage:

In explicating the happiness theory we have focused almost exclusively on welfare rather than illfare, happiness rather than unhappiness. But the analogous treatment of these latter notions is straightforward, resting as it does on the negative counterparts of personal satisfaction and endorsement. A life is therefore going badly for someone when she (authentically) experiences its conditions as unsatisfying or unfulfilling, or disclaims or disowns them (1996, p. 177).

He has since determined that extending the theory of welfare to include illfare is more complex than he describes in his book.<sup>6</sup> In this section, I aim to show why authentic happiness, regardless of its plausibility as a theory of well-being, should be rejected as an account of ill-being. I argue that dropping the authenticity requirement is the best solution. Neither information nor autonomy are defensible conditions of ill-being.

#### 3.3.1 Subjective Adjustments

Welfare as authentic happiness holds that well-being should be discounted if there is information, unknown to the individual, that would cause her to assess her life as going worse than she currently believes it to be. The discounting occurs whether or not she actually ever learns the truth.

For the theory to be structurally symmetrical—for illfare as authentic unhappiness—it must hold that ill-being can be discounted (i.e., one's welfare improves) when a welfare assessment is based on misinformation. The fact that the ill-being is in some part due to misinformation makes it less bad. Consider the following case. Frank loves Jim, but believes that his love is unrequited. This belief puts Frank into a serious depression, but without knowing it, Jim does in fact love Frank back. According to the symmetrical theory of authentic unhappiness, Frank's well-being is actually higher than he thinks, since, were he to find out that Jim loves him back, he would be overjoyed. Further, if the discount rate is symmetrical with happiness—i.e., if it is up to the individual—this will make a significant difference

<sup>5</sup>This result is not clear in the text, but Sumner has confirmed in conversation that this is his position.

<sup>6</sup>Personal correspondence and 'The Worst Things in Life', unpublished manuscript.

to Frank's welfare. Should we discount Frank's unhappiness? My sense is no. Even though Frank would change his attitude about his life by discovering the good news about Jim, until he finds out, it seems to me that Frank is simply just as badly off as if Jim did not love him. (Intuitions might be affected by supposing that Frank will eventually find out, but if we suppose that he never does, that Jim actually loves him seems not to matter.)

For a case with more serious effects, consider *Romeo and Juliet*.<sup>7</sup> When Romeo finds Juliet, believing her to be dead, he is overcome with grief, enough so to kill himself. According to Sumner's theory, we should discount Romeo's grief to the level he would be at were he in possession of the information that Juliet is only sleeping, which is to say, no grief at all (or even great happiness). But this is the wrong answer. There is no difference in Romeo's ill-being knowing that Juliet is dead versus merely believing that she is dead. It is equally bad in both cases.<sup>8</sup> Notice that this is different to the adjustment problem for well-being. There, the issue was that adjusting to the level the agent would be at knowing the truth means there is no change to welfare when he actually finds out. While that conclusion is too strong, we can still agree with Sumner's general approach that discounting by *some* amount, such as half, is appropriate, even though it requires amending his theory to set a different level of adjustment. In contrast, with ill-being I am claiming that no adjustment is justified. I claim that Frank's unhappiness is *just as bad for him* as it would be if Jim really did not love him, and that Romeo's unhappiness is *just as bad for him* as it would be if Juliet were actually dead. This matches Sumner's own claim that pain on the experience machine is just as bad as pain off it.

To get around the issues revealed by these cases, we could introduce a structural asymmetry, whereby information matters for well-being but not for ill-being. Although I have described problems for the well-being side also, adjusting for ill-being is more clearly mistaken. An asymmetry also accords with my general intuition that it is easier to be made badly off than it is to be made well off. But this poses a problem for Sumner's theory, which holds that information matters to ensure that we are evaluating our actual lives and not the ones we believe ourselves to be living. If this point is necessary for well-being, it should matter for ill-being too. So if we are stuck with symmetry, we have good reason to abandon the information condition altogether.

(Sumner's theory also allows for *increasing* ill-being when information would cause an adjustment the other way. This occurs, e.g., when you go to the doctor due to some pain and are informed that the problem is not serious, when in fact it is. If you had that information, you would adjust your welfare lower. Because this is similar to the deceived businessman's case, I will not discuss it further.)

### 3.3.2 Non-Subjective Adjustments for Truth

Another option is to introduce a truth condition that relies on a standard discount rate. In contrast to Sumner's information requirement, which assigns a discount rate based on the individual's subjective assessment of the value of correct information, truth-adjusted hedonism assigns the discount regardless of one's subjective evaluation of information. The size of the discount can vary according to non-subjective factors, such as the degree of the false belief.

Sumner's problem with this sort of account is that it defines our welfare independently of our assessment of it. He gives his religious belief as a young man as an example (1996, p. 158). When he stopped

<sup>7</sup>Thanks to Scott Lightheart-Heit for this example.

<sup>8</sup>For another example from Shakespeare, nearly the same course of events happens in *Antony and Cleopatra*, where Cleopatra pretends to die to win back Antony's love. Instead, Antony is so upset that he commits suicide.

believing in God, he did not retroactively change his assessment of his believing years; his happiness at that time still made him well off. This leads Sumner to conclude that, when we reassess our retrospective lives based on new information, there is “*no right answer to the question of what our reaction should be*” to our welfare level (1996, p. 159). A truth condition would require us to rewrite our pasts whenever we discover that our beliefs did not match the world, but he thinks that doing so is unnecessary.

This move is, in part, a response to experience machine-type objections. Suppose I spend a few weeks on the experience machine living out my fantasies. Proponents of the truth condition will claim that, once I come back to reality and realize that none of my fantasies actually happened, I should be disappointed and conclude that, while I was happy, my fantasies lacked value (and regardless of how I actually respond, my welfare was lower on the machine). Sumner rejects this conclusion, asking “Who are we to dictate that the solace someone else finds in a comforting fantasy should count for nothing?” (1996, p. 159). We might decide, in hindsight, that the experience machine was a waste of time, but we might also decide that using it made us genuinely well off.

Sumner’s view retains the subjective element that makes mental state theories attractive in the first place. The cost of this approach is that anyone who is unbothered by false beliefs is unaffected. If someone asked the businessman how he would feel were he ever to find out that his wife were cheating on him, and were he to respond that he would be completely unbothered by this information, then his welfare is unaffected by the deception. Many, however, will insist that his life really is going worse for him, whether or not he believes it to be. The costs and benefits are reversed for truth-adjusted hedonism compared to welfare as authentic happiness. On truth-adjusted hedonism, the discount rate is the same for everyone with the same degree of false belief, regardless of how they would assess their welfare. This allows us to hold that the deceived businessman who says he would be unaffected by deception is simply mistaken about his welfare, but the cost is that the view loses its pure subjectivity.

I will not attempt to resolve this dispute, but it is worth noting a couple of problems for truth-adjusted hedonism. The first is that, despite its simple appearance, it is unclear how we are to establish the discount rate in the case of falsity. Intuitively, the businessman’s welfare should be heavily discounted, whereas smaller errors—e.g., believing I got nineteenth in a race when I actually got twentieth—should be discounted far less (if at all). One might also think that justified false beliefs should receive less discounting than unjustified ones. However, this will cause its own problems. Romeo was justified in his false belief that Juliet was dead, and people on the experience machine are justified in believing that their experiences are veridical. There are also different ways that truth enters into the pleasure assessment. In one type, I take direct pleasure in a belief, such as when I take direct pleasure in my belief that I am a good piano player. In another type, my pleasure is contingent, such as when I take pleasure in my piano playing, not directly, but conditional on my belief that I could have been a professional pianist. Perhaps the role of the belief—whether direct or conditional—makes a difference to the adjustment.

This is not to say that such difficulties are insurmountable, but it does show that the theory is complex. A further problem is that there are popular false beliefs that sometimes seem to improve welfare, as in Sumner’s example of belief in a benevolent deity. This belief is as popular as it is likely to be false, yet, as Sumner put it, it seems harsh to say that the value people derive from that false belief significantly lowers their welfare (1996, p. 91). So there is an open question regarding what sorts of false belief deserve discounting, and the amount of discounting that is appropriate in each case.

What is clear is that the attractiveness of truth-adjusted hedonism for pleasure does not carry over to pain, so there is an asymmetry between pleasure and pain. Whereas pleasures on the experience

machine might be less valuable, pains are just as bad on or off the machine. The same point applies to a variation of the businessman case in which he believes that his wife does not love him and his coworkers hold nothing but contempt for him (and he is deeply pained by these beliefs), when in fact he is simply mistaken. Regarding such cases, Feldman claims that his intuitions are unclear regarding what they reveal about value:

I find that I have no clear intuitions concerning the impact of the falsity of the object of the businessman’s pain. Is his pain made worse by the fact that it is pain taken in something that is not true? Or is it made better? Or perhaps the axiological value of his pain is unaffected by the truth-value of its object (2004, p. 111).

Although Feldman is correct that intuitions are likely to be less strong compared to the pleasure cases, his last option strikes me as the most correct. Consider Romeo again. Romeo sees Juliet and, believing she is dead, is overcome with suffering at the loss of his loved one. Does it matter for Romeo’s welfare that Juliet is only sleeping? Again, my intuition is no. Similarly, adjusting for truth means that pain caused by fictional or imagined cases is less bad. This means, for instance, that crying at a movie based on real events causes more illfare than the same amount of sadness caused by a movie with no basis in reality, even if the observer is unaware which category the film is in. My sense is that the source of the sadness, in terms of its truth content, makes no difference to the badness of the suffering. Although welfare as authentic happiness and truth-adjusted hedonism have differences, on this point they are the same. Falsity of belief just does not have the same force for suffering as it does for happiness.

### 3.3.3 Indoctrination

Sumner puts the value of information entirely in the hands of the subject. Individually, we are each to decide how much information matters to us, and the extent that it matters can change depending on our circumstances. Autonomy is different. In this section I consider one approach to this condition. The view, which Sumner rejects, is that we can discount for non-autonomous happiness, just as we do for lack of information. Sumner’s view, which I discuss in the next section, is that non-autonomous happiness produces an ‘undefined’ welfare amount. Both options produce unsatisfying results. Before discussing Sumner’s view, it will be helpful to consider some of the problems with the first view, defended by Feldman, who claims that welfare should be discounted when the welfare is due to a lack of autonomy (2010, p. 192).<sup>9</sup>

The idea is that there will be some amount of discounting appropriate to bring the person’s welfare assessment in line with reality. ‘Reality’ here could mean the assessment the individual would give if he had not been forced to adjust his expectations based on his situation, if such an evaluation is possible, or the degree to which he is non-autonomous. As with information, on this approach nothing rules out adding value when someone has been indoctrinated into giving an artificially lowered assessment of his well-being.

Being indoctrinated into having expensive tastes might qualify. Suppose that Prince Harry grows accustomed to the finery of Buckingham Palace during his youth. One day he is visiting a friend’s house for dinner, but everything about the meal—which from the perspective of most people is still lavish—is far below the standards Harry is used to: e.g., they serve normal claret instead of the pre-phylloxera

<sup>9</sup>This autonomy-adjusted view has been further defended by Ishtiyaque Haji (2009, chapter 3).

vintages to which he has grown accustomed. ‘My life is going terribly right now,’ Harry thinks to himself. ‘On a scale of one to ten, I used to be a ten, but now I am probably not even a five.’ Of course, Harry is mistaken. He is extraordinarily well off, and he would evaluate it appropriately if he had not been indoctrinated into having such expensive tastes. But there is nothing he can do about that. We can claim that all assessments involving non-autonomous values should be discounted, but this provides the wrong answer in his case. Surely we do not want to claim, when Harry says that his life is going terribly, that he is actually *even worse off* than he says he is. Rather, the correct answer is that he is doing better than he thinks he is.

If we straightforwardly apply this discounting approach of autonomy to unhappiness, someone who has been indoctrinated into believing that her life is going badly might not actually be as badly off as she thinks. As with the information requirement, symmetry says that, because non-autonomous happiness is discounted, non-autonomous unhappiness should be discounted as well. Discounted here means ‘less intense’, so discounting either happiness or unhappiness brings the number closer to the neutral middle point. If person A’s well-being is +10, and person B’s is -10, and they are both non-autonomous to the same degree, then their well-being should be discounted by the same amount (e.g., -5). This will leave A at +5 and B at -5.

Discounting does not mean ‘lowered’ here, although there are good reasons for thinking that it should in some cases. Consider George, who has been indoctrinated into a cult. Before joining the cult, George was always an optimist who loved life. But according to the teachings of the cult leader, Thomas, life is supposed to be solitary, poor, nasty, brutish, and short. And the world is going to end imminently. George is deeply dismayed by all of this—his assessments of his life have become worse each day—but due to the indoctrination, he strongly believes all of Thomas’s claims.

On Feldman’s account, Sumner’s theory should say that George’s assessment should be discounted. George believes that his life is going badly, but he did not autonomously choose his goals, so in an important way George is not really assessing his own life. In other words, George is not authentically unhappy. I see no reason for agreeing with this approach. If anything, George is actually doing worse than he believes. Not only is he faring badly from his own point of view, but he has been *indoctrinated*. He is a member of a cult, a feature of his life which should make him worse off, not better off.

Because autonomy is not a matter of personal preference in the way information is, we cannot claim that the discount rate for non-autonomy—if there should be a discount rate—will depend on the individual’s assessment of her life. Whatever the appropriate formula is, it will apply to everyone equally, provided that their level of non-autonomy is the same. The easiest way out of this problem is to change the formula for discounting. On this approach, lacking autonomy makes you worse off regardless of your welfare, so if your well-being is positive, the discounting lowers it (as with person A above). But if your well-being is negative, the sign flips, so that, rather than bringing you closer to zero, it adds to how badly off you are.

The problem with this approach is that it does not account for expensive tastes. As we saw in the case of Prince Harry, sometimes one’s well-being should be *raised* from one’s own assessment. Accounting for these cases makes the theory more complicated. The formula now involves the degree to which someone’s values are non-autonomous, which will determine the appropriate amount of discounting or adding, and the direction in which the lack of autonomy has affected the person. Cases like Sen’s landless labourer require discounting (i.e., the labourer believes that his life is going better than it actually is), while cases like Prince Harry require adding (i.e., Harry’s welfare is higher than he thinks). How can we know

in which direction and to what amount the adjustment should occur? The only way seems to be to introduce a non-subjective criterion we can use to determine how prudentially well off each person is, but this would introduce a value requirement, which Sumner rejects.

We could instead claim that autonomy is structurally asymmetrical: It applies to happiness but not unhappiness. I am unsure that this approach is better than the previous option, but there is something to be said for it. Perhaps George is just as badly off as he thinks he is, and while his lack of autonomy might say something about his life, from his point of view he is faring just as badly whether or not his goals for his life are autonomous. Feldman's discounting approach for autonomy has serious problems. Sumner is right to reject it.

### 3.3.4 The 'Undefined' Dilemma

So far, all of the autonomy examples I have discussed involve people assessing their lives incorrectly. The labourer, the religious woman, Harry, and George all believe that they are either better or worse off than they actually are, and their mis-assessments are all due to a lack of autonomy. But what is the appropriate response on Sumner's theory when someone is not autonomous, but is still giving a correct assessment of her life? Sumner introduced the autonomy requirement to deal with Sen's examples, but those are all cases in which the person has inflated her welfare.

Consider Simone, who has been indoctrinated, and thus lacks autonomy, but her life assessment is correct. (The details here are not really important: Perhaps she was indoctrinated into believing a plausible set of values, or perhaps the values are ones she would not have otherwise held, but nevertheless she correctly assesses her welfare.) Given this, the theory needs to appeal to the lack of autonomy and not the gap between the agent's assessment and her welfare. Clearly, on Sumner's theory Simone's welfare must be lower than if she were autonomous, but this shows why the adjustment test cannot be the assessment the person would give of her own life were she autonomous, for then the assessments between her autonomous and non-autonomous selves would be identical.<sup>10</sup> But if the theory institutes a standard discount rate for lacking autonomy, it would become non-subjective.

The autonomy requirement is justified because the life satisfaction theory holds that the assessment must be our own. The reason indoctrination is bad is that those who are indoctrinated are unable to assess their own lives, so they are not endorsing their own values. When George assesses 'his' life, he is really using the assessment of the cult leader Thomas, which George has not autonomously endorsed. If we want to insist on the autonomy requirement, one might conclude that the correct answer to the question 'How well is George's life going for him?' is not 'His assessment plus some amount of discounting or adding', but rather 'undefined'. This is Sumner's preferred interpretation. The autonomy requirement says that George must give his own assessment, but his indoctrination makes this impossible. Therefore, there is no answer. Further, we cannot change George's desires or estimate his well-being based on how we think he might feel were he to hold those desires autonomously and were he to subjectively assess his life once he held them. That would be to presume, first, what George's values should be, and second, how well off he should be in holding them. Obviously, that would be to institute a value requirement, which Sumner has elsewhere rejected.

Sumner therefore faces a dilemma. He can either hold that value assessments of non-autonomous

<sup>10</sup>This is the case at least if the revised view is going to stay close to Sumner's original description. As I have described, Sumner's problem with indoctrination is not that it produces the wrong self-assessment of value, but rather that it interferes with the agent's independent, authentic formation of those values.

lives are impossible, or he can endorse the value requirement, according to which one's life is going worse when it is based on desires that are not autonomously chosen. Of the two, the former option is the worse choice. Embracing it means there are some circumstances in which the theory is unable to assess welfare, which violates one of the conditions Sumner places on all theories of welfare: namely, that a successful theory must be complete by accounting for all possible circumstances (1996, p. 13). In other words, a theory can never return 'undefined', which is where the first option leads.

This dilemma arises if we insist on keeping the autonomy requirement for both well- and ill-being. In fact, if we wish to avoid an objective criterion, the best option is to abandon the autonomy requirement for ill-being. Determining the correct adjustment without such a criterion is impossible, and claiming that the subject's life cannot be assessed is unacceptable. Although welfare as authentic happiness is not without problems, it is more promising than illfare as authentic unhappiness. Both information and autonomy for ill-being lead to unacceptable consequences for the theory.

### 3.4 Conclusion

There are two upshots to this investigation of Sumner's theory. The first is that it shows how extending a theory of well-being to include ill-being can be far from straightforward, and there is no guarantee that the best theory of ill-being will be identical to the proposed theory of ill-being. The second upshot is that considering ill-being can provide reasons for re-evaluating the theory of well-being. Although assessing authentic unhappiness is not the only way to reveal the problems with Sumner's theory that I have described in this chapter, it is useful, and indeed necessary, for seeing the whole picture.

# Chapter 4

## Desires

### 4.1 Introduction

The desire theory stands in stark contrast to hedonism regarding ill-being.<sup>1</sup> Whereas pleasure is usually accompanied by a discussion of pain, ill-being for desires is much less discussed. In this chapter I defend the aversion view, according to which there is a negative counterpart to desires which, when fulfilled, decreases welfare. As I will describe in more detail in section 4.4, the aversion view has already been defended as a theory about the nature of desire—or what desires actually are—independent of their value. But the nature of desire does not, on its own, settle the issue, for it could be that, as a matter of psychological fact, the class of desire can be split into positive and negative forms, yet prudentially only the positive side plays a role. So more needs to be said to extend the aversion view to the prudential realm.

Desires, especially as discussed by desire theorists, are usually taken to be a single psychological state. This state can be a desire for something, such as desiring to try a new restaurant ( $p$ ), or a desire that something not obtain, such as a desire that my meal does not contain shellfish, to which I am allergic (not- $p$ ). Despite one desire being positive and one being negative, psychologically they are the same. On the usual understanding of desires, when I get what I want—in this case  $p$  or not- $p$ —my welfare is increased, so they are prudentially the same also.

A brief but persuasive argument for the aversion view as a theory of ill-being comes from Krister Bykvist, who defends it in a survey of theories of well-being (2009).<sup>2</sup> Bykvist points to a problem for the frustration view (as I call it), which remains the most popular theory of ill-being for desires. The view claims that ill-being consists in the frustration of our desires. I elaborate on Bykvist's argument below and give additional reasons against the frustration view. Despite what I will show is its clear attractiveness over the frustration view, no one has given a full defence of the aversion view. Therefore, my goal in this chapter is to demonstrate why it is a more plausible theory than its competitors.

---

<sup>1</sup>Although I will often speak of the desire theory, what I say applies to any view that includes desires as a component of welfare.

<sup>2</sup>By 'defence' I mean that he thinks the aversion view is the most plausible version of the desire theory, not that he is a desire theorist. I am taking the same approach: I claim that desire theorists should adopt the aversion view, but I am not trying to defend the desire theory as the best general theory of welfare.

## 4.2 The Privation View

Consider this basic version of the desire theory. I fare well if and only if my desires are satisfied,<sup>3</sup> regardless of what those desires are and when the satisfaction occurs. Although a number of criticisms have been raised against this unrestricted version, most of them are unimportant here. My focus is on what ill-being consists in for a view involving desires, and one way of getting at this question is to ask what makes one worst off on any version of the desire theory. The answer, presumably, is that I am worst off when none of my desires are satisfied.

This is a type of privation view because it denies the existence of intrinsic badness as a substantive entity. The worst possible life is the one that has the greatest absence of intrinsic goods, and the only ways to lack intrinsic goods are to form a desire and have it go unsatisfied, or never to form a desire in the first place. For example, if my one desire in life is to have children and I never do, then I fail to gain any welfare. Another way of putting this point is that privation for desires involves one psychological state (a desire for  $p$  or not- $p$ ) and one significant evaluative state (that the only way for welfare to change is for a desire to be satisfied).

In Chapter 1, I argued that the privation view is unacceptable because it denies the existence of intrinsic bads. When applied to desires, there are specific problems. Imagine Ryan, who has a huge number of unsatisfied desires and no satisfied ones. Ryan wants to write a novel but fails to get past the first chapter, he wants to have a successful marriage but his partner leaves him, he wants to be in good health but he has chronic pain, and so on. According to the privation account, Ryan's well-being is at zero. What makes this story implausible is that, according to the privation view, no matter how many desires go unsatisfied, Ryan's well-being never gets any lower, yet the satisfaction of one trivial desire makes a positive difference.

It is unintuitive that unsatisfied desires merely fail to contribute welfare but do nothing to make one worse off every time. Call this the additive problem. In a possible world in which my only desire is to run a four-minute mile and I fail (and the strength of that desire was 10 units), it is reasonable to think that such a world is still better than one in which I want to run a four-minute mile (10 units) and have children (10 units) yet neither desire is fulfilled. If the strength of the desires is equal then, plausibly, adding failed desires makes one increasingly worse off, at least some of the time. Later in this chapter I will argue that sometimes failed desires do simply result in no welfare change. However, this problem reveals the larger problem with privation theories, which is that they have no account of intrinsic badness.

The argument for the additive problem relies on the claim that some people can have a greater total amount of desire than others, and that what matters is this absolute total instead of the proportion of one's satisfied desires compared to the total amount of desire, satisfied or otherwise. On this proportional view, if Marcus has half of his desires satisfied, then it does not matter what the total amount of his desires is (but suppose it is 100). Compared to Charlotte, whose total amount of desire is 50 but who also satisfies half of her desires, the fact that Marcus has a higher absolute score—50 over Charlotte's 25—does not matter. All that matters is the proportion. (Assume that the strength of each desire is the same.) Therefore, on the proportional view, Ryan is made no worse off as he adds unsatisfied desires, because his desire success rate does not change. The reason I reject this view is that it entails that someone with a single satisfied desire is better off than someone who has a greater total number of

---

<sup>3</sup>I am using 'satisfaction' and 'fulfillment' interchangeably.

desires but not all of them are satisfied, which I take to be the incorrect answer. If that single satisfied desire is for something trivial, such as getting Jell-O for dessert, then it is obvious that such a person is not faring better than someone with more fulfilled desires but with some unfulfilled ones also. Therefore, the additive problem is a genuine problem.

States such as pain are evidence for the existence of intrinsic bads. Similarly, the additive problem is further evidence that the privation view is implausible. Although there might be ways around these issues, they suggest that we should look for an account that includes intrinsic bads.

While classical privation views have largely claimed that intrinsic badness is merely the absence of goodness, others have gone the other way by claiming that there are no intrinsically good outcomes. One contemporary version of this type of privation view comes from Peter Singer (1979, pp. 101–103). In a review of the first edition of Singer’s *Practical Ethics*, H.L.A. Hart accuses Singer of failing to give an argument against what is now called the Replaceability Thesis, which claims that killing someone is permissible so long as the individual is replaced by another with an equal or greater number (or strength) of desires (Hart 1980). In a letter to the editor responding to Hart’s claim, Singer gives the following response:

Hart says that Preference Utilitarianism is a form of maximizing utilitarianism. This is true in the sense that Preference Utilitarianism directs us to maximize the satisfaction of existing desires, but not in the sense that it directs us to create more beings with desires that we can satisfy. The creation of desires which we then satisfy gains us nothing. We can think of the creation of the unsatisfied desires as putting a debit in the moral ledger which satisfying them merely cancels out. That is why Preference Utilitarianism can hold that it would be bad deliberately to create a being most of whose desires would be thwarted, and yet hold that it is not a good thing to create a being most of whose desires will be satisfied (Singer 1980).<sup>4</sup>

According to Singer’s description, there is nothing good about satisfied desires, but there is something bad about unsatisfied ones. He argues that this explains why the Replaceability Thesis is mistaken. (Singer speaks of ‘preferences’ while I use ‘desires’. There is no important difference between them for my purposes.)

In later editions of *Practical Ethics*, Singer changed his position on replaceability. In the second edition, he points out a problem with the type of ‘moral ledger’ view for which he initially argued (Singer 1993, p. 129). The issue is that even the best-off person will have some unsatisfied desires, leaving a small debit in the ledger. This implies that it would have been better for none of us to have been born, which Singer thinks is absurd, and I agree. Therefore, the ledger view must be rejected.

Luckily, we do not need to settle the complicated issues raised by Singer’s discussion in order to reject the privation view. Just as the classical privation view fails because it offers no reason to believe that our intuitions about the existence of intrinsic badness is mistaken, the view Singer describes fails because it denies the existence of intrinsic goodness.

---

<sup>4</sup>Hurka noted in conversation that this view is similar to the ‘dependence effect’ described in John Kenneth Galbraith’s *The Affluent Society*: “Wants are increasingly created by the process by which they are satisfied” (1998, p. 100). For related implications of this view, see my discussions of negative utilitarianism and Benatar’s anti-natalism (2006) in Chapter 2.

### 4.3 The Frustration View

In the previous section we saw some specific problems for the privation theory of desires, which holds that there are no intrinsic bads, so a total absence of intrinsic goods constitutes the worst possible life. Any attempt to get around these problems must distance itself from the privation view by positing the existence of intrinsic badness. It is possible, though, that we need not travel too far to get a solution. One approach is to claim that the frustration of a desire constitutes an intrinsic bad for the individual. Call this the frustration view. Whereas the privation view has one psychological state and one significant evaluative state, the frustration view sticks with one psychological state but adds a second evaluative state, which is the badness of a frustrated desire. When ill-being is mentioned in discussions of the desire theory, it is usually described in desire-frustration terms. For example, Derek Parfit describes the desire theory this way:

In deciding which alternative would produce the greatest total net sum of desire-fulfilment, we assign some positive number to each desire that is fulfilled, and some negative number to each desire that is not fulfilled. How great these numbers are depends on the intensity of the desires in question. [...] The total net sum of desire-fulfilment is the sum of the positive numbers minus the negative numbers (1984, p. 496).

Chris Heathwood gives a similar description for what he calls subjective desire satisfactionism. A theory of welfare that is purely desire-based (i.e., the desire theory) is subjective in the sense that the desires are, at least in part, up to the agent. But theories involving desires have a different objective-subjective distinction involving the desire's fulfillment. When the theory holds that it is the actual fulfillment of the desire that increases welfare, this is known as objective desire satisfaction. This view is most common. But as Heathwood shows, a subjective version is also possible:

An instance of “subjective desire satisfaction” is a state of affairs in which a subject (i) has an intrinsic desire at some time for some state of affairs and (ii) believes at that time that the state of affairs obtains. An instance of “subjective desire frustration” occurs when (i) above holds but the subject believes that the desired state of affairs does not obtain. The value for the subject of (or the amount of welfare in) a subjective desire satisfaction is equal to the intensity of the desire satisfied. Likewise for frustrations, except that the number is negative (2006, p. 548).

For my purposes, the important feature of Heathwood's definition is not its subjectivism, but its claim that frustrated desires decrease the agent's welfare. We could just as easily make an objective version that involves frustrations. The frustration view is also tentatively endorsed by Shelly Kagan as the most plausible form of ill-being for the desire theory (2014, p. 272).

To motivate this account, consider the case of Hannah, who has the strong, persistent desire to have children. One day Hannah is told that she will never be able to conceive a child. In this case, it would be a mistake to say that Hannah's inability to conceive can be described purely in terms of the absence of welfare she would have received from being a mother. It seems more plausible to say that Hannah has suffered a substantive loss—she is worse off with the frustrated desire for children than if she had formed no such desire at all. As this case shows, at least some of the time, the failure to satisfy our strongest desires is not merely to miss out on a good, but to experience the presence of an intrinsic bad.

The frustration view solves two of the problems with the privation view. The first is the additive problem, according to which adding more frustrated desires made Ryan's welfare no lower according to the privation view. The frustration view avoids this problem, for now the presence of additional frustrated desires contributes a greater amount of badness. On this view, my level of well-being in the world where I only fail to run a four-minute mile is less bad than the one where I fail to run a four-minute mile *and* fail to have children (and desired both and both were of equal strength). The frustration view gives the correct answer in this case, and to that extent it is more plausible than the privation view. I argue in the next section, however, that not all cases fit this model.

### 4.3.1 The Polar Problem

To see a main issue with the frustration view, let us return briefly to the privation view. The problem with the privation view is that the level of badness in someone's life bottoms out at zero. The worst life according to the privation view is one with no satisfied desires, which is possible in two different ways. First, one can have desires but they are not fulfilled—a result in itself which is implausible because it entails that adding unsatisfied desires does not make one any worse off once the individual is at zero—or, second, one can form no desires, in which case it is trivially true that no desires are ever satisfied. This problem reveals itself in different ways. For instance, I stated that a complete theory of well-being should allow for lives in which the bad outweighs the good. We can make sense of the notion that a life can be 'worth not living', that such an individual would be better off dead, so our theory of well-being should account for this. On the privation view, this is impossible. The worst possible life is one without any intrinsic goods, yet we can imagine lives being far worse than this: e.g., someone whose entire existence consists of excruciating pain.

The frustration view suffers from a related problem, except that, rather than allowing two states—positive and neutral—while omitting lives with negative value, on the frustration view, forming a desire leads to two possibilities: it can make one better or worse off without any way of staying the same. But, intuitively, there are some failed desires that merely fail to make us better off instead of adding intrinsic badness. Because it emphasizes the extremes without having a middle, I call this the polar problem.

To be clear, having a welfare level of zero is still possible on the frustration view if, as with the privation view, one never forms desires. Another way is for desires to cancel out. If desire A obtains and desire B does not, and they both have the same strength (and if there are no other desires) then the desires cancel out and the individual remains at zero. The polar problem concerns cases involving single frustrated or unsatisfied desires. The problem is that, despite what the frustration view claims, it is not the case that frustrated desires make us intrinsically worse off every time.

At least sometimes, a frustrated desire merely leads to the absence of a good instead of an intrinsic bad. Consider a case from L.W. Sumner. You are enjoying a bottle of cheap wine, but you consider how you would get more enjoyment from a more expensive bottle. While you might desire to be drinking the nicer bottle, that you are not does not intrinsically hurt you. You are only comparatively worse off (unpublished, p. 10). In other words, there can be situations where we recognize that, were things different, we could be better off than we are. But this on its own does not constitute an intrinsic bad. In terms of intrinsic value, thinking about the nicer bottle is the same as not thinking about it: there is no change. Bykvist gives a case with the same structure (2009, p. 40). Many people enjoy getting unexpected gifts and, on occasion, they reflect on this desire by thinking about the enjoyment they receive when someone surprises them with one. Even though this desire exists, it need not be the

case that one must also have a negative attitude toward the absence of surprise gifts. One can desire unexpected gifts but be neutral toward not getting them. (And, just as above, the welfare implication of not getting a gift is the same as when one never forms the desire for unexpected gifts.) The cases from Sumner and Bykvist show the same thing, which is that a satisfied desire can improve welfare without the frustration of that desire decreasing it.

Another way of understanding the polar problem is to consider what an analogous view would look like for hedonism. According to hedonism, if I miss out on some pleasure, then I miss out on a well-being increase I otherwise would have had. In order for my well-being to decrease, I must have some pleasure I possess taken away, or I must experience some pain. The frustration view according to hedonism would entail that failing to get pleasure is always equivalent to gaining pain, but this is mistaken. More plausibly, at least some of the time, missing out on some pleasure is merely to miss out on that good; there is no intrinsic bad entailed. Of course, we can be pained by missing out on a good. One can be devastated to miss out on a job, for instance. We can make sense of the distinction between failing to improve one's well-being and actually decreasing well-being (or adding ill-being). The frustration view denies this distinction, which makes it unsatisfactory. The polar problem results from the frustration view's claim that frustrated desires always constitute an intrinsic bad. This is implausible. At least in some cases, the frustration of a desire constitutes only the failure to gain an intrinsic good.

### 4.3.2 The Risk Aversion Accounting Problem

Another problem for the frustration view involves risk. Assume that the desire theory is the correct account of well-being and that the frustration view is the correct account of ill-being. Now suppose I am considering some activity, such as going to check out the new restaurant that just opened in my area. I know that the restaurant is likely to be busy due to the reviews of the restaurant and the high number of foodies in my community, so there is a good chance that I will not be able to get a table. Still, I form the desire to eat at the restaurant tonight. Now, according to the frustration view, if I get into the restaurant my welfare will increase  $x$  amount, but if I do not get in it will decrease  $x$ .<sup>5</sup> This is risky! By forming the desire to eat at the restaurant, I know there is a good chance that my welfare will actually drop. Therefore, if the frustration view is correct, I ought to be much more careful about the desires I form than I imagine people actually are about, e.g., deciding which restaurant to visit. This is true regardless of the strength of the desire, but the frustration view is even more risky the stronger one's desire is. My intuition is that it is the theory, not the way people choose their dining options, that is mistaken.

The situation gets more troubling for the frustration view, for consider the sorts of desires people form that have a very low chance of fulfillment, such as becoming an astronaut or winning a Nobel Prize. If the frustration view is correct, then forming these sorts of desires would be a horrible idea, since it is almost certain they will be frustrated. Consider two people, Richard and Susan, who live identical lives as scientists producing equally significant work and who are in all other ways satisfying desires of equal intensities. But imagine Susan has one additional desire: She desires to win the Nobel Prize this year, whereas Richard does not. On occasion, Susan considers just how great it would be for her to win; the thought of winning brings her joy, and even though she knows that winning is extremely unlikely—many brilliant scientists never end up winning—she still has the desire. Suppose too that, while the thought

<sup>5</sup>I am supposing  $x$  is the same value, but there could be different values if there is an asymmetry between the value of a satisfied desire and the disvalue of a frustrated one.

of winning excites her, the thought of not winning does not trouble her.

Now, suppose you are friends with Richard and Susan, and one day Susan tells you of her desire. Most people, I assume, would respond in something like the following way: ‘Susan, that would be so great! What an honour!’ However, a defender of the frustration view should respond quite differently: ‘Susan, that’s so horrible! Even though you’re a brilliant scientist, there’s almost no chance that you’ll win the prize. You’ve doomed yourself to a reduction in well-being. Forming such a desire just isn’t worth the risk, so do whatever you can to rid yourself of it.’ This strikes me as an unjustified response, not because it is inconsistent with the frustration view, but because the risk-averse implication of the frustration view is mistaken. To put it a different way, suppose that the formation of at least some of our desires is out of our control. When choosing between the two possible lives, do we think Susan or Richard is luckier? I lack a clear intuition in this case, but note that the answer according to the frustration view is clear: Richard is much luckier to lack the risky desire.

The problem with the frustration view, which the case of Richard and Susan illustrates, is one of accounting. In effect, the theory favours desires involving little risk. This is due to the view’s defining feature: the frustration of desires constitutes an intrinsic bad, where the riskier the desire—that is, the less likely that it will be satisfied—the greater the chance that the individual will be made badly off.<sup>6</sup>

This strikes me as an implausible claim for every type of desire. As we saw above, the failure to satisfy some types of desires might result in an intrinsic bad, but the case of Susan suggests that this cannot always be the case. It is more plausible that Susan can reasonably say of her failure to win a Nobel Prize that, while she would have liked to have won, she was not made *worse off* by her failure to win. The frustration view tells us not to adopt risky desires in favour of so-called low-hanging fruit, yet this strikes me as the wrong answer. To be sure, the correct answer need not be a life spent chasing high-hanging fruit—presumably some balance is best—but it is the anti-risk approach that makes the frustration account implausible.

It bears noting that the details of the case will affect the objection I am making. Instead of the way I described Susan’s attitude to losing above, we can imagine Susan to be like Hannah, where Susan cares so much about winning the Nobel that she is distraught when she does not win, just as Hannah is distraught at being unable to have a child. While this sort of case is possible, my claim is that not all cases have this form. In particular, it seems false to say that Susan missing out on a Nobel decreases her welfare by the same amount that winning would have increased it, especially given her neutral attitude toward not winning. To get around this, the frustration view can introduce an asymmetry, according to which a frustrated desire is less bad than a satisfied desire is good. This sort of asymmetry might be helpful, but it is still false to say that every frustrated desire makes one worse off.

### 4.3.3 Conditional Desires

There is a potential problem for the argument I have been advancing. I have claimed that, on the frustration view, forming a desire always leads to either satisfaction or frustration. Both the risk aversion problem and the polar problem seem to rely on this being the case: e.g., the risk of forming a desire is that if it is not satisfied then that means it is frustrated, and therefore the agent’s welfare goes down. However, Kris McDaniel and Ben Bradley (2008) argue that this description of desires is false. Some desires are neither satisfied nor frustrated, which might mean that, on the frustration view, someone can

<sup>6</sup>Note that the relevant feature is risk, not strength. Of course, stronger desires that are also riskier are worse according to the frustration view than low-risk, low-strength desires.

form a desire without facing either the risk aversion problem or the polar problem.

McDaniel and Bradley argue that conditional desires can be the sorts of desires that, depending on the conditions, are neither satisfied nor frustrated. They further argue that all desires are conditional. For example, consider the following sort of case involving a desire conditional on its own persistence. Right now I desire to get a beer tonight provided that when tonight arrives I desire to have a beer (2008, p. 271). Suppose that tonight I do not want a beer. If I do not actually end up getting one, it is obviously incorrect to say that my desire has been frustrated. And if I do end up with a beer, it also seems mistaken to say that my desire has been satisfied. McDaniel and Bradley argue that the correct way of thinking about this case is that my desire has been *cancelled*, which is their term for a desire that is neither satisfied nor frustrated (2008, p. 274). In other words, “When a person’s desire that  $p$  is conditional on  $q$ , the desire that  $p$  is cancelled if and only if  $q$  is false” (2008, p. 275). This is analogous to Peter Strawson’s account of the truth-value of sentences with referring sentences that fail to refer, such as ‘The king of France is wise’. On Strawson’s account, such sentences are neither true nor false. Similarly, conditional desires can be neither satisfied nor frustrated when the condition is false (2008, p. 275).

Desires can be conditional on other factors. I might desire beer only if it is not poisoned, that it will not cause the deaths of other people, that it does not contain animal products, and so on. It is extremely unlikely that I desire beer *no matter what*. McDaniel and Bradley correctly claim that most of our desires are conditional in this sort of way, and, for reasons that are less important for the current discussion, they argue that *all* desires are conditional in some form. What we usually think of as unconditional desires are best formulated, they argue, as conditional in the trivial way that *something* must be the case (2008, p. 278). To say that one has an unconditional desire is to say that one desires  $p$  conditional on  $q$ , where  $q$  is satisfied in every possible situation. Importantly, due to the necessary truth of the conditional, these desires cannot be cancelled, so they must be either frustrated or satisfied.

The implication of all this for my argument is as follows. Susan desires to win the Nobel Prize, and her friend admonishes her for forming a desire that is so likely to be frustrated, claiming that Susan has doomed herself to a reduction in welfare. In light of the point by McDaniel and Bradley, we now know that this might not be true. Rather than being frustrated, Susan’s desire might instead be cancelled. After all, the desire to win a Nobel probably has some conditions: e.g., Susan does not want to win if her winning is due to her friend threatening the jury, or if winning will cause her to die of a heart attack immediately after, or even if just before winning she loses the desire to win. When desires are cancelled, the most obvious prudential implication is that nothing happens. The welfare level stays wherever it was before the cancellation of the desire.

Therefore, amendments are necessary to both problems I raised against the frustration view. My previous description of the polar problem said that, on the frustration view, forming a desire means that one’s welfare can either go up or down but it cannot stay the same. Given the conditional view of desires, this is untrue: the welfare level can stay the same if the desire is cancelled. Similarly, the risk aversion problem implied that, e.g., if Susan does not win the Nobel Prize, then necessarily her welfare will drop, when in fact it could stay the same if the desire is cancelled.

I grant these amendments. Nevertheless, reframing desires as conditionals solves neither the polar nor the risk aversion problem for the frustration view. Consider all of the possible conditions Susan could place on winning. It seems to me that none of them will be psychologically realistic while eliminating the risk of desiring such an unlikely outcome. For conditionals to be of any help, it has to be pretty

likely that if Susan does not win, her desire will be cancelled instead of frustrated, but, as far as I can see, there is no conditional that can help achieve this while retaining the spirit of the frustration view: i.e., that sometimes our desires really can be frustrated, and when this happens we will be made worse off. Indeed, the problem with the frustration view can be described with a conditional: if our desires are frustrated then welfare decreases every time, but this is false. The problem is with the claim about frustrations, not that all desires can involve frustration.

The same is true of the polar problem. While it is not strictly true that forming a desire means your welfare can only go up or down, cancelling does not provide a convincing solution to the problem. Consider Sumner's case of drinking a cheap bottle of wine while considering how you would enjoy a more expensive bottle more. To get around the polar problem, the frustration view needs to eliminate the frustration by showing that, despite your desire to have the nicer bottle, the desire is cancelled instead of frustrated. None of the plausible ways of cancelling the desire are going to do the trick. You could desire the expensive bottle provided that you do not have to pay for it, or provided that you are in the company of good friends, or whatever, but these do not eliminate the present problem that you are presently not drinking the expensive bottle when you desire to be, so the frustration view is committed to saying that this constitutes an intrinsic bad, which means we are back to the polar problem. My initial description of the polar problem is that, contrary to what the frustration view claims, not all frustrated desires decrease welfare. Therefore, appealing to conditional desires is no help for the frustration view.

## 4.4 The Aversion View

So far I have discussed the privation view and the frustration view. I argued that there are a number of significant problems with each, and while it would be too quick to say that those problems are insurmountable, we can tentatively conclude that neither option is likely to succeed. The best strategy, therefore, is to keep looking in the hope that we will find something more promising.

Kagan describes another possibility for the desire theorist (2014, p. 270). So far, the views we have considered involve only one psychological attitude—desire. Importantly, this attitude is positive. One way of defining this is to say that to have a desire for something means that one desires that the thing obtains, that one wants to live in the world in which the desire has been satisfied. (This is true even when we desire that something not obtain. If I am allergic to shellfish, I want to live in a world where my meal does not contain shellfish.) According to the new option we are considering, there is another class of psychological attitude, but it is negative. Following others, Kagan calls these aversions. On the aversion view, we can have negative attitudes toward the satisfaction of certain states of the world, the satisfaction of which constitutes an intrinsic bad.<sup>7</sup> The total theory includes desires and aversions, so when the former are satisfied we are made better off, and when they are frustrated we merely fail to receive an increase in well-being. When the latter are satisfied we are made worse off, and when they are frustrated we merely fail to receive an increase in ill-being.

That desires have a positive and negative form is well represented in the literature on the nature of desires. This is to claim that desires reflect related but distinct psychological attitudes (positive and negative) and is not a claim about their evaluative (i.e., prudential) nature. That desires and aversions are different psychological attitudes (but counterparts) is affirmed by some philosophers of

---

<sup>7</sup>This language can be confusing. I am using 'satisfaction' in the sense of 'obtains in the world', so to satisfy an aversion is for it to be fulfilled, which means that one is made worse off.

mind and neuroscientists, including Timothy Schroeder (2004). I am unable to discuss or evaluate Schroeder's complete view, but it will help the present discussion to describe his approach. First, by 'desire' Schroeder includes a broad category of wishes, wants, desires, and goals (2004, p. 5). He defends the Reward Theory of Desire, according to which

To have an intrinsic (positive) desire that  $p$  is to use the capacity to perceptually or cognitively represent that  $p$  to constitute  $p$  as a reward. To be averse to it being the case that  $p$  is to use the capacity to perceptually or cognitively represent that  $p$  to constitute  $p$  as a punishment (2004, p. 131).

Framing desires in terms of reward and punishment can help us make sense of the intuitive distinction between the psychological states of desire and aversion.<sup>8</sup> Psychologically, desiring something is not the same as being averse to its contrary, nor is being averse to something the same as desiring its contrary. Schroeder gives the example of being averse to Adam's lateness (2004, p. 132). It does not follow from this aversion that one has the positive desire for Adam to be on time; the reason is that the desire and the aversion are valued differently. The person averse to Adam's lateness will be relieved instead of delighted when Adam unexpectedly arrives on time, while the person who desires that Adam be on time will be pleased when he shows up. (What about versions of the desire theory which hold that experience is not required to be made better off? In these cases, all such theories need to stipulate how the agent would respond upon discovering that the desire/aversion is satisfied. It is unnecessary for the agent to actually experience the satisfaction.)

Neil Sinhababu does not endorse Schroeder's reward theory, but he agrees that desires have both positive and negative flavours. These flavours are based on the phenomenological nature of each type:

Some desires, like the desire for a delicious meal, give us a delighted happy feeling when we find that we can satisfy them and an unpleasant feeling of disappointment when we discover that we cannot. Others, like the desire not to miss one's flight, give us the pleasure of relief when we find that we can satisfy them and an unpleasant feeling of anxiety or dread when we discover that we cannot. This gives us reason to divide the category of desire into two subcategories, positive desire and aversion (2009, p. 470).

Another clear example of this distinction occurs when we act out of a feeling of obligation, which is often best characterized as an aversion to not satisfying our obligation rather than a desire to satisfy it. Discharging an obligation brings relief, not delight, so the experience of looking forward to discharging an obligation is dissimilar to a positive desire, such as looking forward to a party. What motivates us to discharge the obligation is not that doing so will be particularly enjoyable, but that *not* doing so will be unpleasant (2009, p. 476). Finally, consider the emotional experience of seeing someone else suffer. While in ordinary language we might say that we desire to relieve such suffering, in the language of Sinhababu's aversion view, it is more accurate to say that we are averse to the suffering of others. Most of the time, seeing people not suffer keeps our emotions unchanged, but seeing someone suffer fills us with anxiety and an urgency to help the individual (2009, p. 487).<sup>9</sup>

---

<sup>8</sup>The specifics of the reward theory of desire are unimportant here; for my purposes I am using Schroeder's theory to motivate aversions. Rewards can involve pleasure (and punishment displeasure), but their effects are broader, including emotional, behavioural, and non-emotional psychological effects (2004, p. 40).

<sup>9</sup>I am not committing to either Sinhababu's or Schroeder's psychological claims. Rather, I am describing them merely to show that there is support for the dual-psychological model among philosophers of mind.

Bykvist takes a related approach to Sinhababu, though whereas Sinhababu is only discussing the nature of desire and aversion, Bykvist is discussing their prudential value (2009, p. 40–41). He says that a properly formulated desire theory involves favouring and disfavouring, where the former involves positive orientation toward some object in terms of emotions, feelings, actions, and evaluative responses. Disfavouring involves negative orientation. Given that we can be positively oriented toward something and neutral toward its unfulfillment (and negatively oriented but neutral to unfulfillment), the frustration view about the nature of desire is false.

Bykvist adds the further claim that only positive desires count toward welfare, and gives our attitudes towards headaches as an example (2009, p. 40). We desire that we do not have headaches, but this is not to say, necessarily, that the absence of a headache constitutes an intrinsic good. Rather, having a headache is bad and having no headache, while comparatively better, is not intrinsically good on his view. His justification is that the valence of the attitude matters:

But it seems much more sensible to say that the satisfactions of anti-headache desires are neutral for you, since, if you are like me, you take a neutral attitude towards not having a headache, and a negative attitude towards having a headache. Therefore, a properly formulated desire theory should say that it is good for you to get what you favour. Roughly put, to favour something is to be positively oriented towards it in your actions, emotions, feelings, or evaluative responses. So, if you have a positive attitude towards something, you tend to be motivated to bring it about, be glad and happy when you think it obtains, have pleasant thoughts about it, or see it in a good light (2009, p. 40).

It is not entirely clear what Bykvist means here. A possible difference between my view and Bykvist's is that mine allows for negative desires, whereas it is unclear if Bykvist means to exclude them. His headache example might just be one case where, rather than being properly described as a desire for not- $x$ , it is better described as an aversion to  $x$ . But if he means that all forms of negative desires are actually aversions, then I disagree. I think we can have proper desires for negative things, where their fulfillment means that we are made better off. For example, suppose my favourite baseball team gets eliminated from the playoffs by the team I hate the most, which advances to the finals. In the finals, I might desire that my hated team loses, where the fulfillment of that desire makes me better off on the desire theory. This is not because I am really desiring that their opponent win—suppose I have no attachment to the other team—but rather because I *really* want my hated team to lose.

One might worry that framing well-being in terms of desires and aversions is equivalent to the claim that having a desire for something entails an aversion to its negation. On this view, having a desire for  $p$  means that one is averse to not- $p$ . Framed in this way, the aversion view is no different from the frustration view, which holds that desiring something implies desiring that its negation does not obtain. Of course, if the aversion view collapses into non-satisfaction then we have no reason to prefer one over the other.

But this is not what the aversion view is claiming. It is possible for a view to claim that a frustrated desire constitutes an intrinsic bad every time. That is the frustration view. The aversion view, in contrast, does not claim that frustration entails the presence of an intrinsic bad. Instead, it holds that when a desire is frustrated it simply fails to contribute any good. One's well-being level stays the same, whereas it would have gone up had the desire been satisfied. In order to suffer an intrinsic bad, one must have an aversion to the satisfaction (i.e., occurrence) of the thing that obtains.

Therefore, on the aversion account there are five relations one can have toward some outcome: one can desire it without being averse to it not obtaining, be averse to it without desiring that it not obtain, desire it and be averse to it not obtaining, be averse to it and desire that it not obtain, or have no attitude toward it.<sup>10</sup> As I described above, according to this view, one is made intrinsically better off only if a desire is satisfied and intrinsically worse off only if an aversion is satisfied. The same holds for negative desires. If I desire not- $p$  and it obtains, then I am to that extent intrinsically better off. I must possess a corresponding aversion to  $p$  in order to be made worse off by the fulfillment of  $p$ , otherwise I simply fail to gain an increase in well-being if  $p$  obtains. It is possible that aversions accompany most negative desires (although I doubt it). This is an empirical claim, not a prudential one. The important point for this discussion is that being averse to something and desiring its negation are logically distinct.

The principal difference between the frustration view and the aversion view is that the latter claims it is possible to desire something without being averse to its negation (or averse to something without desiring its negation). Desires and aversions are distinct psychological attitudes, and we can make sense of them independently. If I go for a hike, desire to see a moose but do not, then the lack of moose does not constitute an intrinsic bad for me.<sup>11</sup> Rather, the absence of moose represents one way that the hike could have gone better but did not. Conversely, if I am averse to seeing a bear and I finish the hike bear-free, this does not entail that I have gained an intrinsic good. It means only that I was not made worse off in that specific way. It would be implausible to claim, as the frustration view does, that one suffers an intrinsic bad whenever a desire is frustrated. This logical independence means that the aversion view does not collapse into the frustration view.

Although aversion and frustration are distinct, the logical independence of the two states points to a new potential problem. Kagan describes the issue as follows:

[I]f desire and aversion are indeed logically distinct psychological attitudes, then as far as I can see, nothing rules out the possibility that one might have both a desire for  $x$  and an aversion to  $x$ —indeed both a desire and an aversion to the very same feature of  $x$ —at one and the same time. And if one did, then if it should turn out that  $x$  does obtain, then that very fact will simultaneously be both intrinsically good for you and intrinsically bad for you (2014, p. 270).

Kagan goes on to say that this result makes the theory implausible (but not necessarily unacceptable). This is the only objection he offers against the aversion view.

Although I see the worry, it does not strike me as the serious problem Kagan believes it is. Such cases are discussed by Patricia Marino, who defines the sort of case Kagan has in mind as one of *valuational inconsistency*, which occurs when two valuations essentially conflict: that is, there is no possible world in which the two valuations can both be fulfilled (2011, p. 45). Far from being an implausible result, Marino convincingly shows that such cases are psychologically commonplace. One type is the sort that Kagan describes, of desiring  $x$  and being averse to  $x$ . An example of this is someone who desires an affair while also being averse to it. We might suppose that the reason for her aversion—that she knows it is wrong and it will risk her present relationship—is exactly the reason why she finds it exciting. There can also be cases where two desires are incompatible, which is not the same as Kagan’s point but might raise

<sup>10</sup>Again, an awkward feature of the aversion view is the terminology, especially the strangeness of saying that the frustration of an aversion is neutral.

<sup>11</sup>Provided that certain psychological conditions obtain. It needs to be the case that I would have a neutral response to the lack of moose, which is possible.

related concerns. Marino gives a case involving a friendly rival as an example (2011, p. 46). Someone might desire that his friendly rival win an honour because he is a good friend, while also being averse to his friend's winning because they are rivals. The result in both examples is inevitable disappointment: "For this agent, no matter what course of action he chooses and no matter how things turn out, there will always be a kind of left-over, a something-lost" (2011, p. 66).

The aversion view gives the same result. Intuitively, how do we think someone will feel if she follows through with the affair? Presumably, she will find it exciting, but will also feel guilt and disappointment in herself. In other words, in one sense she will be made better off, and in another sense she will be made worse off. (We need not settle the specific amounts of each for the purposes of this case.) So the aversion view gives the correct answer here. Either way, she will benefit in one way and be made worse off in another. The aversion view says that she will be made worse off by satisfying her aversion and better off by satisfying her desire. If the desire and the aversion are of similar strengths then, according to the aversion view, the good cancels out the bad of the act, leaving her total well-being at about the same level as before she had the affair. That strikes me as the correct result.<sup>12</sup> Finally, note that if this is a problem for the aversion view, it is also a problem for the frustration view: nothing logically rules out desiring both  $x$  and  $\text{not-}x$ .<sup>13</sup> Kagan's claim is that simultaneous desires and aversions cause a problem for the aversion account. Given this short discussion, having conflicting desires need not be a problem.

The aversion account provides the correct answer in other cases. Recall the case of Hannah, who has the persistent desire to have children. I claimed that we can make sense of the claim that she is made non-comparatively worse off when she is told that she cannot conceive. This, I suggested, is not the mere absence of the good she would have obtained by having a child, but rather an intrinsic bad. (Or at least it makes sense to speak in terms of intrinsic badness regarding this case, such is the significance of her loss.) The frustration proponent might claim that the aversion view is at risk of collapsing into the frustration view because we can make sense of Hannah's well-being loss without appealing to aversions—this is what the frustration account successfully does in these cases—so we need to say how the two views are distinct.

Despite these initial concerns, the aversion view has a straightforward explanation for the case which is distinct from the frustration view. Hannah has the desire to have children and is averse to childlessness. Therefore, being unable to conceive means she suffers an intrinsic bad. Note, however, that it did not have to be this way. Hannah could have had the desire for children without the corresponding aversion, in which case she would only have failed to gain any benefit by her inability to conceive. But, as it happens, she was averse to the frustration of her desire, so she is made intrinsically worse off.

I suspect that the initial attractiveness of the frustration view is due to our possessing aversions to the frustration of many of our strongest desires. The badness of wanting to have children but being unable to is explained by the strength of our aversion to its unfulfillment. In these cases, the frustration view and the aversion view give the same answer, which makes the former equally attractive some of the time. However, not all cases of desire involve a corresponding aversion. It remains an open question how common the desire–aversion pairing is—perhaps most of our desires take this form, although I see no reason to think so—but it is enough that the possibility for desires without corresponding aversions exists to make the frustration view untenable.

<sup>12</sup>If it is not the correct result, the view can accommodate this by claiming either that the strengths must have been different or that there is an asymmetry between desires and aversions.

<sup>13</sup>Kagan also notes this problem (2014, p. 271).

It is also an open question how common it is to have aversions without corresponding desires. Suppose that I am about to have some cavities filled, and I am strongly averse to feeling pain. It does not follow that as the cavities are being filled and I am pain-free that I am receiving some intrinsic goodness from this fact. Rather, I have merely failed to be made worse off by the pain. Such cases strike me as commonplace, which means that, once again, the aversion view succeeds where the frustration view fails.

There is one final problem to consider. I argued that the privation view faces what I call the additive problem, according to which adding failed desires makes the individual no worse off, which I suggested is unintuitive. The aversion view also faces the additive problem because, just as with the privation view, failed desires do not lower one's well-being. Rather, they merely fail to contribute intrinsic value.

This is not as serious for the aversion view as it was for the privation view. The additive problem for the privation view is worse because it entails that well-being bottoms out: i.e., that once one reaches zero one cannot be made any worse off, even as failed desires continue to add up. The aversion view does not have this implication. One can continue to be made worse off on the aversion view so long as aversions are satisfied, which means that, in principle, there is no point at which well-being bottoms out.

It is possible, however, for the aversion view to deny that we can keep being made worse off to infinity. Perhaps at some point ill-being does in fact bottom out, or perhaps it gets asymptotically lower. What matters is that, whereas the privation view implausibly claims that well-being bottoms out at zero, so that there are no intrinsic bads, the aversion view holds that, if our well-being can indeed stop getting worse, this does not occur at zero.

#### 4.4.1 Extensional Equivalence

Sumner gives the following objection to the aversion view (unpublished, p. 12–14). While desires and aversions are distinct attitudes, desiring not- $p$  and being averse to  $p$  are extensionally equivalent in that they are satisfied by the same state of the world. Worse still, for each of these pairs, there is a positive desire that is also equivalent. So for the case of Hannah, there are three ways of describing her attitude, all of which are satisfied by the same state of affairs:

1. D (Hannah has children.)
2. D ( $\sim$ Hannah is childless.)
3. A (Hannah is childless.)

Sumner says that “Whatever form of expression we prefer, the fact remains that Hannah will be satisfied by the same state of affairs (Hannah has children) and frustrated by the same state of affairs (Hannah is childless)” (unpublished, p. 13). Given this result, Sumner concludes that we gain nothing by introducing aversions.

I agree that these three outcomes are extensionally equivalent, but I disagree that this shows anything problematic about the aversion view. (1) and (2) are, in fact, the same. While there might be a difference in how the agent formulates these two desires, according to the aversion view their satisfaction comes out the same. I see no problem there. What about (1) and (3)? Here, the state of the world is the same, but the aversion view evaluates them differently. Desiring children and being averse to childlessness are not the same mental state, so they deliver different value claims. The entire point of the aversion view is that one can affirm (1) without also affirming (3). So even though the two states are extensionally

equivalent, they are not equivalent in value. Nothing precludes Hannah from saying “I desire to have children, but I will not mind if I am unable to.” To deny this, it seems to me, is simply to beg the question against the aversion view.

The response regarding (2) and (3) is similar. Desiring  $\text{not-}p$  and being averse to  $p$  produces different movements in value according to the aversion view. If one gets  $\text{not-}p$  then one is made intrinsically better off. (This is why (1) and (2) are value-equivalent: desiring  $p$  and  $\text{not-not-}p$  come to the same thing.) But failing to get  $\text{not-}p$  is not an intrinsic bad on the aversion view—it simply fails to produce the increase in value satisfaction would have. Being averse to  $p$  means that the occurrence of  $p$  constitutes an intrinsic bad. So in one case—the satisfaction of a desire for  $\text{not-}p$ —one’s welfare increases, whereas in another case—the satisfaction of an aversion for  $p$ —one’s welfare decreases.

Returning to Schroeder’s description of desires is helpful here. He gives reasons against adopting Sumner’s suggestion that desires be interpreted as desires for  $p$  while aversions are desires for  $\text{not-}p$ . (If Bykvist is claiming that negative desires are equivalent to aversions, then this is also a response to his view.) On Sumner’s view, having an appetite for pie is having a desire for  $p$ , while being averse to eating pie is having a desire for  $\text{not-}p$ . The problem with this model is that it cannot distinguish between two different psychological attitudes, attitudes given which we will have different reactions to the satisfaction of their objects (2004, p. 26). In other words, moving around the negation sign will keep the outcomes extensionally equivalent, but it will not capture everything we want from a theory of prudential value. Consider a different case from Schroeder. How should we frame the statement that a bodyguard desires that his charge remain safe from harm? On this view, he has a desire for  $p$ , but the bodyguard might not actually have a desire for the charge’s safety, even though he is highly averse to his charge being harmed because it is his professional duty to protect him (2004, p. 26). Framing desires and aversions as distinct, both psychologically and normatively, is the best way to avoid these problems. According to Sumner, “nothing seems to be gained by introducing the negative element” (unpublished, p. 13). I deny this. We have gained the ability to explain different changes in value in a way that more accurately represents our mental states and the best view of prudential theory.

## 4.5 Conclusion

I have surveyed three possible views: the privation view, the frustration view, and the aversion view. In the end, the aversion view is the most plausible. Although I have shown that the desire theory can account for ill-being, it is an entirely different question whether or not the complete well-being account of the desire theory is the best option compared to non-desire theories. I have said nothing to evaluate the complete view. An interesting feature of this discussion is that there has been a relative paucity of asymmetries. The aversion view is structurally symmetrical to the desire theory, and while there might be prudential asymmetries between the good and the bad, it makes more sense to base those values on the strengths of the desires and aversions. Because of this, the aversion view does not accord with my intuitive asymmetry that it is easier to be badly off than well off. Empirically, it might be the case that we have more aversions than desires, but this is neither obvious nor part of the prudential account.

# Chapter 5

## Objective Theories

### 5.1 Introduction

There are two conditions that must be fulfilled in order for a theory to be objective in the way I am using the term. First, any theory is objective if it holds that there are goods and bads that make us well or badly off regardless of our attitudes toward them. Another way of putting this point is that objective theories need not have a resonance condition: If I possess a good then I am to that extent better off even if I do not care about possessing that good, or even know that I have it, and likewise for bads.

Second, the list of goods cannot consist only of pleasure or satisfied desires. The reason for this second criterion is that, without it, hedonism could be objective according to the first criterion, so long as it holds that pleasure is the good and that possessing pleasure makes one better off regardless of how one feels about it. The ascetic who is averse to pleasure but nevertheless gets it is still better off on this interpretation of hedonism.<sup>1</sup> Hedonism could therefore be interpreted as an objective theory with only one good (pleasure) and one bad (pain), which is a conclusion I want to avoid, mostly for simplicity of the discussion.<sup>2</sup> I discuss hedonism in Chapter 2, so in this chapter I will focus on goods and bads other than pleasure and pain. All of the same points apply to the desire theory. If a theory says that the satisfaction of desires make me better off regardless of how I feel about them, then this is the desire theory instead of an objective theory.<sup>3</sup>

In this chapter I first consider two objective goods, autonomy and knowledge, and whether they have counterpart bads. I then consider Richard Kraut's theory, developmentalism. Kraut's book *What Is Good and Why* is rare for having an extended discussion of ill-being. However, I show that developmentalism does not produce a convincing account of intrinsic prudential badness. I then explain why pure perfectionist theories are all implausible due to their inability to account for ill-being. Because of the length of the discussion, I have a separate chapter on achievement and its contraries in Chapter 7.

---

<sup>1</sup>Some versions of hedonism will give different answers to this case.

<sup>2</sup>This is a purely taxonomic distinction. Guy Fletcher (2013) argues that hedonism *is* an objective theory because it claims that pleasure is good for us, even if we do not want the pleasure.

<sup>3</sup>That there is overlap between these theories is one of the shortcomings of Parfit's list (see Chapter 1), but for my purposes these problems are insignificant.

### 5.1.1 The List

A significant difference between the objective theory on the one hand and hedonism and the desire theory on the other is that the former has no agreed upon list of goods in its basic form. In principle, anything could be on the list of goods and bads. What the objective theory has in common with the others is that there has been little discussion of what makes one badly off. Given the variety of potential versions, my strategy in this chapter is to review two well-known items and consider how they might be extended to include ill-being.

Objective theories are often classified in two ways. So-called *objective list theories* are enumerative; they describe the items on the list, the justification of which is usually intuitive. (This is no different from hedonism and the desire theory, as all theories of welfare will bottom out in their explanatory ability.)<sup>4</sup> In contrast, *perfectionist* theories are explanatory. They do not merely list the objective items, but also explain why those items are present. For the perfectionist, items are on the list because their possession perfects our nature. As with other theories I have considered, objective list theories can (and should) in principle give an account of ill-being but usually do not. For reasons I explore below, perfectionism is a special case. Pure perfectionism necessarily denies the existence of intrinsic bads because one cannot perfect oneself a negative amount. I will argue that this makes pure perfectionism implausible.

## 5.2 Autonomy

Autonomy is often mentioned in lists of objective prudential goods. For example, although John Stuart Mill appeals to the instrumental value of autonomy as a source of pleasure, he suggests it also has intrinsic value when he claims that “He who lets the world, or his own portion of it, choose his plan of life for him, has no need of any other faculty than the ape-like one of imitation. [...] It is possible that he might be guided in some good path, and kept out of harm’s way, without any of [the faculties associated with autonomy]. But what will be his comparative worth as a human being?” (1859/1975, III. 4). In *Liberalism, Perfection and Restraint*, Steven Wall argues that “It is intrinsically good for people to make their own choices about how to lead their lives” (1998, pp. 129–30). Similarly, Will Kymlicka claims that autonomy is a precondition for living a prudentially good life (1989, p. 12).

Here I am discussing personal autonomy, or self-governance. Some authors interpret autonomy as the ability to choose freely (e.g., Hurka 1987, p. 363), where autonomy increases the more options one has, while others have in mind self-determination (e.g., Kymlicka 1989), where autonomy increases as one endorses projects or chooses for oneself without interference or limitation. Since everyone faces interference or the restriction of options of one sort or another—e.g., cognitive biases, cultural influence—each theory must explain the sorts of things that thwart autonomy and the impact such thwarting has on prudential value.

In theories of welfare, autonomy can contribute value in one of two ways. The first is that autonomy is a condition for other forms of value. Without autonomy, projects we pursue no longer contribute welfare. This view is endorsed by, among others, Kymlicka (1989, 2002) and Wall (1998, pp. 130, 160).<sup>5</sup>

<sup>4</sup>Where the explanation for each theory is located is, of course, controversial. For example, according to Roger Crisp, who uses the enumerative–explanatory distinction, for hedonism, what makes something a substantive good is its property of being enjoyable (2006, p. 103). We still might reasonably wonder why pleasurable things are the only good though.

<sup>5</sup>There is another type of conditional view that goes the other way, holding that the value of an activity is conditional for autonomy (e.g., Raz 1986). On this view, autonomy has value only if the options available to the individual are good ones.

The second sort holds that deciding how to live one's life makes one better off regardless of what one actually decides to do. The value of autonomy is independent of the value of other goods, so it is the decision to pursue a good on one's own (i.e., self-determination) that gives such pursuit value, regardless of what the object is. I might believe that Christianity is the one true religion and pursue it, but then later in life change my mind and become an atheist. The truth or other value of the belief system is not what matters for autonomy; it is that I pursue the belief freely. On this view, obtaining something valuable *without* autonomy need not take away from the value of the object, but autonomous pursuit is more valuable when all else is equal. This view is defended by Mill (1859/1975, III).<sup>6</sup>

### 5.2.1 The Conditional View

Kymlicka argues that “My life only goes better if I am leading it from the inside, according to my beliefs about value” (2002, p. 216). On this view, autonomy is a condition for all other value. Without it, life has no value for the agent. Kymlicka's claim is too strong. This is clear when we think about people who get achievements while acting purely for instrumental reasons. On Hurka's definition, a pure professional athlete is one who plays only to make money with no love of the game (2001, p. 189). When such people achieve greatness—breaking records, winning major championships, etc.—their achievements are still valuable and therefore increase welfare, even though the athletes themselves do not believe that their playing has value. Kymlicka's view denies this.

Looking at paternalism—restricting someone's autonomy for their own sake—also reveals problems with Kymlicka's view. He claims that paternalism “may succeed in getting people to pursue valuable activities, but it does so under conditions in which the activities cease to have value for the individuals involved” (2002, p. 216). His idea is that activities, such as prayer, are only valuable if the person believes for himself that they are valuable. Again, this is too strong for two reasons. One is that, while extreme coercive paternalism might lead to intrinsic bads by, for example, causing us suffering because we are unable to choose for ourselves, other types of paternalism do not have such a strong effect. A physician who treats me without obtaining my informed consent robs me of the value I would have received had I made an autonomous choice, and certainly robs me of the instrumental benefits of choosing for myself, but the increase in welfare I get from having the correct treatment—e.g., the happiness at knowing that I will be able to return to my normal life—remains. The second reason is that paternalism is quite common, which means that on Kymlicka's view, many activities, and therefore many lives, lack value. Such activities might in fact lack the value of autonomy, but it is false to say they lack all value, and people generally do not treat paternalism as having such a dramatic effect.

Sometimes people argue that paternalism is a bad strategy instrumentally because outsiders cannot choose better than the agent can for herself. Kymlicka's view gives an unacceptable result, which is that paternalism of this form is not just impossible in practice, because we will never choose what is actually best for the agent, but impossible *in principle*, because trying to improve welfare in this way will always lack the agent's own endorsement.<sup>7</sup> Claiming that this form of paternalism is self-defeating is false, and

<sup>6</sup>Autonomy, liberty, and freedom are related concepts that I will not attempt to distinguish here. Though Mill is generally discussing liberty, his views are compatible with a theory of autonomy. For instance, he promotes the “Greek ideal of self-development” (1859/1975, III).

<sup>7</sup>The ‘endorsement constraint’ holds that one must believe for oneself that an activity is valuable for it to be prudentially valuable. Kymlicka puts it as follows: “If I do not see the point of an activity, then I will gain nothing from it” (2002, p. 216). Others have criticized this constraint, holding that valuing an activity is unnecessary for the activity to have value. Hurka, for instance, calls the approach “impossibly high-minded” (2001, p. 187). On Kymlicka's view, it might still be the case that forms of paternalism which are coercive but promote values the agent has already endorsed, such as with

there are many cases that show this, including someone else deciding who you will marry. While this might not always lead to the best outcome—i.e., sometimes we do know what is best for us—it is obvious that it is *possible* for someone else to do a better job than we can for ourselves.

## 5.2.2 The Non-Conditional View

I now turn to autonomy as a non-conditional good to consider if it has a counterpart bad. We have a name for this counterpart, heteronomy, but it is unclear what it amounts to and what its value is. Autonomy and heteronomy are different from, for instance, pleasure and pain. While the absence of pleasure is not equivalent to pain, an absence of autonomy is equivalent to heteronomy: losing autonomy just is gaining heteronomy. Structurally, autonomy and heteronomy have no neutral middle the way pleasure and pain do, so there is no easy mapping to match the way pleasure contributes well-being and pain contributes ill-being. While there might be a state that is neutral—the goodness of the autonomy is offset by the badness of the heteronomy—it does not clearly follow from the structure, nor does it follow that at some point enough heteronomy is intrinsically bad.

The specific structure of the autonomy–heteronomy continuum will depend on the particular version of autonomy. On the non-interference view, the scale is capped at the top and bottom. There is a maximum amount of autonomy one can have as well as a minimum. At the top, once one is completely free of external coercion, one has complete autonomy (or, positively, full autonomy occurs once one is completely self-governed). At the bottom, one is completely heteronomous by having no control over one’s life by being completely under external control. This pair can be represented on a single axis, where we could assign ‘100’ to complete autonomy and ‘0’ to complete heteronomy. However, describing the pair in this way does not entail anything about their prudential values. In terms of value, there are three main possibilities. First, autonomy is intrinsically valuable and heteronomy is worth zero. This would match the structure just described, so the value scale goes from +100 to 0. Second, autonomy is worth zero while any lack of autonomy contributes ill-being. This would mean the value scale goes from 0 to –100. Third, autonomy has some positive value while heteronomy has some negative value, which means there is a neutral point where the autonomy and heteronomy cancel out, so the scale goes from, say, +50 to –50. Because I am assuming that autonomy is intrinsically good to see if it has a counterpart bad, in what follows I will not discuss the second option.

The scale according to the options view is capped on the bottom but not at the top. This is because one can have zero autonomy by having zero options, but in principle there is no limit to the number of options one can have, so autonomy can increase forever. Although this is structurally different from the self-determination view, the same points apply. That is, this structural description has no implications for prudential value. It could be that there is some point where enough lack of autonomy becomes intrinsically bad, but there is no natural point where this should occur, so defenders of such a view must defend their approach in a way that is different from, say, the pleasure–pain continuum.

Testing for autonomy is difficult because there are other factors relevant for well-being. Even if autonomy is a great good, it is not the only one. The conditional view claims that autonomy is only a component of the good life, such that on its own it contributes nothing. For the non-conditional view, the problem is that many cases do not arbitrate between the view that heteronomy is intrinsically bad and the view that less autonomy becomes less good without going negative (i.e., losing autonomy is

---

libertarian paternalism, which involves ‘nudges’ toward doing what we want to do but are too weak-willed to actually do, are unaffected by this argument.

just the loss of a good). Imagine John, who is forced by his parents to become a philosopher. This thwarts his autonomy both because he has fewer options available and because he did not freely choose his path. Nevertheless, suppose that he is content with his life. Indeed, suppose that he is as happy as he would have been had he freely chosen philosophy. (We are holding his happiness fixed so that it does not influence our beliefs about John's welfare.) The problem is that the difference in John's welfare is compatible with the view that heteronomy is intrinsically bad and the view that less autonomy becomes less good without going negative, as even if the forced profession adds some negative value, the total could still be positive because of other factors. So what we need is a way of determining if there is negative value or just the loss of positive value.

One result that would demonstrate the intrinsic badness of heteronomy is if, even given the intrinsic goods from philosophy, including his happiness, people had the intuition that John's welfare is negative overall. Some people might have the intuition that the total welfare level becomes negative in this sort of case and other cases, such as the happy but non-autonomous Epsilons in *Brave New World*. Their autonomy is close to zero, though they are generally happy. Yet another case is Truman in the film *The Truman Show*, who is happy but significantly manipulated (and has fewer options because, e.g., he cannot leave his town). Assuming that the other sources of prudential value remain, a negative total welfare level would entail that heteronomy is intrinsically bad. I do not have this intuition. The lives of the Epsilons, though far from perfect and lacking in many types of value, are still worth living, as is Truman's. But my intuition here does not mean that heteronomy is not intrinsically bad. After all, it could be that the loss of autonomy is not merely a loss of something good, but a gain in something bad. Contrariwise, when Truman discovers at the end of the film that he has been manipulated, it might be that he is not just gaining a good he previously lacked, but also *losing* a bad he previously possessed. The issue is that this sort of case cannot tell us which option is correct.

The cases of John, the Epsilons, and Truman involve pleasure while lacking autonomy. They are unhelpful for settling the value of heteronomy. The same problem arises in cases of lacking autonomy with pain. For example, drug addiction is bad for the addict, but it is hard to separate the lack of autonomy from other bad things involved in addiction, such as the pain of craving the drug or the ways it tends to destroy health, relationships, and finances. Similarly, being indoctrinated into a cult usually comes with the same set of negative effects as addiction. So just as with good effects such as being happy, bad effects such as suffering mean we cannot hold fixed the disvalue of heteronomy, so this sort of case is also unhelpful.

That leaves one other type of case we can try. We can make everything else neutral and then adjust autonomy. The problem with this approach is that it is either impossible to imagine or unhelpful at motivating intuitions. We can try to imagine someone with no enjoyment, achievements, or knowledge, and then adjust only autonomy, but I simply have no clear intuitions about this sort of case. I can grant that the value of such a life is higher if it is autonomous, but I have no sense of where on the welfare scale such a person falls as autonomy goes up and down. So this sort of case is similarly unhelpful.

Of course, this is an unwelcome result, for it means that all of the obvious tests for the intrinsic badness of heteronomy fail to produce a clear answer. That leaves us with little to go on. My intuition is that heteronomy is not intrinsically bad, so there is no negative counterpart to autonomy. Given that perfect autonomy (on the self-determination view) is so difficult, if not impossible, to obtain, this approach best accords with our intuitions that an autonomous choice is better than a less autonomous choice, and as autonomy decreases eventually there is no welfare contribution because our action is no

longer self-governed to any degree.

On my view, a complete lack of autonomy will at most have zero value but will not be negative (in terms of the autonomy). One's welfare drops as one becomes less autonomous, until a threshold is reached and there is no welfare contribution from autonomy. A complete lack of autonomy is compatible with a life of negative value, as such a lack can be accompanied by intrinsic bads such as pain, but on its own, lacking autonomy is neutral, not intrinsically bad. One intuition that leads me to this conclusion is my sense that most of the ill-being work in cases of indoctrination and slavery is being done by other bads, including suffering. In cases of happy heteronomy, such as Truman's, my intuition is that his life is still fairly valuable, showing that, while lack of autonomy might still be intrinsically bad, it is not a significant intrinsic bad. He would benefit from autonomy, but his lack of autonomy is only comparatively bad for him. I have no response to someone who claims that there is a mild intrinsic bad involved in the loss of autonomy, though I will note that there is little difference in implication between the two options. Additionally, I think a major part of the intuition that autonomy is bad can be explained instead by appealing to its wrongness.

### 5.2.3 Wrongness

In this dissertation I have focused entirely on goodness and badness while saying nothing about rightness and wrongness, which concerns appropriate action or what we ought to do. I want to briefly discuss wrongness here because I think it partly explains our intuitions about heteronomy. Indoctrinating someone is a serious wrong, and this point can help make sense of the repulsion we feel toward thwarting autonomy. This can be so even though thwarting autonomy only robs the individual of the good of autonomy she would have otherwise received, instead of making her intrinsically badly off. This is often revealed in the bioethics literature on informed consent. For example, Benjamin Freedman claims that "Perhaps the worst which we may do to a man is to deny him his humanity, for example, by classifying him as mentally incompetent when he is, in fact, sane" (1975, p. 32). I take it that Freedman's point is that it is *wrong* to rob people of the ability to choose for themselves. To deny them this ability is to treat them wrongly. If he means that the worst thing we can do to someone prudentially is thwart autonomy, then this, it seems to me, is plainly false.

The point that indoctrination is wrong and not necessarily intrinsically bad also makes sense of paternalism. Interfering with someone's autonomy for her own benefit might be justified if she is only missing out on the value of the autonomy instead of suffering an intrinsic bad. Of course, it might still be wrong to act paternalistically, even if there is no substantial loss of welfare, but this assessment is more plausible than Kymlicka's claim that paternalism completely robs the activity of welfare. Similarly, to fail to get informed consent from patients is wrong because it robs them of the value of choosing for themselves, but when a doctor chooses for a patient it need not make a significant difference to how well off the patient is. Indeed, I see little difference in terms of intrinsic value. Failing to get informed consent, can, however, make a significant difference for instrumental value by thwarting other values the patient has. So the wrongness of paternalism is separate from its badness.

Wrongness also helps answer a different structural question. Does it matter whether the lack of autonomy is due to absence of control (such as being an addict) versus external interference (such as paternalism and indoctrination)? A defender of the options view might hold that having options taken away is intrinsically bad, whereas losing control for other reasons—i.e., where someone else has not taken the options away—is merely the loss of something good. In terms of welfare, I think there is no

difference. The absence of autonomy is comparatively bad relative to more autonomy, but the source of the loss makes no difference to one's welfare. I suspect, once again, that what leads some to say that interference by another is worse is actually a claim about wrongness. Being an addict involves a loss of autonomy only, while being psychologically manipulated by another person involves both the loss of autonomy and the wrongness. It would be a strange position, unlike any other claim about prudential value that I am aware of, to hold that it is the manner of the loss, and not the loss itself, that makes a difference prudentially.

### 5.2.4 Me and Your Life

Some philosophers claim that losing autonomy means that one's life is no longer one's own, and is thus not something that can be evaluated in terms of welfare. Sumner, for example, defends his authenticity condition, which consists of information and autonomy, on the grounds that lacking autonomy means that evaluating values one has not freely chosen means no longer evaluating that life: "If a subject's endorsement of some particular (perceived) condition depends on a factual mistake, or results from illusion or deception, then it is not an accurate reflection of her own underlying values. And if those values have been engineered or manipulated by others then they are not truly *hers*" (1996, p. 174). As I described in Chapter 3, for Sumner, this means that we cannot evaluate how well such a person's life is going because we are no longer evaluating *her life*, but someone else's instead. Concerning autonomy as an objective good, John Harris makes a similar claim when he says that "The point of autonomy, the point of choosing and having the freedom to choose between competing conceptions of how, and indeed why, to live, is that it is only thus that our lives become in any real sense our own" (1995, p. 11).

In one way, this is a valuable point: autonomy concerns self-determination, and when we lack autonomy, we lose control over our lives. But when the claim is interpreted in the way Sumner describes and Harris suggests, it has the flaw I discussed in Chapter 3. If it is the case that we cannot evaluate an element of someone's life because it is not actually part of that life—it is not in the scope of consideration—then that produces the incorrect result by saying that the welfare level becomes undefined. It means that considering the factor is irrelevant, not bad or good. By analogy to their view, if we are trying to establish someone's financial situation, looking at the bank account of some unrelated person is not useful, regardless of how well or badly that other person is doing. But the insight Sumner and Harris are trying to capture is that the irrelevance—i.e., the lack of control—itself is a problem. We should not claim, as Sumner says we should, that losing autonomy means there is less of one's life to assess, but rather that the loss of scope is itself worse (though not intrinsically bad) for the person. This is what my view says. An analogy to my view is the board game Risk, where having full control of all the territory is the goal, and losing any territory to someone else is almost always bad.<sup>8</sup> The correct assessment of losing control is not undefined, but comparatively negative. (That is, there is a loss of value. As on my view about autonomy, in Risk the worst off one can be is zero.)

## 5.3 Knowledge

The investigation of autonomy showed that there is no compelling reason to conclude that autonomy has a negative counterpart. Along with autonomy, many objective theorists claim that knowledge is good, so

<sup>8</sup>Though not always. In Risk, one might intentionally give up some territory for some long-term benefit. Similarly, losing some autonomy might be justified for the long-term benefit of an individual.

we can similarly ask if knowledge has a negative counterpart. I will make the simplifying assumption for the purpose of this discussion that knowledge is justified true belief.<sup>9</sup> One structural difference between autonomy and knowledge is that some of the conditions for knowledge have contraries that are not the mere absence of the positive condition. For example, a belief can be justified, unjustified (contrary), or simply lack justification. Similarly, truth has the contrary of falsity. Is the counterpart to knowledge intrinsically bad, and is there just one such counterpart, or many? There has been some discussion of this point in the literature. With the assumption that knowledge is justified true belief, we can assess each possible combination to determine its value. If justified true belief is the best epistemic state to be in, then there are eleven alternative states one can be in (justified false belief, unjustified true belief, and so on).

Kagan and Hurka give different partial rankings. Hurka claims that there might be positive value in having justified false beliefs. Kagan does not say if justified false beliefs have value; he only says that unjustified true beliefs are better than justified false beliefs, and that “false beliefs in the face of evidence to the contrary” are the worst form (2014, p. 280). Regarding this state, he says that “it seems to me not merely the case that this is the worst epistemic state to be in, it also seems to me to be the case that being in this state is indeed bad for you” (2014, p. 280). On this view, a young-Earth creationist has a false belief in the face of evidence to the contrary, and is to that extent badly off.<sup>10</sup> Further, although Kagan does not claim this directly, it follows from his view that the more evidence one possesses to the contrary, the worse off one is. A climate scientist has more justification for accepting climate change than a non-scientist, so his denial, despite so much contrary evidence, makes his false belief worse. Notice that unjustification comes in two forms. One can have insufficient evidence for a proposition, such as believing that there are other intelligent life forms elsewhere in the universe, or one can have evidence that points in the *opposite direction* of justification, which is what Kagan means by ‘evidence to the contrary’. (We could call the latter ‘anti-justification’ to denote the difference.) A present-day young-Earth creationist or a climate change denier is making a serious error, not because they have insufficient evidence, but because the evidence supports the opposite position. A state of *anti-justified false belief* is the worst state to be in. As Kagan points out, there is no common name for this worst type of anti-knowledge. In Chapter 6 I will argue that the worst type of anti-achievement, which I call bungling, is analogous with this.

Determining the full ranking is difficult. For instance, is it better to have a true belief without justification or a false but justified belief? In some cases, for instrumental reasons we might hope for the former. Consider Parfit’s example of a physician who has to decide on the best medical treatment (1984, p. 25). The objectively right treatment is the one that will in fact work, while the subjectively right treatment is the one which, given the evidence, the doctor has most reason to perform. In such a case, one might prefer a true belief without justification because it will be the one that saves the patient. But which option makes the doctor’s life go better? At least most of the time, I think it is prudentially better to have a false but justified belief. This is because such a belief best accords with reason, and we should want to believe what the evidence tells us instead of believing what is true purely by chance. As it happens, this approach is also best in the long term, as it is better to use evidence to do what is rational instead of using a less reliable method, even though the rational method will sometimes cause errors.

<sup>9</sup>This follows Kagan’s discussion of ill-being (2014, p. 277) but not Hurka’s, which adds an anti-Gettier condition in his discussion of the intrinsic value of knowledge (unpublished (b), p. 4).

<sup>10</sup>Hurka agrees (unpublished (b) p. 11).

Appealing to reason also explains why false belief in the face of evidence to the contrary is the worst form. It is bad to go so strongly against reason, to act irrationally. That there is so much justification in support of the opposite position means that reasonable expectation says one should have the opposite belief. It is the inability to respond to evidence when one is able to that makes it so bad.

Kagan mentions a principle that goes against the conclusion he and Hurka reach. Kieran Setiya proposes that, other things being equal, it is better if the available evidence points toward the truth instead of being misleading (Kagan 2014, p. 281, fn. 5). The idea is that we should want the evidence to lead to the truth because it is more likely to make us better off this way. If the direction of the evidence is out of our control, we should want epistemic luck to be in our favour. This strikes me as correct, but only instrumentally speaking. If we have bad luck and the evidence frequently leads us towards false beliefs then we will, in the long run, be worse off because we will make bad decisions. But this does not capture what is actually intrinsically good about knowledge (and bad about a lack of it). Using reason is valuable; misusing it is disvaluable. Setiya's condition fails to capture this.

Hurka argues that an epistemic state's value also depends on the type of knowledge involved (2010, pp. 218–219). He uses Robert Nozick's discussion of the experience machine to show this distinction. One of Nozick's reasons why life on the experience machine would be less good than life off it is that people attached to the machine are disconnected from reality. This disconnect means that they will have false beliefs about the world and their place in it: e.g., someone might believe that he has just won the World Series, while in reality he is still floating in the tank. Even though these beliefs might be justified by the evidence, their falseness makes machine life lack value.

Hurka argues that life on the experience machine does not involve merely the absence of the good of true beliefs; he claims that it is an evil to have false beliefs about one's current environment, which is what explains the badness of the experience machine. On Hurka's view, the badness of the false belief is greater than the goodness of having correct beliefs about your life (which at most will be a small good and might not be good at all) (2010, p. 218). So, for Hurka, not only is this one type of counterpart bad to the good of knowledge, but the badness of this type of false belief is greater than the goodness of knowledge is good in certain circumstances. Although Hurka is discussing agent-neutral value, the same point plausibly applies to prudential value for the objective theorist.

This asymmetry applies to beliefs about one's own life, but Hurka does not think all knowledge is asymmetrical in this way. False beliefs about one's place in the world have a special badness. Another form of knowledge concerns external facts, including scientific ones. Consider Aristotle, who held false beliefs about physics and biology. According to Hurka, Aristotle's mistaken scientific beliefs are not bad, especially compared to the beliefs of his contemporaries. This is in part because Aristotle's beliefs might have been justified (2010, p. 219). But knowledge about laws of nature does seem significantly good. Therefore, in the case of external knowledge, there is an asymmetry between the good and the bad, but it is the opposite of the asymmetry for internal facts (i.e., facts about ourselves). Whereas it is significantly bad to have false beliefs about ourselves, it is much less bad (or not bad at all) to have false beliefs about the external world. That there are different degrees of badness shows that there is not just one type of negative counterpart, but many.

## 5.4 Kraut's Developmentalism

Richard Kraut's *What is Good and Why* is in the minority of accounts of well-being because it also contains an account of ill-being. Kraut's view differs from some theories of well-being (although not from some other theories of objective welfare) by expanding the range of possible welfare recipients to include all living things and not just humans or sentient animals.<sup>11</sup> However, he gives a specific set of conditions for humans that does not apply to other living things.

Kraut begins by giving an account of 'good for'. On his view, something is good for someone if that thing is well suited or serves the person well (2007, p. 94). This approach is the starting point for Kraut's arguments against the desire theory and hedonism, which are based on the perspective of the individual instead of suitability. Although Kraut endorses a form of what he misleadingly calls 'diluted hedonism', this is not a form of hedonism *per se*, but rather a way of adding a subjective component to his objective theory. According to this view, if something is good for an individual, then it is good in part because she enjoys it, but not *only* because she enjoys it (2007, p. 127). In other words, enjoyment is a necessary but not a sufficient condition for welfare. The thing that is enjoyed is also relevant, and if enjoyment is not doing all the explanatory work, then the theory needs some additional explanation for the goodness of certain states of affairs. (According to Kraut, traditional hedonism requires no such explanation. Pleasure is the only good and no explanation is needed for why it is good.) Kraut's complete view of well-being is hybrid, for he denies that goods such as knowledge are good for us if they are unenjoyed, and later adds the stipulation that not all forms of pleasure are good for the individual (2007, pp. 128–129). A full evaluation of Kraut's theory will take place in Chapter 7, where I discuss subjective-objective hybrid theories, which claim that welfare is taking a positive attitude to an objective good. My purpose here is to assess just his account of the objective side, though I will describe the roles of enjoyment and suffering insofar as they are necessary to understand his theory.

Kraut believes that, because he rejects hedonism, he needs to give an account of why pleasure taken in certain things is good for us. According to his account of intrinsic goodness, well-being involves flourishing, where what it means for something to flourish is defined by its species membership.<sup>12</sup> When something is good for someone, that thing is good because it is part of the person's flourishing. He says that "when we consider the good of any living thing, we should look to the process of growth and development that best suits things of its kind" (2007, p. 136, fn. 4). For humans, flourishing involves both physical and psychological components, and while someone can still be well off even if she suffers a debility in one of these respects, both are important.

Kraut calls this 'developmentalism'. There is no single way for humans to develop, but rather many that accord with the sorts of beings we are. One person can develop her capacities fully by being a chef, while another can do so through computer programming. In either case, the individual develops herself in a way that involves characteristic human capabilities: intelligence, creativity, emotions, sociability, and so on. On Kraut's view, we proceed from the recognition that physical, cognitive, affective, and social skills contribute to flourishing and use induction to determine which activities are best for each of us. The flourishing of other species will be similarly defined by the sorts of things they are.

According to developmentalism, our enjoyment of activities often tracks their value. This is true of friendship, for instance:

---

<sup>11</sup>The full scope is unclear. Kraut includes plants (e.g., orchids on p. 135), but he does not mention simpler organisms such as bacteria.

<sup>12</sup>This is described in chapter III, especially parts 34–37.

[O]ne reason we find this so enjoyable is that being a good friend is in some respects like having a good job: it offers abundant opportunities to be a comforter, a helper, a companion; and when we do these things, we put into play the sophisticated psychological skills that gradually took shape as we emerged from childhood (2007, p. 143).

Similarly, playing sports is enjoyable because they involve both physical and psychological skills, as do non-sport physical activities such as walking (2007, p. 144). Each time an activity contributes to flourishing, enjoyment tends to occur.

However, nothing makes this connection necessary on Kraut's view. There are plenty of possible cases where we develop ourselves according to our species membership, yet we get no pleasure from the result. Suppose I choose to develop my intelligence and spatial skills by learning chess, even though I get no enjoyment from it. In this case, I flourish, but my lack of enjoyment means it does not improve my welfare.<sup>13</sup> Pleasures taken in bad objects are also not good for us. He gives the example of a rapist getting as much pleasure from rape as one normally does from consensual sex (2007, p. 166–168). Hedonism is forced to claim that the rapist's pleasure is just as good for the rapist as pleasure during consensual sex, but the developmentalist can claim that wanting to commit rape and getting pleasure from it is a sign of faulty development. It would be much better for the rapist had he not developed in the way he did; Kraut says that "We should call the rapist's pleasure a bad pleasure" (2007, p. 167).

### 5.4.1 Un-Flourishing

The feature of Kraut's discussion that distinguishes it from nearly every other account of well-being is that he dedicates a considerable amount of space to ill-being (2007, pp. 148–169). He recognizes, as the epigraph at the beginning of this dissertation illustrates, that ill-being does not trivially follow from a theory of well-being: "A theory about what is good is not, all by itself, a theory about what is bad. It can easily seem otherwise, because we might suppose, unreflectively, that what is bad for someone is simply the absence of what is good for him" (2007, p. 148–149). He uses hedonism as an example to show that what is bad (pain) is not simply the absence of the good of pleasure. We cannot assume, in any case, that the absence of the good is equivalent to the bad. (For example, in Chapter 4, I argue that many desire theorists commit this mistake.)

While we cannot assume that the bad is just the absence of the good, Kraut notes that this might be a natural claim to make about developmentalism. What is bad for a plant, for instance, is just whatever amounts to it not flourishing, where that is some type of failure: to grow, to reproduce, to produce foliage, etc. (2007, p. 149). We might then move from claims about plants to claims about humans, such that, in Kraut's example, if a swimmer can no longer swim, that constitutes a bad instead of merely the absence of a good.

Kraut says that sometimes these absences can constitute bads, but he thinks that the range of intrinsic bads is much broader than whatever amounts to non-flourishing. (Following hedonism again, he thinks that a complete theory should allow for goodness, badness, and neutrality.) He also argues that physical pain is insufficient as an account of ill-being because there are many ways that one can suffer physically without having what we ordinarily think of as physical pain. For example, we can shiver uncontrollably due to cold, sweat uncomfortably due to heat, be nauseated, exhausted, dizzy, thirsty,

<sup>13</sup>In the jargon I will introduce in Chapter 7, this makes Kraut's view a restrictive hybrid theory. It holds that the enjoyment and the objective good are both necessary for welfare.

and hungry. Our senses can be assailed in a variety of ways, including disgusting tastes, unpleasant noises, and blinding lights (2007, p. 150).

According to Kraut, when we experience unpleasant sensations, “the powers of [our] organs are being used to ill effect” (2007, p. 150). Kraut then claims that such unpleasantness is a form of un-flourishing because we are not developing appropriately or we are being prevented from developing. To Kraut, our sensory system has let us down. He says that, from the point of view of our well-being, “the sensory system we have been given by nature is disordered and not functioning as it should” (2007, p. 150). It is this organ disorder, which manifests as foul tastes, unpleasant noises, and so on, that causes the un-flourishing.

I am not the first to point out that Kraut is committing an error here (Hurka unpublished (a), p. 10). Kraut claims that, when we feel pain, our sensory system is not functioning as it should, and so the badness of the pain is rooted in our faulty development. But this is precisely the opposite of what is happening much of the time. If I place my hand on a hot stove, then the response—perhaps shortly after I automatically move my hand—will be pain, which indicates that I have done something harmful to my body and that I should avoid being so silly in the future. So, rather than a sensory system that is not functioning as it should, pain felt in my hand is *exactly* the correct response from a functioning body, so rather than being a component of ill-being, Kraut’s view should actually hold that this sort of pain is good for us. It is true that sometimes pain can be the result of a malfunctioning nervous system, which is one of the causes of chronic pain. When pain is caused for this reason, Kraut is correct that the sensory system is disordered, but he will not want to claim that only this type of pain and not other, ‘justified’ pain is bad.

According to Kraut, when we suffer pain, our organs are disordered “from the point of view of our well-being” (2007, p. 150). Maybe what Kraut means here is not that our organs are disordered in the sense of not functioning properly in a general way, but that they are functioning poorly just in terms of our well-being. However, this move will be unsuccessful for two reasons. First, the developmental part of Kraut’s argument, which involves reference to nature, will now be unhelpful, for he will not be able to speak of functioning generally. That was supposed to be the justification for the view. Second, it would be question begging to argue that the way we know that our organs are disordered is that they cause pain when it was supposed to be the organ malfunctioning that explained the way that organs contribute to our ill-being.<sup>14</sup>

Kraut’s view runs into more problems. He correctly claims that a negative emotion is not always a developmental malfunction. For instance, when someone feels sorrow at the death of a loved one, “one is functioning well, not badly, and the sorrowful feelings one experiences are not a debit in one’s well-being, not even a small one” (2007, p. 155). Other negative emotions, such as anger, can be appropriate, while the converse—positive emotions that are not good for our well-being— can occur, such as inappropriate, uncontrollable laughter. Kraut’s point with these examples is that the context matters when determining whether or not an emotion is justified, and that the absence of a justified emotion or the presence of an unjustified one can affect us for better or worse.

The problem here is that developmentalism offers us no way to determine which of these emotions are justified and which are not. They both feel bad. Sorrow, even when it is appropriate, is unpleasant, so that alone cannot settle the issue. But when we turn to the plausible explanations, none of them,

<sup>14</sup>A further issue, which foreshadows Chapter 7, is that we might reasonably hold that a malfunctioning organ is bad regardless of the presence of pain.

including the ones Kraut gives, relies on developmentalism to explain the difference. I lose my loved one and feel sorrow, and while the sorrow is unpleasant, I am glad that I feel it instead of feeling nothing. (Indeed, Kraut thinks that the sorrow does not decrease welfare at all.) But how do I know that sorrow is the correct response here? Whatever the reason, it is not going to be that I know that this is how a well-functioning brain responds to such events. To claim that would be to beg the question.

Kraut's response is based in part on the frequency of the emotion. If we feel sorrow at appropriate times and not too frequently, then that is good, but if we feel it more often, then that is bad for us. He is aware that this is a strange result, but claims that it is what we should want from a theory that is based on development. He says that a good life has a certain shape, and that certain things that can be good (or neutral) for us in small amounts can become harmful in greater amounts (2007, p. 169).

This is true of activities such as watching television (Kraut's example), but is implausible for an emotion like sorrow. Once again, developmentalism is unable to handle certain cases. Suppose that the village where I live is destroyed by an earthquake and I am the only survivor. Extreme sorrow is the appropriate response here, and, once again, this is not due to any fault with my development, but rather the correct response to my awful situation. Or consider Sen's landless labourer who develops coping strategies to deal with the unfortunate condition of his life. In such cases we are responding appropriately—exactly what we should expect and want to happen should these tragedies befall someone—yet we do not want to claim that these people are well off. Kraut does not want to claim that these people are well off either, but his theory fails to provide the apparatus we need to conclude this.

Developmentalism shows the difficulty with extending a theory of well-being to incorporate ill-being. I have argued here that his theory is unable to match our intuitions about cases, and he gives no reason for preferring the outcome of the theory over these intuitions. Not all of the problems with Kraut's theory arise from his account of ill-being, but considering ill-being has revealed more clearly what some of the principal issues are.

## 5.5 Narrow Perfectionism's Problem

Objective theories are characterized by their claim that there are goods and bads, the presence of which leave us better or worse off regardless of how we feel. States of affairs are good or bad independently of attitudes or desires. What makes the items on the list intrinsically valuable varies between theories, and the explanation is often unspecified. Perfectionism is a type of objective theory that gives an explanation for the goods on the list. It says that what is good for something is whatever perfects that thing's nature.

The problem for perfectionism is as follows. For narrow perfectionism, according to which the perfectionist values are the only morally relevant ones, perfectionism denies the existence of intrinsic evils (Hurka 1993, p. 100). This is because essential properties cannot be developed to a negative degree.<sup>15</sup> So whatever the perfection is, someone can possess it perfectly or to lesser amounts, but these lesser amounts will always be differences of degree instead of kind. Further, this structure is a necessary feature of perfectionist theories and not, it seems, something perfectionists can reject.

This structural feature goes against the basic structure I defended in Chapter 1 and elsewhere, according to which there are good, bads, and a neutral middle. Moreover, I argued that there are intrinsic bads that a total theory of well-being must account for, and that any theory that is unable

<sup>15</sup>This type of privationism has been defended by many perfectionists, beginning with Augustine (1998, book XI, chapter 9) and later by Aquinas (1485/2012, 1a, q. 48, arts. 1–2), and Leibniz (1710/1952, pp. 135–36, 140–41, 352).

to do so is imperfect. Many theories lack this structure due to a failure to consider ill-being, but perfectionism lacks this structure necessarily. These are not new problems for perfectionism. In his book *Perfectionism*, in which he argues the perfectionist goods are agent-neutrally good (instead of good for their possessors), Hurka recognizes that narrow perfectionism has this problem. He says that

Narrow perfectionism [...] does not permit talk of intrinsic evils. Because an essential property cannot be realized to negative degrees, the theory's scale of quality must have zero as its lowest point. It must say that every state that passes the tests of number has positive perfectionist value, and that, considered on its own, every human life is worth living (1993, p. 100).

Hurka notes that denying the existence of intrinsic evils does not entail denial of the things we often regard as intrinsically evil. For instance, perfectionists need not deny that pain exists. However, they cannot claim that pain is intrinsically evil, or that a human's life can be worth not living. This result—and the problems I presented above—is only a problem for narrow perfectionism. Broad perfectionism, which includes non-perfectionist values, can allow for the existence of evil. Such a view might combine perfectionism as one type of good with another type of good, and these other goods could have negative counterparts (or there could be bads with no positive counterparts) (1993, pp. 100–101). So a theory incorporating perfectionism can include evils, but it is the non-perfectionist elements that would enable that. Therefore, it is narrow perfectionism that must be rejected.

Others have criticized perfectionism for similar reasons. Sumner points out that narrow perfectionism cannot include evils, and that this is a troubling feature. He argues that a theory of the good should be able to make sense of the seemingly obvious fact that lives can be scarred by tragedy (1996, p. 195). Sumner also criticizes perfectionism for being unable to conclude that a life can be worth not living (i.e., that we can be doing so badly that it would be better for us to die). We can only make sense of acts like euthanasia—a medical procedure many people believe is morally justified some of the time—if someone can judge her life have more bads than goods to make continued living evil.

## 5.6 Conclusion

It is difficult to make any general conclusions about objective theories. Not only is the range of defended views extremely broad, but the range of possible views is close to infinite. Some objective goods, including knowledge, have negative counterparts. Others, such as autonomy, do not. I have also argued that Kraut's developmentalism, one of the only prudential theories that attempts to develop an account of ill-being, fails to produce a plausible account. Finally, I showed that pure forms of perfectionism, which Hurka calls narrow perfectionism, should all be rejected because they make ill-being impossible.

# Chapter 6

## Anti-Achievement

### 6.1 Introduction

Philosophers in recent years have given increased attention to achievement as an intrinsic good.<sup>1</sup> Despite this attention, far less has been written on the possibility that achievement has a negative counterpart or counterparts. In this chapter I make a start toward correcting this discrepancy by considering the various states of anti-achievement and assessing whether any of them are intrinsically bad. I argue that, insofar as achievement is a plausible intrinsic good, there is at least one intrinsically bad form of anti-achievement.

In many ways, achievement is similar to knowledge, which Kagan calls a *structured good* because there are many conditions that must be met for knowledge to obtain (2014, p. 281). In my discussion of knowledge, I described how, even if we make the simplifying assumption that knowledge is justified true belief, there are many ways of falling short of knowledge, and each way has positive, neutral, or negative value. I agreed with Kagan and Hurka that the worst state of anti-knowledge is what I call anti-justified false belief, which involves a false belief when there is evidence to the contrary. Achievement is also a structured good. It involves multiple conditions, so, as with knowledge, we can ask whether any state of anti-achievement is bad, and if so, which state is the worst and why. Achievement also shares the feature of knowledge that there is no common term to capture the nuances of the various ways of falling short. For knowledge, ‘ignorance’ captures more than one way, and there is no common term for ‘anti-justified false belief’. Similarly, ‘failure’ involves any type of lack of success, but as I will show, mere lack of success does not necessarily constitute an intrinsic bad. This means we are stuck with the unappealing term ‘anti-achievement’.

My strategy here is not to work out a complete theory of anti-achievement, but rather to show some of the different possibilities. Looking at theories of achievement is a good starting point, but philosophers have suggested many factors that affect the value of achievement, including competence, difficulty, effort, and self-sacrifice. There are also a variety of proposed relevant features concerning the goal one aims at, including its intrinsic value, rationality, and comprehensiveness. Others, meanwhile, argue that a goal is not a necessary condition of achievement.<sup>2</sup> It would take a much larger project than the one I am presently undertaking to assess the plausibility of each of these factors for achievement

---

<sup>1</sup>Three notable examples are Gwen Bradford (2015), Thomas Hurka (2011), and Hasko von Kriegstein (2014).

<sup>2</sup>Most notably Bradford (2015).

and its comparative plausibility for anti-achievement. To make the task manageable, I will propose a minimal account of achievement and then consider the opposites of each condition. I am not arguing for a particular theory of achievement or arbitrating between different theories, and much of what I say applies to anti-achievement regardless of the correct theory of achievement.

In this discussion, I consider how achievement and its contraries affect prudential value, or how they contribute to us being well off or living better lives. Others prefer to think of achievement non-prudentially.<sup>3</sup> Much of what I say in what follows is compatible with either interpretation, but I will stick to prudential value for consistency. One of my goals is to show that anti-achievement does not trivially follow from achievement. As I have been arguing throughout this dissertation, we err if we assume that determining the correct theory of well-being will, without further investigation, produce the correct theory of ill-being. I demonstrate this by showing that achievement and its contraries are structurally asymmetrical: i.e., that features that apply to one do not apply to the other.

## 6.2 Achievement

Achievements all have the same structure: there is a process that results in a product (Bradford 2015, p. 11). A composer writing a beautiful song involves the process of her envisioning the piece and writing down the notes. The product is the completed composition. A runner finishing a marathon involves covering the marathon distance by foot. Exactly what should be included in the process for each achievement depends on what the product is. Achievements such as running marathons and composing songs involve many sub-processes, such as training for weeks or years, learning the various components of music theory, and so on. Demarcating the boundaries of the process is important for determining the value of the achievement, but for my purposes the process-product condition is sufficient. This can also be described as a success condition, given that something is brought about (though speaking of success might imply goals, which the minimal account does not include).

The second necessary condition is competency. In order for something to count as an achievement, the product cannot be the result of pure chance. Winning the lottery is not an achievement, nor is finding twenty dollars on the street. Even though these events involve a process and a product, the reason they are not achievements is that the agents do not play the correct role in determining the outcome. In such cases, the outcome involves too much luck, so to rule out such events, a theory of achievements needs an anti-luck condition to rule out cases involving pure luck (von Kriegstein 2014, p. 49). There are many possible interpretations for such a condition, which represent the scale from ‘no luck’ at the one extreme to ‘almost pure luck’ at the other. Setting the condition at ‘no luck’ is too strong, as nearly everything we think of as an achievement involves some element of luck, but just how much luck can be tolerated is another issue for the achievement theorists to sort out.<sup>4</sup>

There are also different interpretations of competence. Some have defended epistemic accounts, according to which one’s competence increases the more knowledge the agent has about what she is doing (e.g., Bradford 2015, chapter 3; Hurka 1993). A different approach locates competence in the abilities of the agent (Greco 2010), while another view holds that competence concerns the relationship

<sup>3</sup>This is Hurka’s preference, for instance (unpublished (b), p. 1).

<sup>4</sup>The full story is slightly more complicated. Sometimes luck can play a significant role in the outcome without undermining the achievement. For example, non-elite runners need to win a lottery to run the New York City marathon, so luck plays a major part in any achievement in that race. However, this type of luck is not undermining. Perhaps the better way to frame the anti-luck condition is that it matters that we avoid having the wrong *kind* of luck, not too much luck. For the purposes of my discussion I will assume that all luck is undermining.

between the agent and the likelihood of success (von Kriegstein 2014, pp. 75–78; Hurka unpublished (b), p. 8). The details of the competence account are unimportant for my purposes here, but the correct account might play a role in determining anti-achievement. What is important is that, all else equal, the greater the agent’s competence, the less the outcome was due to luck and, therefore, the greater the achievement. Although Greco’s version of competence concerns the agent’s general abilities (what I will call *global* competence), for achievement it is better to think in local terms. By that I mean that it is the agent’s actions in the specific situation leading to the achievement that determine its value, not the agent’s general abilities (i.e., his global competence). To see the difference, consider a case from von Kriegstein involving a baseball pitcher, Ricky, who usually screws up in important situations (2014, p. 75). However, in one such situation, Ricky throws three perfect pitches to strike out the batter. Even though Ricky is normally incompetent in such cases, his competence in this specific case makes his strikeout an achievement. Therefore, for achievement local competence is more important.

The final necessary condition for the minimal account is difficulty. Different definitions of difficulty have been proposed.<sup>5</sup> For simplicity, I am going to stick to the account that says difficulty is equivalent to effort. Why is it not an achievement for me to brush my teeth, tie my shoes, or make a cup of coffee? Intuitively, the answer relies on the lack of difficulty involved in my bringing about those products. They involve little effort for me. Our assessment would change if my circumstances changed. Suppose, for example, that I break both my arms. Now all of those products would be more difficult, probably to the point where they would count as achievements. In describing this case, notice that I have only appealed to *my* abilities, so the type of difficulty I have in mind is agent-relative. This follows other accounts of achievement, although, unsurprisingly, there is disagreement concerning the conditions of agent-relative difficulty (von Kriegstein 2014; Bradford 2015).

In sum, the necessary conditions of the minimal account are as follows: there must be a process that results in a product (success), and the size of the achievement is determined by how difficult it is for the agent and how competently the agent brings it about. I will conclude this section by mentioning two other conditions that are not included in my minimal account, but which might be added to a more robust account.

The first is goals. Many defenders of the intrinsic value of achievements have included goals as a necessary condition, including Hurka, who says that to achieve something is to “intend to produce some goal and then do so” (2011, p. 98). However, following Bradford, I have left goals out of the minimal account. This is because there are cases where, intuitively, a product has value as an achievement yet the agent’s goal was not fulfilled. Sometimes this is due to the agent setting unrealistically high expectations. Suppose a reasonably competent runner enters a race with the goal of winning. Lining up for the start, she says to herself “All I care about is first place—if I am going to get second I might as well get last.” Despite her best effort she ends up getting second against a competitive field. To me, this runner getting second is still an achievement, despite her own beliefs. Of course, this sort of example is unusual because most athletes do not have the sort of all-or-nothing approach I have described, but *some* athletes do, and even *their* unmet goals can count as achievements.

Non-athletic cases show the same point. Sometimes we set out on paths without having a clear goal in mind, and sometimes we have some goal in mind but not the one we eventually achieve. Consider Roy Plunkett, the DuPont scientist who discovered Teflon. Plunkett was trying to find a way to improve

<sup>5</sup>Accounts include probability, where a less-likely outcome is all else equal a greater achievement (von Kriegstein 2014) and complexity, where a more complex goal is all else equal a greater achievement (Hurka 1993).

refrigerators, but it was through his experiments that he discovered a material that made for better frying pans. Plunkett's discovery is an achievement even though he was not aiming at that particular goal.<sup>6</sup>

The other condition worth mentioning is agent-neutral difficulty. Some think that agent-relative difficulty is insufficient (e.g., von Kriegstein 2014, p. 124). The problem with a purely agent-relative account is that products that are easy for the agent are not achievements, even though producing that product is difficult for everyone else. For example, on Bradford's view, it is not an achievement for a virtuoso violinist to play a complex piece if he finds it easy, even though all the non-virtuosos find it extremely difficult. I see the value in adding an agent-neutral difficulty condition. However, for the purposes of developing an account of anti-achievement there is no need for me to settle on a complete account of achievement. I will therefore stick with the minimal account as I have described it.

## 6.3 Failure

With the minimal account in place, we can ask whether each condition has an opposite, and if so, what effect that opposite has on prudential value. The opposite of the product condition is failure, which is also the first thing that comes to mind when people think of 'the opposite of achievement'.<sup>7</sup> At least in some cases, many people will have the intuition that there is something bad about failure. This is so even if you try as hard as you can, and even if you act competently. Just as success makes us happy, which for the objective theorist can be a sign of value, failures cause us pain or rob us of valuable mental states. For an objective theory, the badness is not fully explained by the pain of failure, but rather failures are painful because they are bad for us. A life filled with failure can be worse even if, somehow, the agent never discovers them. Some will doubt this. Even a defender of the intrinsic objective goodness of achievement might be skeptical that failures are intrinsic objective bads.

This foundational point is difficult to establish, for while I can point to signs of objective badness—e.g., that failures are often extremely painful—just as achievementists can use pleasure as a sign for the value of achievement, one can be a skeptic on either side by holding that our attitudes fully capture the value of either side (e.g., failure is only instrumentally bad because it feels bad, not because it is actually intrinsically bad). My sense is that failure can make us worse off than we would have been not pursuing some outcome, and even though we might use the failure as an opportunity to improve, it can contribute ill-being. The issue is that some failures do not seem bad at all. I might try my hardest to help my team win a game of recreational soccer, only to come away with a loss. This does not strike me as intrinsically bad, so, minimally, failure on its own is only doing some of the work.

### 6.3.1 Goals and Attempts

Failure involves the absence of success (and therefore achievement), but that is not the entire story. It is true that there are many achievements that I am failing to get right now, but it is false—or at least misleading—to say that these each constitute some failure on my part. If this were the case, my life would be filled with failure, for there are many possible achievements that I am missing out on, and each

<sup>6</sup>Those who are attracted to a goal condition can offer ways around these sorts of cases. One could claim, for instance, that Plunkett had a goal of scientific discovery, or helping people, or keeping his job.

<sup>7</sup>Failure is the opposite of success, but instead of mentioning competence and difficulty every time, I will assume that other relevant features of anti-achievement are held fixed. Failure here just means lack of success, even when the agent is competent and pursuing a difficult product.

of those would mean that I am failing to get them. In a trivial sense I am failing to get them, but they do not appear to be proper failures (or disvaluable ones). For example, suppose that with appropriate training and sufficient funding I could climb Everest. It would be an achievement were I to make it to the top, yet that I have not done so does not mean that I have failed to climb Everest. Similarly, if we suppose that Plunkett had not discovered Teflon, we should want to say that he lacked that achievement, but it does not constitute a failure on his part.

To make something a failure, one necessary condition is an attempt at some goal. We can amend Muddy Waters's claim that "you can't lose what you ain't never had" to "you can't fail to get what you ain't never aimed at". Something counts as a failure if one aims at some outcome and fails to bring that outcome about. Therefore, my lack of Everest summits does not presently constitute a failure, but were I to have the goal and attempt the climb but fail to reach the summit that would, obviously enough, be a failure. Achievement theories that omit goals, such as Bradford's theory and my minimal account, are therefore structurally asymmetrical. A condition of failure does not apply to success.

Bradford's discussion of failure points to the same structural asymmetry: "To be precise, what's at issue here is a *failed attempt*, which is a matter of trying to achieve some product, but not bringing that product about" (2015, p. 171). Elsewhere in her discussion of failure, Bradford omits the goal condition because she thinks that failures usually still involve achievements: e.g., the person who gets second in the New York City marathon still competently causes a difficult outcome, even if his only goal was to win.<sup>8</sup> Similarly, it would be silly to say that the person who finishes last failed to win the race, even though in one sense that is the case. This is presumably because the person who finishes last did not intend to win.

For Bradford, what we typically think of as failure often involves achievement. But suppose that, for reasons of structural symmetry, someone were to deny the goal condition. This would mean that we fail a lot more than we think we do. This account makes the theory symmetrical with the minimal account of achievement. Achievement involves success at producing a product, while failure means lack of success. Our actions do not culminate in products all the time, and if not to an infinite amount, there are certainly many failed products. I might take consolation in my failure to reach the summit of Everest by reflecting that I almost made it, but my climbing involved many other failures: to win a Nobel Prize, discover a new antibiotic, achieve world peace, and so on. This might not be a problem if failures are not intrinsically bad, but it implausibly concludes that we are failing pretty much at every moment.<sup>9</sup> This account would hold, for instance, that it would have been a failure had Plunkett not discovered Teflon, which is implausible. So goals are required.

With the goal condition, Bradford's theory holds that failure involves failing to bring about an intended outcome that is both difficult and competently caused. A theme in her discussion of failure is that what we perceive as failure often involves a lot of other achievements which are nearly as good as the intended outcome, and her main point is that failures often produce achievements regardless of how we view the failure. Almost reaching the goal is nearly the same as reaching it.

In contrast to Bradford's account of failure is the more ordinary sense of the term: the sense in which I intended to climb Everest and failed and am now (seemingly appropriately) disappointed. As she says, this type of failure involves failing in our own eyes (2015, p. 172). This is the sense she refers to when

<sup>8</sup>Compare Bradford's claim to James Lovell, the commander of the Apollo 13 mission, who called the outcome a 'successful failure'.

<sup>9</sup>Which brings to mind Samuel Beckett's line from *Worstward Ho*: "All of old. Nothing else ever. Ever tried. Ever failed. No matter. Try again. Fail again. Fail better."

she mentions how few runners expect to win the marathon, and so do not consider their mere finishing to be a failure.

My restriction to goals in this section is not to say that it is impossible to fail without a goal. Consider a father who abandons his young child and never re-enters the child's life. In such a case, the father has failed to be a good parent, and, more importantly, perhaps this is bad for him, even though he never had the goal of good parenting. Similarly, suppose I run into a burning building to save my prized Muddy Waters guitar instead of my trapped mother. I succeed at saving the guitar, but not my mother. Surely I have failed in some important way, and this failure could be bad for me. These examples both involve moral conduct, so perhaps the account should be amended such that failures involving morality—to be a good person, do the right thing, and so on—do not require goals. This is because we *should* have had the goal to do the right thing, morally speaking. Non-moral goals such as running marathons are different, as there is no direct sense in which they should be chosen over other pursuits. In what follows I am only discussing non-moral goals. Minimally, to fail means intending to bring about a desired *non-moral* outcome and not doing so.

### 6.3.2 The Value Jump

Consider two cases. One: Al Gore lost the United States presidential election by 537 votes. Two: Gary Robbins missed finishing the 2017 edition of the Barkley Marathons by six seconds. The Barkley Marathons is an ultramarathon event in Tennessee, consisting of 100–120 miles of trail running and off-trail navigating that athletes must complete in under 60 hours. Only fifteen people have finished since 1986, making completion of the race a substantial achievement. In the 2017 race, Robbins went off course in the last two miles and missed the 60-hour cutoff by six seconds.<sup>10</sup>

If the minimal account of achievement included goals, Robbins and Gore both achieved nothing. In contrast, on Bradford's account, Robbins missing the cutoff by six seconds is less of an achievement than finishing just in time, but only just barely. Missing it by one second would have been very nearly the same achievement as finishing within the allotted time (i.e., the difference in value is whatever the difference one second makes, perhaps regardless of where that second is located).<sup>11</sup> Similarly, there is little difference in the size of the achievement when Gore lost the election by 537 votes than if he had won. Both of them almost succeeded, making their achievements very nearly as good as if they had succeeded on Bradford's view.

Failures of this sort point to a jump in value, which we can get in different ways. The first is to amend Bradford's account of achievement so that achieving one's goal plays a part in determining the achievement's value. This would mean that falling short of the goal might still be an achievement, so long as it is difficult and competently caused, but the 'goal value' only occurs when the goal is achieved. This would be a large departure from Bradford's theory. She is willing to grant the value of completing goals, but maintains that this value is separate from the value of the achievement.

I propose that the best way to account for this jump is through the badness of failure. Six seconds and 537 votes makes all the difference. This is captured by baseball player Frank Robinson's remark

<sup>10</sup>Close calls happen regularly in ultramarathons, although not always as misses. In the 2015 Western States 100-mile race, an athlete finished six seconds within the thirty-hour cutoff. That same year, an athlete in the Hardrock 100-mile race finished *a single second* ahead of the sixty-hour cutoff.

<sup>11</sup>Bradford is not committed to a linear increase in value, but reaching the goal is not supposed to contribute value on its own to the achievement. She says that "As far as achievements go, my view entails that there is not such a gap in value, other things being equal" (2015, p. 172).

that “Close don’t count in baseball. Close only counts in horseshoes and grenades.”<sup>12</sup> Cases such as Gore’s and Robbins’s show that the total value is sometimes non-linear, or perhaps even discontinuous. Bradford’s view is not committed to linearity. It is more difficult to run the last kilometer of a marathon than it is to run the first one, so difficulty can increase non-linearly. But her view denies that there is a substantial jump in the value *of the achievement* when the goal is fulfilled. In other words, Bradford only denies that the value jump is a feature of achievements. Her view does not rule out the value jump for some other reason, such as the intrinsic value of fulfilling one’s goals (2015, p. 172). So the lack of a value jump in her account of achievement is not necessarily a flaw, but it does point to a need to determine how to include the jump, which I think is best accomplished by adding failure.

The value jump is not always due only to achieving the personal goal one has set. At least in some cases, the personal goal is not doing all of the work in explaining what makes the outcome a failure. Sometimes failure is defined in terms of the activity. The length of a modern marathon (42.195 kilometers or 26 miles and 365 yards) is an arbitrary distance.<sup>13</sup> There is nothing besides tradition that makes the distance special, yet finishing a marathon is a significant achievement in a way that running a slightly shorter distance is not. Suppose I enter a marathon and make it to 42 kilometers and cannot continue, so I drop out of the race. On Bradford’s account, this is nearly as great of an achievement as covering the additional 195 meters. Her view is correct that as I cover more distance in the race my achievement increases. This is for the obvious reason that it is harder to run farther. We want the overall view to explain why there is a value jump upon completion of the goal, which Bradford’s account of achievement does not do on its own. (Again, this does not rule out the addition of other features that will produce the jump.) When looking at the total value of running a marathon, the value increases steadily as one approaches the finish, then finishing the race leads to a large spike in value. This will be true for other distances and other activities: dropping out of a ten-kilometer race at nine kilometers is not nine-tenths as valuable as finishing, nor is getting nine strikes in a row while bowling nine-tenths as valuable as bowling a perfect game. We need the value jump to explain why, e.g., we care about reaching the top of the mountain instead of stopping a few meters from the summit.

There is a difference between some goal we might set for ourselves (bowl over 200 in this game) and having a goal necessary for the activity. This is what distinguishes running a marathon from simply running for a long time, and it is also what makes stopping at 42 kilometers less good than the 195-meter difference suggests. Compare this to an achievement where the value increases more linearly, such as playing the game of running as far as you can in an hour. There, each little bit of distance covered increases the value of the achievement, and the goal is simply to run as far as possible. Similarly, the goal of winning an election captures the value jump: Gore losing by 537 votes does not mean that the total value was only very slightly less good than if he had won. Winning would have added a large jump (See Figure 6.1). The graph continues rising after the victory has been established because it is a greater achievement to win by more votes than by fewer.

Granting Bradford’s account of achievement, with the addition that some failures are intrinsically bad, produces the jump we are looking for: when one achieves one’s goal there is all achievement and no failure, but falling short of the goal means, on her view, that there is still some achievement but, by adding in failure, also some disvalue to the outcome. So failing by any amount will lead to the jump in value. Exactly how big the jump is will be determined by the independent values of the achievement

<sup>12</sup> *Time Magazine*, July 30th 1973. It is often misquoted as “. . . horseshoes and hand grenades”.

<sup>13</sup> It was extended from 40 kilometers so that the athletes would pass the royal box in the 1908 London Olympics.

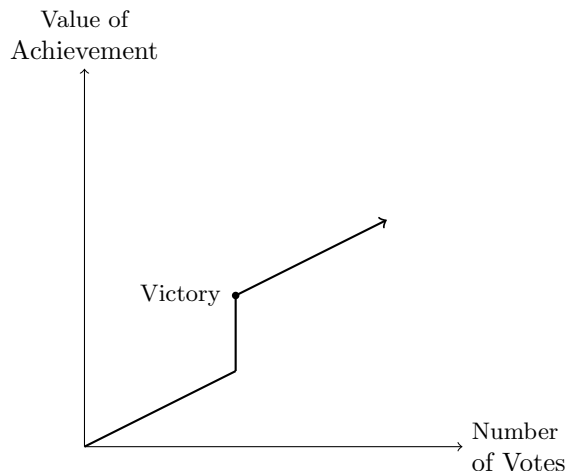


Figure 6.1: Election Value Jump.

and failure.

Considering cases of failure reveals something we might have missed were we to focus only on achievement. The first is the role of goals. The value jump does not show that Bradford’s account is mistaken, but it does show that goals need to enter the picture somewhere, either as something with independent value or as a feature of failure. The second is a new justification for the intrinsic badness of failure: it can lead to a value jump when one fails to achieve one’s goals. While there are ways of accommodating the value jump without claiming that failure is bad, for theories such as Bradford’s introducing the claim about badness produces the correct result without amending her theory.

Hurka argues that even in cases of failures that involve little achievement, the failure seems not to constitute an intrinsic bad (2010, pp. 219–220). (As usual, he is discussing non-prudential value.) For him, even if failures are bad, they are less bad than the good of achievement. In an amendment to Tennyson, Hurka claims that “Tis better to have sought and failed than never to have sought at all.” If this claim is correct, it means that there is a value asymmetry between success and failure, such that an achievement of some amount  $x$  is more valuable than a failure of amount  $x$  is disvaluable. I am open to such an asymmetry. Hurka does argue that one type of failure is significantly bad, a view which I discuss in the final section.

### 6.3.3 The Outcome Gap View

Goals are required for non-moral failure. Minimally, you must aim at something and not bring it about. So what makes some failures worse than others? I suggest that one factor is how far the outcome is away from the intended goal. In the 2014 FIFA World Cup, Brazil, the host country and a favourite to win, lost 7–1 to Germany in the quarterfinal.<sup>14</sup> They had the intention of winning, so the amount they lost by makes their failure worse than if it had been a close match. Or consider the famous ‘contender speech’ from the film *On the Waterfront*, where Terry Malloy, played by Marlon Brando, reflects on his failure to become a top boxer: “I coulda had class. I coulda been a contender. I coulda been somebody, instead of a bum, which is what I am. Let’s face it.” He had the goal of being the champion but did not even get a chance to prove himself, which makes the failure worse. Call this the outcome gap view,

<sup>14</sup>The German players, in contrast, had no problem achieving their goals.

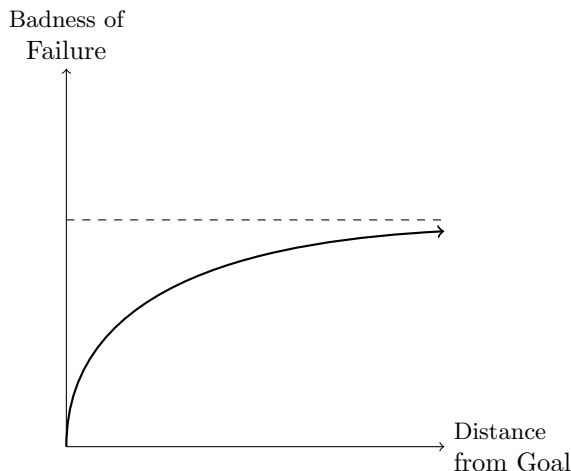


Figure 6.2: Decreasing Marginal Disutility of Failure.

where the greater the gap between the goal and the outcome, the greater the failure. I affirm this view. It strikes me as obviously correct.

The outcome gap view reveals another structural asymmetry between the minimal account of achievement and failure, for there is no similar feature regarding the distance from one's goal that makes something a greater achievement. (This is both for the trivial reason that goals do not count on the minimal account, but also because the gap is unlikely to contribute more value.) If there were such a feature, we would have a reason to choose easier goals just because we are likely to greatly surpass them, but we cannot affect our welfare in this way. Even if the increased value of surpassing the goal were outweighed by the decreased value in the achievement being lower, this still seems implausible. So the outcome gap view applies to failure but not to achievement, making the two structurally asymmetrical.

There is a potential problem for this view, however. Consider someone who sets impossibly high standards for herself: e.g., she tries bowling for the first time and aims to bowl a perfect game the next time she plays. This fact alone means her failure will be greater, given that there will be a larger gap between her unlikely goal and the outcome. Perhaps just as we should not be able to significantly increase our welfare by setting low goals, we should not be able to lower our welfare by setting them too high. Some might not find this to be a serious problem. If someone has unreasonable expectations, we should want that unreasonableness to show up somewhere, which it does in the outcome gap view.

However, I believe the problem is serious enough to warrant considering how to amend the outcome gap view to avoid it. The strategy I propose is to change the view so that the badness of the failures does not increase linearly, but rather has diminishing marginal disvalue (see Figure 6.2). The disvalue might increase linearly when it is within a certain range close to the goal, but past a certain point additional distance from the goal will produce a smaller amount of badness.<sup>15</sup> This amendment is intuitively plausible. Suppose Brazil had lost to Germany by 36 goals instead of six. As the number of German goals increased, each goal would make the Brazil failure worse, but it would make less of a difference. Certainly the first six make more of a difference than the last six. The amended outcome gap view accounts for this while the linear one does not. On this view, once you have failed, a lot of additional failure is less important.

<sup>15</sup>This is also plausible in goods such as pleasure (e.g., Hurka 2010, pp. 204–205), although perhaps less often in bads.

Within the outcome gap view, there are actually two possibilities.<sup>16</sup> So far I have been describing the *absolute outcome gap view*, according to which the badness of the failure is determined by the absolute gap between the goal and the outcome. But another possibility—the *relative outcome gap view*—holds that the badness of the failure is determined by the relative gap between the goal and the outcome. To see the difference between these options, suppose that a soccer team ambitiously desires to score ten goals but ends up scoring only two. Compare this to a team that less ambitiously desires to score five goals but ends up scoring only one. The absolute gap view says that the first team’s failure is worse, given that there is a larger spread of goals between what they wanted and what they achieved, whereas the relative gap view says that the failures of the two teams is the same. In some cases, such as voting, the relative gap view fares better. Gore lost by 537 votes, which as a percentage of the total is very small. Suppose someone running for student council also loses by 537 votes, but in his school there are only 600 voters. The absolute gap view says that the badness of the student’s loss is equal to Gore’s, but that is incorrect. So at least in voting cases, the relative view is better.

### 6.3.4 The Proximity View

Here is a second component that affects the badness of failure. This view is less intuitive than the outcome gap view, but I will argue that it is plausible nonetheless. According to this new option, in some cases, getting close to the goal without achieving it is *worse*. Comedian Jerry Seinfeld captures the sentiment:

I think I have a problem with that silver medal. I think if I was an Olympic athlete I would rather come in last than win the silver. [...] You win the silver, that’s like “Congratulations! You almost won. Of all the losers, you came in first of that group. You’re the number one loser. No one lost ahead of you”.

On the view under consideration, it is one’s proximity to success that makes one’s failure worse than doing less well up to some point. Call this the proximity view. It has explanatory value, including an explanation of the tragedies that befell Gore and Robbins. They nearly succeeded but did not, and we feel appropriately bad for people in these situations, more so than with those who get less close. And, of course, these failures attract a lot more attention than ordinary, less-close ones.<sup>17</sup>

I claimed above that one way of getting the value jump is to posit the intrinsic badness of failure. If the proximity view is correct, the value jump will be larger than it would be without it. A feature of combining Bradford’s account of achievement with the proximity view is that achievements and failures can increase in the same direction at the same time (i.e., a greater achievement does not mean less of a failure and vice versa).<sup>18</sup> Robbins and Gore both got close to their respective goals, and to that extent their achievements were greater because it is a greater achievement to get more votes or to run farther. But on this view their *failures* are also worse because they got so close but missed their goals. Graphically, the maximum proximity failure point is just short of success: e.g., losing by a single vote. So

<sup>16</sup>Thanks to von Kriegstein for pointing out this distinction to me.

<sup>17</sup>Indeed, they often get more attention than successes. Robbins’s failure garnered significantly more media attention outside the ultrarunning community than success would have. A similar case is the amateur bowler Bill Fong. While perfect games (300 points, or 12 strikes in a row) are relatively common, even among amateurs, a ‘perfect series’ of three perfect games in a row (900 points or 36 strikes) has only been accomplished 21 times. Fong nearly achieved this amazing feat in 2010. He got 35 strikes in a row, but on the last shot he failed to knock down a *single pin*, getting 899 as a result. The failure caused a stroke shortly thereafter. See Mooney (2012) for an account.

<sup>18</sup>von Kriegstein notes the same point as a general feature of achievements and failures (2014, p. 28).

as the line moves away from the origin, the achievement increases in value, but so does the size of failure should the goal not be achieved. This is one example where failure does not require lack of achievement. Of course, it requires *one* missed achievement, but they do not work in opposite directions all the time.

There is evidence from psychology research that we subjectively respond to perceived failure in accordance with the proximity view. Psychologists have used Olympic athletes to measure responses to success and perceived failure. The format of the studies is usually the same: undergraduates are shown videos of medal ceremonies and are then asked to rate the happiness displayed by each medalist in facial expression. Intuition says that happiness should steadily decrease from first place down through the rankings, but this is not what the studies find. Instead, they find that the gold medalist is happiest, followed by the *bronze* medalist, then silver (Medvec et al. 1995; Totterdell 2000; Kudrna et al., 2016). Indeed, in a study of Olympic judo athletes, over half of silver medal winners displayed facial expressions of either sadness or contempt (Matsumoto and Willingham 2006).<sup>19</sup> The proposed explanation for this effect is that the athletes engage in counterfactual thinking. The bronze medalist compares herself to all of the athletes who did not medal, while the silver medalist compares herself to the gold medalist. There are different explanations for why the comparisons go in these directions, but in any case, counterfactual thinking helps explain the subjective response of the proximity view. Gore and Robbins think of what might have been had events gone slightly differently. Gore will not think of all the people who did not make it out of the primaries, just as Robbins will not think of the people who covered less distance or who did not even start. This is because they get close enough to success to make achievement the likely comparison.

The Olympic medal studies produce surprising results and show the role subjective response plays. However, that people consistently respond in this way does not mean that the intrinsic value of success and failure matches our subjective assessments. It is possible that achievements like an Olympic silver medal possess greater objective value than a bronze, but our subjective assessment also contributes to our welfare, just as a different good. (Here I am deviating from Bradford's account. On her view, there is no sense in which silver is objectively better than bronze. The achievements of each place from first to last could be equally valuable if the difficulty for each athlete was the same and if each outcome was competently caused. As an achievement there is no sense in which doing better means achieving more, where 'better' means placing closer to first.) The distinction between the value of the achievement as an achievement and our subjective assessment might also explain the proximity view. The *pain* of failure can be bad, especially when we nearly succeeded, but the subjective response need not affect the independent value of the achievement.

However, our subjective response can also signal whether something is good or bad for us. We do not treat all failures as prudentially neutral, and while some mild failures might produce little to no badness, significant failures can be devastating. It is difficult to determine the plausibility of the proximity view because it is hard to separate our subjective disappointment at the failure from the failure's potential intrinsic badness. One way of testing the proximity view is to consider a case of failure where the individual is unaware that he almost succeeded. Suppose that Anthony has the goal of climbing Everest.

---

<sup>19</sup>Judo might be a poor sport to study, given that the final match guarantees at least silver and the ceremony occurs shortly after. Medvec et al. (1995) studied all sports for which NBC broadcast the medal ceremony at the 1992 Barcelona Olympics, including sports such as swimming and running, which have a different structure to judo (many people race at the same time). They found the same result. They also note that regret is long-lasting. They cite an interview with Abel Kiviat, who was leading the 1,500m running race in the 1912 Stockholm Olympics, only to fade in the closing meters to lose by one-tenth of a second, taking silver. In an interview when he was 91 years old, Kiviat said that he still frequently thought about that loss (Tait and Silver 1989, p. 351).

A snowstorm develops on his way up the mountain, forcing him to turn back. When he returns to base camp, he is dejected that he did not make it to the top. In his despondency, he hands his GPS watch to a colleague, saying he does not want to know how close he got. Consider two possibilities. In the first, the colleague looks at the data and sees that Anthony made it halfway to the summit. In the second, the colleague sees that Anthony was only 50 feet from the top. In neither scenario does Anthony find out. Which scenario is worse for Anthony? I would rather the first scenario be true than the second. Both are bad, but the second is worse. I would find it more heartbreaking that Anthony got so close but failed than if he had made it only halfway. Other cases show the same thing. Suppose that while in Nepal, Anthony finds out that he was rejected from a top university. It turns out that he was one spot away from acceptance, which he never discovers. This is in some way worse than being further down the list. Cases of this sort show that knowledge of the level of failure is unnecessary to make the failure bad, so it is not the agent's subjective assessment that is doing the explanatory work.<sup>20</sup> The same point applies to situations where an individual does not know whether she succeeded or failed due to unusual circumstances, or perhaps because she dies before she can know the truth. In cases of achievement and failure, it is not the subjective assessment that matters.

The proximity view is actually an example of a more general account of close calls. (I am using 'close call' atypically here. I mean near successes. We have so many phrases for something bad almost happening—near miss, close call, narrow escape—but no standard phrase for something good that almost happens.) Hurka provides such an explanation in his account of appropriate attitudes, where he defends what he calls the 'first modal condition', according to which concern for a nonexistent object has less value as it becomes a more remote possibility (2001, p. 118). On this account, we are justified in regretting some good we missed out on provided that there is a close enough possible world where we obtained it. The psychology studies I described above show that we do in fact regret close calls, while Hurka explains that it is appropriate for us to do so. Robbins is justified in being pained by his near success, given that (presumably) had things gone just slightly differently he would have achieved his goal. This is in contrast to some further possible world where his regret is less justified because he loses because he gets less close to the cutoff.

This might pose a problem for the proximity view, given that Hurka's account applies to all goods and bads, not only failures. Someone who returns from a tropical vacation during which it uncharacteristically rained the entire time is justified in regretting the extra pleasure she would have obtained had it been sunny instead (2001, p. 118). However, to use Hurka's example, she is not justified in regretting that no "aliens abducted her and [took] her to an intergalactic pleasure palace" (2001, p. 118), or, if she is, only a very small amount of regret would be justified. Robbins can regret his failure, but that need not be because of any feature about the badness of the failure. In the case of Anthony failing to climb Everest, even though he does not find out, he *would be* justified in regretting the lack of success more in the case where he got closer to the summit.

Rather than a problem, I think it is a strength of the proximity view that it is an instance of the more general modal condition Hurka proposes. Another possibility is that our intuitions are overdetermined in these cases. The modal condition explains them, but so does the proximity view, so they might both be true. In either case, that the two approaches align is not a problem. I grant that intuitions for the proximity view are fuzzier than the outcome gap view. The proximity view seems plausible to me, but

<sup>20</sup>It is possible, though, that we are considering how we would feel were we in their place but with the information. But insofar as this is a problem, it is a problem for many arguments for objective prudential value.

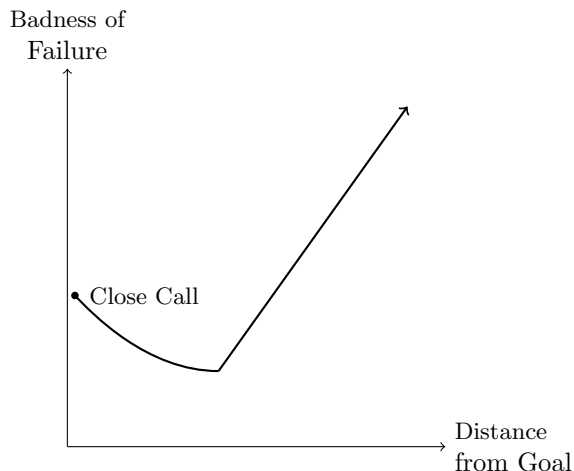


Figure 6.3: Outcome and Proximity (Outcome-Weighted).

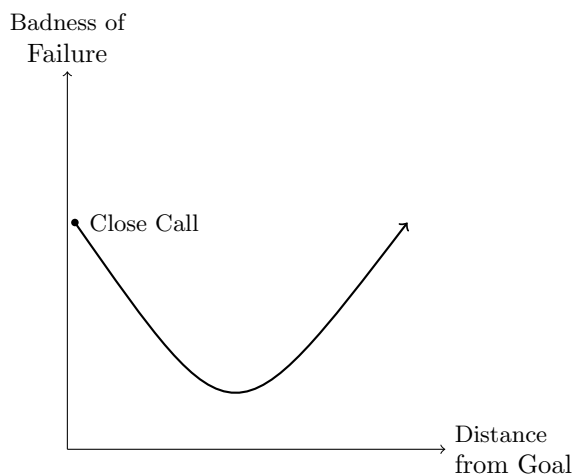


Figure 6.4: Outcome and Proximity (Equal Weight).

there might be some other way to explain it.

Perhaps both the gap view and the proximity view are correct. If so, they would produce graphs of the following form. In Figure 6.3, the outcome gap has more weight than proximity, so badness decreases as one gets closer to the goal and then kicks up at the end.<sup>21</sup> Some might feel that proximity should be given more weight. This is shown in Figure 6.4.

### 6.3.5 Missed Value View

One final possibility is that the badness of failure is related to the goodness that would have resulted from the achievement (i.e., the intrinsic value of the achievement, not of the product). This means, perhaps naturally enough, that the greater the prospective achievement, the greater the badness if failure results. This is analogous to what is sometimes claimed about ill-being for the desire theory, which I discussed in Chapter 4 as the frustration view. Derek Parfit describes the desire theory this way:

<sup>21</sup>In both cases, badness occurs until the goal is reached, at which point there is none.

In deciding which alternative would produce the greatest total net sum of desire-fulfilment, we assign some positive number to each desire that is fulfilled, and some negative number to each desire that is not fulfilled. How great these numbers are depends on the intensity of the desires in question. [...] The total net sum of desire-fulfilment is the sum of the positive numbers minus the negative numbers (1984, p. 496).

When one forms a desire, one's welfare will go up if the desire is fulfilled and down if the desire is unfulfilled. If the theory is symmetrical in value the welfare will go down by the same amount it would have gone up by, otherwise the theory is asymmetrical. There is an analogous form of this view for achievement and failure, where the disvalue of the failure is equal (or related) to whatever the value of the achievement would have been.

Although the values might often correlate, this theory is implausible in its claim that non-achievement always produces disvaluable failure. Some missed achievements are simply neutral and not intrinsically bad. As I argued in Chapter 4, the same is true of desires. If I go for a hike, desire to see a moose but do not, then the lack of moose does not constitute an intrinsic bad for me. Rather, the absence of moose represents one way that the hike could have gone better but did not.<sup>22</sup> Similarly, suppose that in a game of recreational soccer I try to score but the keeper makes a great save. It would have been an achievement had I scored, and so it is a failure that I did not, but it is not a disvaluable failure. I simply miss out on the welfare of the achievement. This asymmetry is in keeping with Hurka's claim that it is better to have tried and failed than never to have tried. If playing recreational soccer could lead to a significant decrease in welfare due to missed shots, we should be much more cautious about participating, but it is implausible to think that such caution is warranted. So while failure and achievement are connected, we cannot conclude that the value of one will affect the value of the other to the same extent each and every time.

## 6.4 Competence

So much for failure. Lack of success seems to play a role in our welfare, and at least some of the time failure seems to constitute an intrinsic bad. However, so far I have said nothing about the other conditions for achievement, namely competence and difficulty. The competent causation condition affirmed by Bradford and Hurka for achievement matters for anti-achievement too. Consider the case of Vanderlei de Lima, a Brazilian runner who was leading the 2004 Olympic Marathon, and was likely to win, when he was tackled by a spectator. De Lima got bronze as a result. He clearly intended to win the race, but does his third place constitute a failure to achieve his intended goal? In one sense, yes, but the cause of the failure means, minimally, that it is less bad than if the third place had been due to some reason within his control (e.g., he was up late drinking the night before). So we can say that, all else equal, a failure is worse when its causes are within one's control (i.e., due to one's incompetence) and less bad when the causes are outside of one's control. This explains both why winning the lottery is not an achievement—the result is not competently caused—and why not winning the lottery is not bad.

<sup>22</sup>Some people are unconvinced by this example. No matter. All I need is one case where, intuitively, the lack of fulfillment of a desire merely fails to deliver increase instead of making one intrinsically worse off. Consider another case: I see a concert by my favourite band and desire that they play my favourite song. That they do not does not constitute an intrinsic bad for me; it merely represents one way I could have been made better off. Krister Bykvist gives the case of wanting an unexpected gift and having a neutral attitude toward not getting it (2009, p. 40). I believe cases involving a neutral attitude toward non-satisfaction are commonplace.

Failures due purely to bad luck are not as bad as failures due to poor planning. This means that in games involving chance, such as poker, losing a hand is only a failure if the player makes an error. She could play flawlessly and still lose due to bad luck with the cards that are dealt. Games that do not involve chance, such as chess, will involve more failure because losing is, in some sense, always within one's control.

Incompetent failures are worse than competent ones, but what effect does incompetency have on success? An incompetent success is obviously worse than a competent one. After all, the former is an achievement (provided that it is sufficiently difficult) while the latter is not. Here, lack of competency is distinct from *incompetency*. This is analogous to justification for knowledge. Lack of justification can involve mere insufficient justification, but what I called *anti-justification* is worse. Both types have a positive, neutral, or negative value. One merely lacks competency when one wins the lottery, which is neither good nor bad. But consider a case of incompetency. Suppose Annie's poker hand has a one percent likelihood of success. Unaware of this, Annie bets all the money she has, and by a complete fluke, she wins. Is her play merely neutral, or intrinsically bad? Hurka thinks it is bad because Annie has "flouted rather than follow[ed] rational norms" (unpublished (b), p. 11). Annie's play made failure probable; whether or not failure occurs is less important on Hurka's view. Notice as well that these forms of incompetency are local. Saying that Annie and de Lima act incompetently concerns only their behaviour during a short slice of time: i.e., when de Lima decides to go out drinking and when Annie pushes all her chips into the pot.

The badness of incompetency helpfully explains some of the pure failure cases I mentioned earlier. For example, I said that, in the case where I am playing recreational soccer, try to score, but fail, this is not intrinsically bad for me. Perhaps what is doing the work in that case is that I played as competently as I could have. While a better player could have made the shot, my lack of success is not bad. However, if we suppose that I failed because I tried an irrational play—suppose I tried to carry the ball between my legs while hopping toward the goal—then this failure is bad for me. Hurka thinks that it would be bad for me even if I had, through some miracle, succeeded. My intuition on that point is less clear, but incompetent failures are certainly worse than competent ones.

## 6.5 Difficulty

Difficulty also plays a role. This condition is much less complicated. As the difficulty of the goal increases, the badness of the failure decreases and vice versa. Suppose the Olympic marathon trials for my country are taking place tomorrow, and despite my lack of training, I intend to qualify. It would be an incredible achievement were I to do so,<sup>23</sup> but it would not be a disvaluable failure not to. Part of the reason for this is the difficulty of the goal. Similarly, an easy failure is worse than a more difficult one. If I fail to win the potato sack race against a group of five year olds, this is a bad failure, given that it should have been easy to win.

---

<sup>23</sup>Perhaps my ability to qualify without training means that I have incredible innate talent, in which case, at least on Bradford's account, my qualification would not be an achievement. Suppose instead that I have average ability but put in all my effort.

## 6.6 Bungling

In Chapter 5 I described Hurka's and Kagan's claim that the worst epistemic state is false belief that is contrary to the evidence, or what I called anti-justified false belief. An explanation for the badness of this state is that the person should have known better. There is a 'reasonable expectation' condition for knowledge that goes unmet when, for example, people think that climate change is a myth. This is made worse when the facts of the individual increase the anti-justification, such as when a climate scientist denies climate change. We are now able to more clearly describe what makes some failures particularly bad, and it has similarities to the worst form of epistemic belief.

Above I affirmed two views involving gaps. The outcome gap view concerns the gap between the goal of one's attempt and the result, while the proximity view holds that a very small gap can be worse than a larger one. These views, I noted, involve none of the features of the theories of achievement I described. Now I wish to describe another type of gap, which involves the combination of three factors: global competence, local competence, and difficulty. Until now, global competence, which refers to the agent's general ability, has played no role in either achievement or in my discussion of incompetency.<sup>24</sup> On the view I am now introducing, it is worse for a globally competent person to fail than a globally incompetent person, and it is even worse for a globally competent person to fail due to incompetently pursuing a goal. I call anti-achievement when one is ordinarily competent bungling, which occurs when someone fails at something she is reasonably expected to do.<sup>25</sup> Bungling is made additionally worse through poor planning or execution (a local form of incompetence).

First the claim about global competence. Consider people who are extremely competent at something: a professional baseball player at fielding ground balls, a basketball player at shooting free throws, a chef at cooking omelettes. For each of these people, their skill in their particular area increases their likelihood of succeeding at their particular task. Here, competence is equivalent to ability, and increased ability means success involves less luck. Suppose the basketball player can regularly make fifty free throws in a row; she is extremely competent at this task. One day the player is practicing her free throws and it is not going well. She *misses* fifty in a row. This is not due to any irrational pursuit (i.e., local incompetence). We are not supposing that she has all of a sudden decided to shoot behind her back or through her legs. Rather, it is an off day. My claim is that a globally competent person's failure is a worse failure for her than it would be for someone with no competence at free throw shooting, and the reason is that the player is reasonably expected to succeed because of her ability. The same applies to the other cases. An error in baseball is defined as a failure to make a play that, based on the competency of the players in the league, the player was reasonably expected to make. In other words, an error for a major league player might not be an error for someone playing in a lower league, and the easier the play for the major league player the worse the failure.

Adding local incompetence makes the failure even worse. Take two professional chefs each needing to make toast. One decides to use a toaster while the other decides—on a whim—to use a knife inserted in an electrical socket. The first chef fails because she forgets to check the dial on the toaster; the second fails because he chooses a bad way to try to make toast. Clearly the former plan is more competent, and the reason is the same as other cases of competency: the action makes success more likely (even though success did not happen). In this case, both agents are globally competent chefs. Either chef's failure

<sup>24</sup>Global competence is a claim about one's ability at some particular thing over time. It does not refer to someone who is good at many things. For example, a professional baseball pitcher is globally competent at pitching because he has a high likelihood of throwing strikes, but this does not mean he is good at other things, such as baking.

<sup>25</sup>I owe the term to Wayne Sumner.

to make toast would be much worse than a young child's failure (which is not bad at all). But of these two professionals, the latter's plan is far less locally competent. Here, competence is equivalent to local likelihood of success—likelihood due to this plan—and different plans increase or decrease likelihood. Both forms of bungling are bad, but this latter type is worst of all. The reason is that, through local incompetence (a bad plan), the agent further increases the gap between the reasonably expected outcome and the actual outcome. Part of being capable is knowing the best way to achieve success, so we expect capable people to choose effective means of achieving their goals. When they do not, they bungle.

In one way this is a strange result. After all, if incompetence is bad for achievement (it decreases its value), then it is strange to say that incompetence can help an agent regarding failure. Should it not be worse to be incompetent and lack success than to be competent without success? My answer is no. To see why, consider which is worse in the following pairs:

- Someone who has never used a bow missing a target or a professional archer missing.
- Someone with no training in statistics making a statistical mistake or a professional statistician making the same mistake.
- A child burning toast or a professional chef burning toast.

Clearly, in each case it is worse for the expert to fail. This is because they are reasonably expected to succeed but do not. It is no surprise for a child to burn toast, and might even be an achievement for one to successfully toast bread. But making toast is trivial for most adults, let alone professional chefs. Here, competence is equivalent to global likelihood of success. Knowing only general facts about each agent—one's a child and one's a chef—you should place your money on the chef.<sup>26</sup>

Difficulty also plays a role in bungling. According to the minimal view of achievement I described, agent-relative difficulty is the only difficulty that matters. Bradford's account also claims this. If agent-relative difficulty is the only relevant type of difficulty, it means that, had the agent not bungled, the result would not have been an achievement. So the values are asymmetrical. An error in baseball is by definition failing to make a play the player should have been able to; this is agent-relative (or at least relative to the league). On Bradford's view, easy plays are not achievements, given that, from the agent's point of view, they are not difficult. It is an asymmetry that not bungling is not an achievement. For example, in a May 2017 baseball game, Rougned Odor, the second baseman for the Texas Rangers, failed to make it to first base because he tripped over himself and fell. This is an agent-relative bungle because professional baseball players know how to run, and Odor is particularly competent.

The features that make bungling bad are analogous to the features that make anti-justified false belief bad. The climate change denier responds incorrectly to the evidence, and by so doing fails to use his capacity for rational reflection. The bungler fails to use his capacity for whatever he is capable of doing, and so in the same way mismanages his competence. They both behave contrary to reasonable expectation. This assessment works with Bradford's view of competence, which involves having knowledge of what you are doing (where the more knowledge you possess about the activity the more competent you are). On this view, knowing how to shoot a free throw means that your achievement is greater (when all else is equal) than someone who just happens to get the ball into the net.<sup>27</sup> Missing a bunch of shots

<sup>26</sup>Though there is no guarantee that the chef will succeed, given that there are such things as bunglers, and that bunglers bungle.

<sup>27</sup>Increasing one's knowledge can also make the task easier, meaning that the achievement is overall less valuable because it is less difficult.

in a row is a bungle for the person who knows what she is doing but not for a person who lacks that knowledge. (One can construct analogous explanations for non-epistemic theories of competence.)

Recall the case of de Lima, the Brazilian marathoner who was tackled at the Athens Olympics when he was on his way to winning gold. I claimed that this failure is not as bad for de Lima, if it is bad at all, as it would be if the reason for his failure to win had been within his control (e.g., he stayed up late the previous night drinking). It matters, in other words, why the failure comes about. If the failure occurs for some reason that the runner should have reasonably expected (e.g., he trips on his shoelaces) then it is a bungle. If failure occurs for some other reason (e.g., a spectator tackles him) it is not a bungle, because there is no reasonable expectation that players will be prepared for it. Some bad luck can still result in a bungle in some sports because professional players can make plays even when they get an unlucky bounce, but the expectation is that they will be able to handle the unluckiness and make the play. The same point applies, *mutatis mutandis*, to non-sports.

I have been discussing bungling so far to always include failures, but we might wonder whether bumbles can occur with success. As Hurka argues, local incompetency on its own is bad, which might suggest that failure is unnecessary for bungling. Sometimes it seems worse when the two forms of incompetency combine with failure than with success. This is reflected in the baseball error statistic, which requires the fielder to misplay the ball in a way that allows a play that would not have occurred but for the error. One can act incompetently but then, in a way, make it up by still making the play, which in baseball is not an error. Similarly, we might imagine Usain Bolt having an uncharacteristically bad start, but despite this he wins the race in a respectable time. In a marathon, a runner might trip on her laces, but the effect on her performance will be small. It would be worse if the error cost her the race. But in other cases the outcome seems irrelevant. Consider Annie's bad poker play, but add that Annie is among the world's best players. In this case, the play is bad regardless of the outcome. If she miraculously ends up winning the hand it does not take away from her bad play. (Poker players use the term 'resulting' for the fallacy of basing the quality of the decision on the quality of the outcome.)

I think much of this comes down to how we think of outcomes. Chess and poker plays are discreet in a way running and baseball plays are not (and of course baseball is more discreet than running, for which there are no plays). Perhaps this means there is no difference, so we ought to ignore outcomes altogether. But that strikes me as leaving out important information, which is that failure sometimes makes a difference.

I propose that bungling is the worst type of anti-achievement. It is bad because it involves a large gap between the reasonably expected outcome and the actual outcome, and in this way it is another view involving gaps.

### 6.6.1 Combining the Accounts

Although bungling does not require failure, we can now examine what happens when they are combined. Sometimes the bungle will result in a failure drastically below the goal, which invokes the outcome gap view. A fielder could intend to throw the ball to first base but instead throw it into the ground or the stands. In other cases, the bungle will result in a failure *just* below the goal, which invokes the proximity view. The fielder might fumble with the ball but still manage to throw it, only to see the runner get to the base just before the ball. Which is worse?

For those attracted to giving more weight to the outcome gap view over proximity, bungling and outcome have a positive relationship. The greater the bungle, the greater the gap between the goal and

the actual outcome. They will say that failing by such a large amount worse. However, the issue is more complicated for those who give more weight to proximity, for on that combination, bungling and proximity have a negative relationship. A smaller bungle will lead to an increase in proximity (a closer call), but will be worse overall than if the gap were larger (at least to a point).

## 6.7 Conclusion

I have argued that insofar as achievement is an intrinsic prudential good, there are intrinsically bad counterparts to it. I have also shown that a theory of achievement does not trivially determine the correct account of anti-achievement and its badness. Indeed, I have shown that, while the badness of anti-achievement is related to the goodness of achievement, the one does not simply mirror the other. Instead, there are six factors that determine intrinsic badness: global incompetence, local incompetence, lack of difficulty, a large gap between the goal and the result, a small gap between the goal and the result, and the distance between the result and the reasonably expected result. I argued that the worst type of failure is bungling, which involves a large gap between the result and the reasonably expected result. This is due to a lack of difficulty and the gap between global and local competence.

# Chapter 7

## Hybrid Theories

### 7.1 Introduction

In this chapter I discuss ill-being for subjective-objective hybrid theories, which claim that well-being consists in directing a positive attitude (pleasure, love, desire, enjoyment, etc.) toward something objectively good (something that is excellent, worth loving, etc.). These elements combine non-additively in an organic unity, which is G.E. Moore's term for when elements combine to have a total value that is different from the sum of the values they would have on their own. For example, and to simplify matters slightly, a defender of additivism (a non-hybrid theory), which is a type of pluralism, will claim that the total value of pleasure taken in an achievement is just the sum of the value of the achievement and the value of the appropriate attitude of pleasure in the achievement. Combining the elements does not change the value. In contrast, according to the hybrid theory, the elements combine to make a total value that is greater than the sum of the parts, such that an achievement on its own might have no value, as might pleasure in an object lack value when the object does not exist, but when combined they produce significant well-being.

In contrast to the other theories I have discussed,<sup>1</sup> hybrid theories are rarely categorized or discussed as a group.<sup>2</sup> Nevertheless, they have been defended as a way of avoiding the shortcomings of purely subjective or objective theories on their own. Hybrids raise unique and difficult questions for ill-being, which makes them worth considering.

### 7.2 The Structure of Well-Being

As a theory of well-being, a hybrid theory combines two elements such that their interaction produces a non-additive value (which distinguishes it from pluralism). The majority of hybrid theories that have been defended combine subjective and objective factors, so in what follows I will focus on this subjective-objective type.<sup>3</sup> On this combination, one is well off to the extent that one has a positive attitude toward some existing objective good, so the two goods are the positively directed attitude and the objective

---

<sup>1</sup>It will be useful to have a term for all non-hybrid theories, so I will call them 'pure'. Pure theories can have more than one good, as is the case with most objective theories, which are *pure pluralist* theories.

<sup>2</sup>Some exceptions are Woodard (2015) and Kagan (2009).

<sup>3</sup>I will now only speak of 'hybrids', by which I mean 'subjective-objective hybrids'. For examples of hybrids that do not take this form, see Woodard (2015).

good. The hybrid can add other goods, such as the value of pleasure as pleasure, but these are not part of the organic unity.

An application of the organic unity principle is Moore's defense of the goodness of aesthetic appreciation, which is one of his principal goods. Though Moore is not discussing well-being specifically, his view of aesthetic appreciation provides a good analogue to welfare. Initially, Moore argues in *Principia Ethica* that beauty is good. We are asked to imagine two worlds. The first is as beautiful as possible; the second is a heap of filth. Moore says that it would be rational to bring about the first world over the second, even if no human or other conscious being will ever perceive it (1903a, pp. 83–84).

But later on Moore says that beauty on its own is not very valuable. The two principal goods on his view are the “pleasures of human intercourse and the enjoyment of beautiful objects” (1903a, p. 188). Both of these goods are complex in that they involve pleasure directed at an appropriate object. While Moore thinks that pleasure without a good object (i.e., because the object does not exist) is still somewhat good on its own, pleasure as pleasure is not very good. When pleasure and the object combine, they become the greatest possible goods. The attitude must match the object. On Moore's view, pleasure taken in an ugly object is bad (1903a, p. 190). Another example of the principle of organic unities is knowledge, which on its own has little or no value according to Moore. However, if an agent has a belief in the reality of the object she is admiring, this increases the total value if the belief is true and decreases the value if it is false (1903a, pp. 194–199).

The principle of organic unities means that objects, which on their own have no value, can combine to be intrinsically valuable. Independent bads can also be combined to produce an outcome that is more valuable than the sum of the elements. For example, Moore thinks retributive punishment has this form (1903a, p. 216). Pain is independently bad, but when it is appropriately had by a deserving person—someone with an evil character—the total value is better than when an evildoer goes unpunished (though it is still not positively good).

Organic unities can also apply to welfare. For example, Richard Kraut's developmentalism, which I discussed in Chapter 5, expands on a subjective-objective hybrid he previously defended, which has the following conditions: “one must love something, what one loves must be worth loving, and one must be related in the right way to what one loves” (Kraut 1994, p. 44), adding that “for a human life to go well one must love something worth loving” (1994, p. 47). Susan Wolf also defends a hybrid theory, holding that meaningfulness is a constituent element of well-being. On her view, “Meaning arises when subjective attraction meets objective attractiveness” (Wolf 1997, p. 211). In other words, for our lives to go well we must be actively engaged in worthwhile projects. Derek Parfit also describes one form of the hybrid view in *Reasons and Persons*:

What is good for someone is neither just what Hedonists claim, nor just what is claimed by Objective List Theorists. We might believe that if we had either of these, without the other, what we had would have little or no value. We might claim, for example, that what is good or bad for someone is to have knowledge, to be engaged in rational activity, to experience mutual love, and to be aware of beauty, while strongly wanting just these things. On this view, each side in this disagreement saw only half of the truth. Each put forward as sufficient something that was only necessary. Pleasure with many other kinds of object has no value. And, if they are entirely devoid of pleasure, there is no value in knowledge, rational activity, love, or the awareness of beauty. What is of value, or is good for someone, is to have both[:] to be engaged in these activities, and to be strongly wanting to be so engaged (1984, p. 502).

Defenders of hybrid theories have described the subjective element in different ways. The attitude can mean, among other things, being pleased by, taking happiness in, loving, or enjoying the good in question. Whatever the positive attitude is, it must be *directed at* a genuine objective good for it to contribute significant value. One interpretation is that the agent must actively pursue the good in question. For example, to love knowledge is to actively work to bring it about. Further, the pleasure must be directed at the object for its own sake instead of as a means. A professional athlete who loves winning because it increases his salary is not loving his achievements as achievements, so the organic unity does not occur. Finally, to distinguish the hybrid theory from additivism, the total value must be greater than the mere sum of the two separate values. Misdirected attitudes might still contribute to well-being because they are good as pleasure, but the greatest amount of welfare results from a positive attitude toward an objective good. In what remains, I will generally talk of pleasure or positive attitudes, but these can be substituted for any preferred subjective state.

Standing in the correct relation to the good in question is also important. It is not enough that one has pleasure at the same time as an objective good; the attitude must be directed at the good, as when one takes pleasure *in* one's achievement. The organic unity would not occur if one had pleasure merely at the same time as one's achievement but for some other reason, such as from a backrub. Further, the good must actually exist. Suppose I take pleasure in what I believe to be a new weightlifting personal best for me, which, suppose, is an achievement. My pleasure is the good of an appropriate attitude, which is a single source of value, and a view can hold that this contributes welfare to me regardless of whether or not I actually did break my previous best. But the hybrid theory says that I get no (or at least less) value from it when I miscount the weight and my belief is false. In other words, on the hybrid view, the organic unity comes from the appropriate attitude directed at an actually existing object.

In addition to the connection between the attitude and the object, the type of good also matters. One possibility is that only goods that relate to one's life or that one has obtained count. So if one takes pleasure in one's own achievement (and achievements are objectively good) then one is better off than if the achievement occurred with no pleasure or pleasure in something that is not objectively good. Following Kagan, I call this the *possession condition*, according to which we need to possess the goods in question for them to contribute to our welfare (2009, pp. 254–257).<sup>4</sup> Endorsing the possession condition means that taking pleasure in genuine goods that are not one's own does not contribute to one's well-being. Suppose I take pleasure in *your* achievement. If the possession condition is true then I am not better off, even if I might be more virtuous.<sup>5</sup>

To deny the possession condition means that any good counts, provided that I take pleasure in it.<sup>6</sup> Kagan describes the following case (2009, p. 256). Suppose I stop to admire a grove of redwoods. The trees are not mine in any sense, yet suppose that their beauty makes them objectively good. If we deny the possession condition, my taking pleasure in redwoods makes me better off so long as the redwoods actually exist. Similarly, my taking pleasure in the achievements of others also contributes to my well-being, again with the caveat that the achievements must actually exist. Denying the possession condition changes the scope of the theory. In Kagan's initial description, he talks about the objective list theory, which necessarily involves possession because it is a theory about *one's* life. But with the redwood example, the scope changes to objective goods *generally*, not necessarily things that are only

<sup>4</sup>For his part, Kagan denies that possession matters. What I claim applies regardless of whether the possession condition applies, but sometimes I will speak of 'possessing a good' simply as a shorthand for it being relevant to one's welfare.

<sup>5</sup>As will become clear below, this type of case is possible for only one type of hybrid theory (permissivism).

<sup>6</sup>A more complex theory is possible, where the possession condition applies to some objects but not others. For simplicity I will set this possibility aside.

good for the individual who does the possessing.

It is the combination of a subjective attitude directed at an existing objective good that makes one well off. On one type of hybrid theory, this is the only way one can be made better off (i.e., a positive attitude directed at an existing objective good). Both elements are necessary, and they must combine in the correct way. I call this the *restrictive view*. Endorsers of this view include Kagan (2009, p. 255) and Kraut (1994, p. 44).<sup>7</sup> The above quotation from Parfit also leans toward restrictivism.

Some define hybrid theories as necessarily restrictive. Christopher Woodard calls this definition the *joint necessity model* (2015, p. 164). This formulation means that hybrid theories are usually a special form of monism, where the individual elements on their own are not good, but the *combination* of the two elements produces a single good. Woodard says that hybrid theories can endorse the joint necessity model, but a more general defining feature of hybrids is holism, which, as I described above, is best illustrated by Moore's principle of organic unities. According to holism, the value of the whole depends not just on the individual values of the parts, but also on how the parts interact. In other words, the contribution one element makes to one's well-being depends on at least one other element (Woodard 2015, p. 168).

Holism also has a permissive option. This version says that maximal well-being is the result of a positive attitude directed at an existing objective good, but pleasure in a *non-existing* good—i.e., just the appropriate attitude—is sufficient for some welfare, as is the presence of an objective good, regardless of the presence of a positive attitude. For permissivism, there can be three factors: (1) The object (e.g., achievement), (2) the intentional attitude (e.g., pleasure in the achievement), and (3) other goods, such as the pleasure as pleasure. This means that, on permissivism, appropriate pleasure and an objective good will always make some positive contribution to well-being regardless of the correct connection between them. In contrast, restrictivism denies that an appropriate attitude or the object on its own can ever be good. Instead, there must be an attitude directed at an existing object. Restrictivism also denies that pleasure as pleasure can ever be good, but permissivism allows this (though it is not part of the organic unity). In fact, permissivism allows any additional good.

Another distinguishing feature of permissivism is that it must accept the possession condition if the objective good can contribute any value on its own. This means the list of objective prudential goods and bads will more closely resemble a pure prudential objective theory of the sort I discussed in Chapters 5 and 6. After all, while one can have an appropriate attitude toward an unpossessed beautiful object such as the Grand Canyon, that still requires an attitude. No good can come to me if you achieve something but I have no attitude toward it. In contrast, because on restrictivism the attitude and the object are jointly necessary, goods such as the achievements of others can count even though they would contribute nothing on their own to the non-achiever.

So if a subjective or objective element on its own is good without the other then the theory is permissive; if both are necessary for value then the theory is restrictive.<sup>8</sup> Taxonomically, I am open to both views being called hybrid theories. Some proponents reject the permissive option—I discuss this

<sup>7</sup>Kagan's view is complicated by his distinction between being well off and having one's life go well (1994). He thinks that possessing an objective good with no corresponding pleasure increases the value of one's life but does not increase well-being (2009, p. 257). Also, although he focuses on restrictivism, he does not endorse the hybrid theory over other theories of welfare (2009, p. 253).

<sup>8</sup>In fact, there are three types of permissive theory: (1) appropriate pleasure contributes welfare but the possession of an existing objective good without pleasure keeps welfare unchanged; (2) the opposite, where appropriate pleasure on its own does not change welfare but an objective good without pleasure increases welfare; and (3) where either element on its own without the other contributes welfare. This is an interesting question deserving of further investigation, but in what follows I will stick to the third option.

below—but I am being inclusive for the purpose of thoroughness. As we will see, each option produces different accounts of ill-being.

One difference between adjusted subjective theories (the topic of Chapter 3) and hybrid theories is that AS theories give the subjective side priority by endorsing what Kagan calls the hedonistic constraints. The first constraint says that pain, regardless of its object, can never be good for someone. The second says that pleasure, regardless of its object, can never be bad, though it can be neutral (Kagan unpublished, p. 8). This is brought out in the way AS theories are usually framed in terms of discounting pleasure. A pleasure taken in an object might be discounted, but never such that it becomes bad for the agent. In contrast, a hybrid theory can reject both constraints, so pleasure can sometimes be bad (when it is misdirected) and pain can sometimes be good (when it is properly directed). This is because hybrid theories are not purely subjective. They use the independent value of the object at which the subject directs an attitude to determine the combined total. Another difference is how each camp characterizes the object of the pleasure. Even for a restrictive hybrid theorist, who thinks that experiencing the good is only valuable if pleasure is directed at it, the good will still be characterized by being worthy of value (e.g., deserving of love, excellent). The AS theorist, meanwhile, does not treat the object as valuable or worthy of value.

Some might be more inclined to group AS and hybrid theories together, and perhaps they are both species of a larger family. While I believe there is value in treating them separately, nothing significant turns on this taxonomy.

### 7.2.1 Attractiveness

It will be helpful to describe the attractiveness of the hybrid theory to show why proponents have taken the stances they have on some of the issues just discussed. This is a good place to make the same pronouncement I have for all the other chapters: It is not my goal to defend or even assess the general plausibility of the hybrid theory as an account of well-being. Instead, my interest is to explore its plausibility as a theory of ill-being. By doing so, I will also explore the plausibility of some asymmetries between the good and the bad.

The purported advantage of hybrid theories is that, by combining two elements, they avoid some problems non-hybrid views face. A pure, non-adjusted subjective account ‘gets the insides right’: being happy, having a positive attitude toward life, and so on. The downsides of such theories are also familiar. For instance, they hold that all pleasures of the same strength, regardless of their source, are equally good, so pleasure taken in watching reality television alone is equal to a quality meal with friends. They also hold that experiences rooted in false beliefs, such as when one is attached to the experience machine and those of the deceived businessman, are as valuable as if one’s beliefs were true.

In contrast, pure objective theories accord with the intuition that there are objectively better and worse ways for our lives to go, and that these are due to the goods and bads we possess. Having friends makes us better off, not because we are always happier, but for some non-subjective reason (e.g., we are social beings, or having friends is good for its own sake). The cost of a purely objective account is its claim that we can be well off without a corresponding positive attitude. I can have achievements, friends, and knowledge, yet get no enjoyment from these goods. There is something strange about claiming, as objective theories do, that one can have a high level of welfare without believing that one’s life is going well, or even while being unhappy.

Hybrid theories offer a way out of these problems. If well-being consists in taking pleasure (or having

some other positive attitude) toward an existing objective good, then the theory can avoid the conclusion that reality television is as valuable as friendship. Similarly, on the hybrid theory, it is no longer the case, as it was with the pure objective theory, that I can be well off (or at least *very* well off) through the possession of some goods while being miserable or getting no enjoyment from them. On the permissive view, we can get some increase in welfare by possessing an objective good, but the greatest value also requires a positive attitude.

We are now in a position to notice the source of the disagreement about what constitutes a hybrid theory. Permissivism does not entirely avoid the two criticisms I have been discussing, for both unintuitive results of the pure theories are still possible. However, it improves on pure theories by holding that the greatest amount of welfare cannot occur while watching reality television or while possessing objective goods but feeling miserable (or nothing). Only the restrictive option, which holds that a positive attitude directed at an existing objective good is necessary for well-being, fully avoids the criticisms (though as I describe below, it has its own problems).

Hybrid theories purport to avoid common criticisms of pure subjective theories such as the experience machine, and it is worth exploring exactly how hybrid theories might avoid this type of objection. The objection is that the person attached to the machine believes she is having some desire fulfilled (e.g., climbing Everest) when in fact she is just floating in a tank, so her life is not really going as well as she believes it is. I described the hybrid theory's common subjective-objective form as having a positive attitude directed at an objective good. However, on closer inspection, that description on its own will not do. After all, when attached to the experience machine, one can have a positive attitude directed at some objective good, which in the Everest case is the good of achievement. What makes the experience machine less valuable is not that the purported good is not actually good, which, we might imagine, is the problem with reality television. That is, the hybrid theorist is not claiming that the agent is mistaken in her belief that achievements are objectively valuable (supposing, of course, that she has this belief). Instead, the hybrid theorist is claiming that machine life is less valuable because the good *does not actually exist*.

This means there are two different ways for the connection to fail between the attitude and the object. One can be mistaken about the object's value, or one can be mistaken about the object's existence. In his discussion of aesthetic appreciation, Moore calls these mistakes 'errors of taste' and 'errors of judgement' (1903a, p. 193). To cover a broader range of objects, we might instead call errors of taste 'errors of orientation', or having an inappropriate attitude toward some object. An example of an error of orientation is taking pleasure in your lack of autonomy; an example of an error of judgement is believing that you have climbed Everest when you are hooked up to the experience machine. Moore thinks that errors of orientation are worse, but both errors reduce value. So there are two required connections: One has to have the correct attitude, and the object must actually exist. This produces the following definition of the hybrid theory:

*Subjective-objective hybrid theory:* In order for someone to have the highest amount of well-being, one must have a positive attitude that is directed at an object, where the object meets the following conditions: (1) it must be good and (2) it must exist.

To this list, some add (3) the object must be possessed by the agent.

The type of object plays a role in the type of failed connection. Cases involving beauty usually involve mistaken value (an error of orientation). I might hear a song or see a piece of art and believe it

to be beautiful, when in fact it is ugly. Maybe the painting is kitsch and so possesses no value, or maybe I am just inflating the value above where it should be. In contrast, knowledge claims will usually involve mistaken existence. This is the type of mistake that occurs both on the experience machine and with the deceived businessman. In all of the knowledge cases, we can suppose that the agent is correct about the value of the object. That is, it really would be valuable for the deceived businessman to be in a successful marriage and have children who love him, and therefore the enjoyment would be appropriate. The mistake is not about the value, but rather the existence of the object.

The difference between mistaken value and mistaken existence makes the permissive hybrid theory additionally complex, as it might be the case that the different types of failed connections have different effects on welfare (as Moore thought). This is only true of permissivism; restrictivism says that anything less than the full combination is valueless. Consider again the experience machine, which we might describe as a case of taking pleasure in the object of a false belief. On closer inspection, it would be a mistake to diagnose this case as one involving a failure to take pleasure in knowledge. This is a mistake because in non-machine cases we do not describe the value of climbing Everest as having a justified true belief that one has climbed Everest.<sup>9</sup> It involves a different good: achievement, where it is the case that the achievement occurs regardless of the agent's beliefs about it. This is the good that does not exist on the machine. The hybrid theory also has to address non-existence cases. For example, while hedonism says that taking pleasure in having proved Fermat's last theorem is just as good whether or not I have actually proved it, the hybrid theory claims there is a difference.

Switching back to the positive formulation, we can ask what role each criterion makes in the overall value of a positive attitude directed at an existing object. Regarding (1) from the definition above—that the object must be valuable—this leads to a difference between permissive and restrictive accounts. Both agree that (1) is necessary for full value, but permissivism claims that an appropriate attitude on its own is still good, while restrictivism denies this. What about (2), that the object must exist? To avoid the experience machine objection, non-existence of the good must lower the total value by some amount. So the question is how much less good results from taking pleasure in something that does not exist. The total value is either partially or completely reduced. Here, the first option is preferable. Note that it matters what you think the experience machine thought experiment shows. Jennifer Hawkins says that, while Nozick seems to be arguing only that experiences are not all that matter for well-being, many people go further to conclude that we should never plug into the machine (Hawkins 2015, p. 359). To reach this stronger conclusion, it has to be the case that machine life is either worthless or has negative value. Restrictivism claims this, as the pleasure of machine life is valueless because there is not the correct connection between the good (which, in reality, does not exist) and the attitude. But this is too strong. While being attached to the experience machine and having blissful experiences is less valuable than experiences that match reality, it would still be better to be on the machine having appropriate attitudes than having no experiences at all. Further, and also in contrast to restrictivism, it is plausible that pleasure as pleasure is still good, which makes machine life somewhat valuable. Such pleasure is outside the organic unity that makes hybrid theories distinctive, but it is present whenever one part of the organic unity (i.e., the pleasure) is present.

Existence without goodness might still produce some value, as permissivism can claim is the case with pleasure directed at a neutral object (because the pleasure is simply good as pleasure). It is also

<sup>9</sup>Of course, in most cases when one has climbed Everest one has the belief that one has climbed Everest. Usually attitudes involve a belief about the object's existence (see Hurka 2001, p. 162).

possible for pleasure directed at a good object to be valuable, even though the object does not exist. This is for two reasons. First, because the pleasure is good as pleasure, and second, because the pleasure is appropriately directed. This is the case with some theories of appropriate attitudes, such as Moore's, which hold that one can act appropriately by taking pleasure in a good object, whether or not that object actually exists. The clearest case of this is with fiction (1903a, p. 193). Being pleased that the hero succeeds is good on Moore's account even though the hero does not really exist. While it is less good than being pleased in a factual case, it is still good. We can say the same of the experience machine: even though my friends are not winning all the Olympic medals and curing all the diseases in real life, it might still be good (because it is appropriate) to be pleased by the experience of their success on the machine. Such accounts hold that, whether or not one's beliefs are veridical, having the appropriate attitude is good for the agent. Hurka goes farther by claiming that the falsity of a belief does not affect its value, though he is discussing virtue, not welfare (2001, p. 162).

To conclude this section, I will describe two general criticisms of the hybrid theory to push back against the attractiveness claim (but not to resolve the dispute one way or the other). The first comes from L.W. Sumner, who gives two different objections (1996, pp. 163–166). First, because Sumner defines well-being as whatever makes a life better for the one who leads it, he is skeptical of any attempt to introduce some independent non-subjective value requirement. Such a requirement entails that there is a fact of the matter regarding how prudentially valuable different pursuits are, which Sumner finds implausible. (I discussed his theory in detail in Chapter 3.) Speaking about perfectionist value in particular, he says that “it just does not seem true that my life automatically goes better for me if the goals I am pursuing rank higher rather than lower from this external standpoint” (1996, p. 165).

Sumner's second criticism is that an external imposition of value is “objectionably dogmatic.” In a case I discussed in Chapter 3, Sumner has us imagine training for years to become a priest, during which time our happiness fluctuates depending on how well we are doing in our sacerdotal mission. But then, after growing older, we place less value on the pursuit of the priesthood than we once did. Sumner claims that we typically would not re-evaluate our level of happiness when we look back on those years, but he further asks if we should re-evaluate our well-being:

Will you automatically conclude that your life was going worse at that time than you then thought it was? Or might you conclude that, while you were engaged in a pursuit which you could not now take seriously, none the less that was not a bad way to spend your life *then*? As in the case of gaining retrospective empirical enlightenment, there seems to be no right answer to the question of how to respond to shifts in personal values or standards: at the extremes you can either write off some earlier part of your life as a complete waste or accept it as an essential component of your past identity [...]. Many intermediate options between these extremes are also available; which one you choose to embrace is surely up to you (1996, pp. 165–166).

Alexander Sarch (2012) argues in the other direction by focusing on the absence of objective value on its own.<sup>10</sup> This criticism applies to AS theories and restrictive hybrid accounts (and shows why hybrid proponents might be more attracted to the permissive version, despite it having its own problems). Sarch considers the case of a person who pursues and achieves ambitious and worthwhile goals, yet due to a

<sup>10</sup>Recall from Chapter 3 that Sarch is discussing both adjusted subjective theories and hybrid theories together.

genetic condition is unable to feel enjoyment or suffering.<sup>11</sup> He thinks it is obvious that this person's life is valuable despite its lack of positive attitudes. Restrictive hybrids give the exact same welfare level whenever one's subjective level is zero, regardless of any objective factors. He finds this implausible: "[A] life of no enjoyment or disenjoyment that is spent exclusively watching paint dry would seem to be much worse for the one whose life it is than a singularly successful life that contains a range of remarkable experiences and achievements, but that also happens to contain no enjoyment or disenjoyment" (2012, p. 445).

### 7.3 The Structure of Ill-Being

We can now begin to explore what the hybrid theory says about ill-being. The distinctive feature of any subjective-objective hybrid is that the most ill-being comes from the organic unity of an objective bad and an inappropriate attitude, such as pleasure in one's failure. This is the converse of the hybrid claim that the most well-being comes from the combination of an objective good and an appropriate attitude. Enjoyment is the correct attitude toward objective goods, so for full well-being the good must exist and one must enjoy it. Likewise, full ill-being results from taking enjoyment in an existing bad. Although there is still pleasure, its appropriateness has changed, so we now have the contraries of the two conditions for full welfare. Hybrids can once again be separated into restrictivism, which claims that both subjective and objective elements are necessary for ill-being, and permissivism, which claims that either element on its own can contribute ill-being, but the greatest illfare comes from an organic unity between them.

The defining feature of both forms of the hybrid theory is the claim that there is an organic unity, so the main question is whether such a unity is plausible. At this point, it is an open question. We cannot assume there is one for illfare just because it is plausible for welfare.

#### 7.3.1 Restrictivism

According to restrictivism, ill-being only arises from an objective bad and an inappropriate attitude directed at the bad. This is the analogue of its claim that well-being only results from an objective good and an appropriate attitude. To assess the plausibility of this view, we can consider whether ill-being should result from any other combination. First consider an objective bad with an *appropriate* attitude. According to restrictivism, this combination produces no ill-being. (It is the contrary of the claim that an objective good with an inappropriate attitude, such as pain at one's achievement, contributes no well-being.) Consider some cases involving this combination. So as not to base these assessments on my previous discussion of objective bads in Chapters 5 and 6, I include heteronomy as a possible bad. For each of them, assume that the attitude is the appropriate strength. 1) Edmund fails to climb Everest, and as he is walking back to base camp he hates his failure. 2) Truman becomes aware that he is being manipulated and is angry about it, even as the manipulation continues. 3) Sen's landless labourer recognizes his awful situation and responds with appropriate grief. 4) After years of media appearances in which she denied climate change, Karen realizes she was mistaken and feels bad about her anti-justified false belief. 5) Kayla's beloved dog dies and she is deeply saddened.

<sup>11</sup>If this example is too difficult to imagine, consider Hurka's definition of the pure professional athlete, who plays the sport only as a means to make money (2001, p. 189). Even though she gets no pleasure from playing, her achievements still contribute to her welfare.

If restrictivism for ill-being is correct, each of these cases should produce zero ill-being because the combination of an objective bad with an inappropriate attitude does not occur. Yet it is clearly the case that these people have ill-being. In the case of Truman, how could continued manipulation paired with extreme anger nullify the badness of the heteronomy? Similarly, concluding that Kayla has no illfare as she cries over her dead dog is absurd, so restrictivism produces the wrong result. There are two ways that Kayla's case is distinct from the others. First, the death of her dog does not meet the possession condition. Second, hers is the clearest example of the badness of pain as pain. In Truman's case, we think it is the badness of the manipulation that is doing the work, but in Kayla's case, although her dog's death is certainly bad, her sadness is doing more work. One problem for restrictivism is that it does not include the badness of pain as pain, just as it does not include pleasure as pleasure. Pain as pain is outside the organic unity, which, in Kayla's case, produces a particularly unappealing result by claiming that only the organic unity matters. As the other cases show, it is also a problem that restrictivism does not include the badness of the objective bad on its own.

In contrast, additivism delivers a better result in these cases. Kayla's sadness is bad for her as pain but good for her as an appropriate attitude; the additivist can decide how much weight each factor receives to produce the correct result. Similarly, Truman's manipulation is bad for him, as (perhaps) is his anger, but there is some goodness to his appropriate attitude. Again, the exact total value is an open question, but additivism has the resources to deal with these cases in a way restrictivism lacks.

Kagan gives an argument that, while not directly responding to the objection I have raised, is an attempt to avoid similar implications. Kagan's claim is that, while an objective bad paired with an *appropriate* attitude is not as bad as an objective bad paired with an *inappropriate* attitude, it is still bad. He makes this move by defending the first hedonistic constraint, according to which pain, whether appropriate or not, can never contribute to an organic unity to produce neutral or positive welfare (unpublished, pp. 7–8). Recall that Kagan only discusses restrictivism in his paper. For example, he says this about well-being:

Perhaps well-being requires both the actual possession of various objective goods and the taking of pleasure in the possession of those goods. Perhaps neither ingredient taken alone suffices for well-being, but taken together they do. [...] According to this idea, even if you have an objective good in your life, the mere presence of that good doesn't enhance your well-being unless you also take *pleasure* in the good (unpublished, p. 3).

He mentions what I am calling permissivism, but then sets it aside (unpublished, p. 4). To take pleasure in an objective good is to have an appropriate attitude toward such a good, and it is the existence of the good and the appropriate attitude that make the organic unity for well-being. Contrariwise, ill-being comes from an objective bad with an *inappropriate* attitude, as those are the opposites of the good and the appropriate attitude. Kagan's claim is that pain is always bad for the agent, which would mean that all of the cases I give above produce ill-being, whereas I claimed that they are neutral according to restrictivism. Kagan's move does address my objection, but the result is no longer restrictivism, for on his view pain as pain now trumps everything else. But pain as pain is not part of the organic unity, so he has switched to permissivism or additivism. While one of these views might be correct, switching is not, of course, a defence of restrictivism.

The second type of case for restrictivism concerns the existence of the object. Restrictivism holds that pleasures on the experience machine, though they might be appropriate, do not produce well-being because the object, which is the source of the pleasure, does not actually exist. Machine life is

valueless according to restrictivism (and the same is true of other false pleasures, such as the deceived businessman's). Restrictivism for ill-being makes the same claim: the object must actually exist for it to produce ill-being.

To test the plausibility of this claim, consider the following case. Vicious Romeo sees Juliet sleeping and believes she is dead. He is delighted, which is an inappropriate attitude. According to restrictivism, Vicious Romeo being happy at Juliet's death is not bad for him unless she is actually dead. In other words, the object must actually exist to get the organic unity. This is the wrong result. An inappropriate attitude such as pleasure in the death of a loved one should produce ill-being whether or not the lover is actually dead, but restrictivism claims that Juliet being asleep makes all the difference to Romeo's illfare, which is implausible.

Other cases show the same result, such as Hurka's variation on the Milgram experiments (2001, p. 163). Restrictivism holds that someone who takes pleasure in inflicting pain on someone else is badly off only if the recipient is actually being pained, but this is the wrong result. While it might be worse if the pain is real, it is still bad even if there is no actual pain. The previous two cases deny the possession condition, but we get the same result when we switch to bads within a life. Suppose that Roger attempts a four-minute mile but is pleased when the clock shows 4:00.01, even though the clock is faulty and he actually ran 3:59.99. Restrictivism says that the badness of the inappropriate attitude is nullified because the bad does not actually exist, but this is implausible.

I have described restrictivism as holding that ill-being only results from an objective bad and an inappropriate attitude, but someone might dispute this, holding that at least one other combination—namely, pain taken in an objective good—also produces ill-being. This combination also involves a mismatch between the object and the attitude, so one might reasonably expect it to be bad. If so, this means that restrictivism is structurally asymmetrical, for it holds that a condition for ill-being does not apply to well-being. More importantly, notice that this does not help restrictivism avoid the problems I have described, for we can easily generate analogous cases to reveal the implausible implications of claiming, for instance, that pain in a good is only bad if the good actually exists. A complete account of restrictivism can add other ways of generating ill-being, but that just adds additional ways the theory is mistaken.

Restrictivism claims that the most (or all) ill-being results from an inappropriate attitude directed at an existing objective bad. I have shown that restrictivism is implausible both in its claim that an appropriate attitude in an objective bad is neutral and in its claim that an inappropriate attitude in a non-existent bad is neutral. Therefore, restrictivism for ill-being should be rejected. This is a significant result, for nearly all defences of the hybrid theory of well-being are restrictive. We also saw that additivism fares better. Although the details will matter, it has the ability to weigh the factors individually to determine the correct total value.

### 7.3.2 Permissivism

Permissivism avoids the aforementioned problems with restrictivism. Instead of claiming that an appropriate attitude in an objective bad produces no ill-being, permissivism holds that the total is less bad than an inappropriate attitude in an objective bad, but the objective bad is still bad. Truman's manipulation is still bad for him even though he is appropriately angry about it, as is Karen's anti-justified false belief. The same is true of inappropriate attitudes in non-existent bads: though not as bad as when the object exists, the combination is still bad. Further, permissivism can include the badness of pain as

pain, though it cannot include any separate badness in non-possessed bads, such as the death of Kayla's dog. There, the resulting illfare is because she is pained by it, but this is not part of an organic unity (and in this way is no different from additivism).

The main competitor for permissivism is not restrictivism, which I have shown to be implausible, but additivism. The question is whether an organic unity between an inappropriate attitude and an objective bad produces the most ill-being, as permissivism and restrictivism both hold, or whether it is more plausible to simply add up the bads. If an organic unity for ill-being is implausible, both forms of the hybrid theory fail. Unfortunately, this issue is difficult to settle. The hybrid theory says that pleasure in a bad produces an organic unity, so given some case, it should be obvious that the total disvalue of the combination is greater than the mere addition of the elements. After all, the organic unity is the defining feature that makes the hybrid theory an attractive theory for well-being, so we should expect some similarly-attractive result to carry over to ill-being.

Consider Karen, whom I mentioned above. Now, instead of feeling bad about her anti-justified false belief that climate change is a myth, she takes pleasure in it. I will introduce some values to show the difference between the two views. Karen has a bad, the belief (-10), and the inappropriate attitude (-10). Additivism says that this episode is worth -20, while the hybrid theory says that the total value is less than -20 (i.e., worse than mere addition). Which result is correct? I have no clear intuition. When I think about similar cases, I find the same lack of intuition that there is an organic unity between the two elements. If the hybrid theory is correct, it should be clear that there is an organic unity, but it is not obvious that there is one. This is a problem for the hybrid theory.

An explanation for this result comes from returning to well-being. The motivation for the organic unity between an objective good and an appropriate attitude is that the approach avoids the problems of the pure theories. One way of thinking about the hybrid theory is that, compared to a pure theory, it reduces the value of the individual elements when they are in isolation, while increasing them when they are together. For welfare, there is a big difference between mere addition and the organic unity. So on permissivism, a good on its own might be +1 and a positive attitude toward the good is +1, but together they equal +20. But when we switch to ill-being, there is no similar result. It is not obvious that an inappropriate attitude would be worth -1, an objective bad would be worth -1, but the combination totals -20. This means that there is a structural asymmetry.

The hybrid theory plausibly claims that an organic unity occurs for welfare, but there is no clear reason for holding that the organic unity carries over to illfare. Therefore, the hybrid theorist for welfare should be an additivist for illfare. This makes for an interesting asymmetry, for there appears to be a boost in welfare from the organic unity that does not apply to illfare, perhaps with the result that there is a greater possible amount of welfare than illfare. There might be a higher amount of welfare possible, but this is not a necessary feature of the overall view. Instead, the hybrid theorist can claim that there is a prudential asymmetry between the good and the bad, so while the organic unity gives a total value greater than the value of summing the individual elements, the negative elements on their own are greater than the positive elements on their own. This approach is defensible based on other arguments I have given, such as the existence of a prudential asymmetry between pleasure as pleasure and pain as pain.

I described the intuitive asymmetry in Chapter 1, which holds that it is easier to be badly off than it is to be well off. Although not a condition as such, I said that we should expect this asymmetry to show up in theories of ill-being, which is what we find with hybrids. Possessing an objective bad decreases

our welfare whether or not we direct *any* attitude toward it, and the contribution of the bad seems more independent of any attitude than is the case with objective goods. Anti-justified false belief is bad on its own, and while we can make our welfare even worse by taking pleasure in it, there does not seem to be any special connection that results from the combination of these two elements that we cannot explain through mere addition. If this is true, permissivism is unpersuasive as a theory of ill-being.

In a way, this result should not be surprising. The hybrid theory says that it is more difficult to reach full well-being than pure theories do, because on the hybrid two elements must exist and interact in the correct way. When they do, there is an organic unity. But the intuitive asymmetry I described in Chapter 1 holds that it is easier to be badly off than it is to be well off, so we should expect a theory that makes it harder than other theories to be badly off to produce implausible results, as is the case with restrictivism, or to be an otherwise tough sell, as is the case with permissivism. Despite their attractiveness as theories of well-being, hybrid theories fail as theories of ill-being.

Let us take stock. I have shown that some theories produce structural asymmetries between the good and the bad. Sometimes these make it easier to be badly off, such as hybrids and adjusted subjective theories. In other cases, particularly pain, I argued that there is a prudential asymmetry between pleasure and pain, such that, in equal intensities, pain is a worse bad than pleasure is a good. Structurally, other pairs come out neutral. For instance, there is no reason to think that aversions are easier to satisfy than desires, though it depends on the agent. Other pairs are what we might call *good-favouring*, for they make it easier to be well off, sometimes by denying that there is a negative counterpart to the good. Autonomy most clearly has this form. I argued that a complete lack of autonomy is prudentially neutral, not negative. Although I did not develop a complete ordering, knowledge might also have this form. Anti-justified false belief is bad, though not clearly worse than knowledge is good, and perhaps even better.

Nearly every theory I surveyed has some form of asymmetry. The only one that clearly breaks this pattern is the desire theory and its negative counterpart, aversions. All the others are asymmetrical in some form, either structurally or prudentially. For example, while achievement and bungling might have the same weight prudentially, I showed that there are conditions for anti-achievement that do not apply to achievement, including how far one is away from the goal and the global competence of the agent. Similarly, adjusted subjective theories are attractive as theories of welfare, but they fail as theories of illfare because adjustments for ill-being are implausible. Hybrids are the same: while an organic unity is a plausible feature of welfare, it is far less plausible for illfare.

Of course, it would be nicer if this investigation revealed a pattern that applies to all prudential theories so that we can see how ill-being follows from well-being every time. But as I have shown throughout this dissertation, we have no reason to assume that the landscape of prudential value is so simple. Rather than assuming that ill-being trivially follows from well-being, we have to take a closer look.

# Bibliography

- Acton, Henry Burrows. "Negative Utilitarianism," with John William Nevill Watkins, *Aristotelian Society Supplementary Volume* 37:1 (1963): 83–114.
- Allemang, John. "Philosopher Thomas Hurka's Top 10 of Happiness," *The Globe and Mail* January 21, 2011. <https://www.theglobeandmail.com/life/philosopher-thomas-hurkas-top-10-of-happiness/article562929/>
- Augustine. *The City of God Against the Pagans* (trans. R.W. Dyson). Cambridge: Cambridge University Press (1998).
- Aquinas, Thomas. *Summa Theologiae*. Rochester, NY: The Aquinas Institute, 2012 (1485).
- Benatar, David. *Better Never to Have Been*. Oxford: Oxford University Press, 2006.
- Benatar, David and David Wasserman. *Debating Procreation: Is It Wrong to Reproduce?* Oxford: Oxford University Press, 2015.
- Benatar, David. "Kids? Just Say No," Aeon October 19 (2017): <https://aeon.co/essays/having-children-is-not-life-affirming-its-immoral>
- Bradford, Gwen. *Achievement*. Oxford: Oxford University Press, 2015.
- Brandt, Richard. *A Theory of the Good and the Right*. Amherst, New York: Prometheus Books, 1979.
- Brueckner, Anthony and John Martin Fischer. "Why Is Death Bad?" *Philosophical Studies* 50:2 (1986): 213–221.
- Broad, C.D. *Five Types of Ethical Theory*. London: Routledge & Kegan Paul, 1930.
- Bykvist, Krister. *Utilitarianism: A Guide for the Perplexed*. New York: Continuum, 2009.
- Crisp, Roger. *Reasons and the Good*. Oxford: Clarendon Press, 2006.
- Feldman, Fred. *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford: Clarendon Press, 2004.
- Feldman, Fred. *What Is This Thing Called Happiness?* Oxford: Oxford University Press, 2010.
- Fletcher, Guy. "A Fresh Start for the Objective-List Theory of Well-Being," *Utilitas* 25:2 (2013): 206–220.
- Freedman, Benjamin. "A Moral Theory of Informed Consent," *The Hastings Center Report* 5:4 (1975): 32–39.
- Galbraith, John Kenneth. *The Affluent Society, Fortieth Anniversary Edition*. New York: Houghton Mifflin Company, 1998.
- Greco, John. *Achieving Knowledge*. Cambridge: Cambridge University Press, 2010.
- Griffin, James. *Well-Being: Its Meaning, Measurement and Moral Importance*. Oxford: Oxford University Press, 1989.
- Haji, Ishtiyaque. *Freedom and Value: Freedom's Influence on Welfare and Worldly Value*. New York: Springer, 2009.

- Hanser, Matthew. "The Metaphysics of Harm," *Philosophy and Phenomenological Research* 77:2 (2008): 421–450.
- Harris, John. "Euthanasia and the Value of Life," in *Euthanasia Examined* (ed. John Keown). Cambridge, UK: Cambridge University Press, 1995.
- Hardin, Russell. *Morality within the Limits of Reason*. Chicago: University of Chicago Press, 1988.
- Hart, Herbert Lionel Adolphus. "Death and Utility," *The New York Review of Books* May 15, 1980.
- Hawkins, Jennifer. "The Experience Machine and the Experience Requirement," in *The Routledge Handbook of Philosophy of Well-Being* (ed. Guy Fletcher). New York: Routledge, 2015.
- Heathwood, Chris. "Desire Satisfactionism and Hedonism," *Philosophical Studies* 128 (2006): 539–563.
- Hurka, Thomas. "Why Value Autonomy?" *Social Theory and Practice* 13:3 (1987): 361–382.
- Hurka, Thomas. *Perfectionism*. Oxford: Oxford University Press, 1993.
- Hurka, Thomas. *Virtue, Vice, and Value*. Oxford: Oxford University Press, 2001.
- Hurka, Thomas. "Asymmetries In Value," *Noûs* 44:2 (2010): 199–223.
- Hurka, Thomas. *The Best Things in Life*. Oxford: Oxford University Press, 2011.
- Hurka, Thomas. *British Ethical Theorists from Sidgwick to Ewing*. Oxford: Oxford University Press, 2014.
- Hurka, Thomas. "APA Paper on Kraut," unpublished manuscript (a).
- Hurka, Thomas. "The Intrinsic Goods of Knowledge and Achievement," unpublished manuscript (b).
- Kagan, Shelly. "Me and My Life," *Proceedings of the Aristotelian Society* 94 (1994): 309–324.
- Kagan, Shelly. "Well-Being as Enjoying the Good," *Philosophical Perspectives* 23 (2009): 253–272.
- Kagan, Shelly. "An Introduction to Ill-Being," *Oxford Studies in Normative Ethics* 4 (2014): 261–288.
- Kagan, Shelly. "What is the Opposite of Well-Being?" unpublished manuscript.
- Kraut, Richard. "Desire and the Human Good," *Proceedings and Addresses of the American Philosophical Association* 68:2 (1994): 39–54.
- Kraut, Richard. *What Is Good and Why*. Cambridge, MA: Harvard University Press, 2007.
- Kudrna, Laura, Georgios Kavetsos, Chloe Foy, and Paul Dolan. "Without My Medal on My Mind: Counterfactual Thinking and Other Determinants of Athlete Emotions," *Center for Economic Performance Discussion Paper No 1436* (2016): 1–33.
- Kymlicka, Will. *Liberalism, Community, and Culture*. Oxford: Oxford University Press, 1989.
- Kymlicka, Will. *Contemporary Political Philosophy: An Introduction*, 2nd edition. Oxford: Oxford University Press, 2002.
- Lazari-Radek, Katarzyna and Peter Singer. *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford: Oxford University Press, 2014.
- Leibniz, Gottfried Wilhelm. *Theodicy* (ed. Austin Farrer, trans. E.M. Huggard). New Haven: Yale University Press, 1952 (1710).
- Marino, Patricia. "Ambivalence, Valuational Inconsistency, and the Divided Self," *Philosophy and Phenomenological Research* 83:1 (2011): 41–71.
- Matsumoto, David, and Bob Willingham. "The Thrill of Victory and the Agony of Defeat: Spontaneous Expressions of Medal Winners of the 2004 Athens Olympic Games," *Journal of Personality and Social Psychology*, 91:3 (2006): 568–581.
- Mayerfeld, Jamie. "The Moral Asymmetry of Happiness and Suffering," *The Southern Journal of Philosophy* vol. XXXIV (1996): 317–338.
- Mayerfeld, Jamie. *Suffering and Moral Responsibility*. Oxford: Oxford University Press, 1999.

- McDaniel, Kris and Ben Bradley. “Desires,” *Mind* 117:466 (2008): 267–302.
- McKerlie, Dennis. “Dimensions of Equality,” *Utilitas* 13:3 (2001): 263–288.
- Medvec, Victoria, Scott Madey, and Thomas Gilovich. “When Less is More: Counterfactual Thinking and Satisfaction Among Olympic Medalists,” *Journal of Personality and Social Psychology* 69:4 (1995): 603–610.
- Mill, John Stuart. *On Liberty* (ed. David Spitz). New York: Norton, 1975 (1859).
- Mooney, Michael. “The Most Amazing Bowling Story Ever,” *D Magazine* July (2012).
- Moore, G.E. *Principia Ethica*. Mineola, NY: Dover Publications Inc., 2004 (1903a).
- Moore, G.E. “Mr. McTaggart’s Ethics,” *International Journal of Ethics* 13 (1903b): 341–370.
- Nord, Eric, Jose-Louis Pinto Prades, Jeff Richardson, Paul Menzel, and Peter Ubel. “Incorporating Societal Concerns for Fairness in Numerical Valuations of Health Programmes,” *Health Economics* (1999): 25–39.
- Otsuka, Michael and Alex Voorhoeve. “Why It Matters that Some are Worse Off Than Others: An Argument Against the Priority View,” *Philosophy and Public Affairs* 37:2 (2009): 171–199.
- Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.
- Parfit, Derek. “Equality and Priority,” *Ratio* 10:3 (1997): 202–221.
- Popper, Karl. *The Open Society and Its Enemies, vol. I*, 5th edition. Princeton, NJ: Princeton University Press, 1971.
- Raz, Joseph. *The Morality of Freedom*. Oxford: Clarendon Press, 1986.
- Ross, W.D. (Sir David). *Foundations of Ethics: The Gifford Lectures Delivered in the University of Aberdeen 1935–6*. Oxford: Clarendon Press, 1939.
- Sarch, Alexander F. “Multi-Component Theories of Well-Being and Their Structure,” *Pacific Philosophical Quarterly* 93:4 (2012): 439–471.
- Scheffler, Samuel. *Death and the Afterlife*. Oxford: Oxford University Press, 2014.
- Schroeder, Timothy. *Three Faces of Desire*. New York: Oxford University Press, 2004.
- Sen, Amartya. *On Ethics and Economics*. Oxford: Basil Blackwell, 1987.
- Sidgwick, Henry. *The Methods of Ethics*, 7th edition. London: Macmillan & Co. Ltd., 1972 (1907).
- Sinhababu, Neil. “The Humean Theory of Motivation Reformulated and Defended,” *The Philosophical Review* 118:4 (2009): 465–500.
- Singer, Peter. *Practical Ethics*, 1st edition. Cambridge: Cambridge University Press, 1979.
- Singer, Peter. *Practical Ethics*, 2nd edition. Cambridge: Cambridge University Press, 1993.
- Singer, Peter. “Right to Life?” *New York Review of Books* May 15, 1980.
- Smart, John Jamieson Carswell. *An Outline of a System of Utilitarian Ethics*. Melbourne: Melbourne University Press on behalf of the University of Adelaide, 1961.
- Smart, Roderick Ninian. “Negative Utilitarianism,” *Mind* 67:268 (1958): 542–543.
- Sumner, L.W. *Welfare, Happiness, and Ethics*. Oxford: Oxford University Press, 1996.
- Sumner, L.W. “The Worst Things in Life,” unpublished manuscript.
- Tait, Rosemary and Roxane Cohen Silver. “Coming to Terms with Major Negative Life Events,” in *Unintended Thought* (eds. James Uleman and John Bargh). New York: Guilford Press, 1989.
- Totterdell, Peter. “Catching Moods and Hitting Runs: Mood Linkage and Subjective Performance in Professional Sport Teams,” *Journal of Applied Psychology* 85 (2000): 848–859.
- von Kriegstein, Hasko. *Shaping the World in One’s Image: An Essay on the Nature and Value of Achievements*. Toronto: University of Toronto Doctoral Thesis, 2014.

- Walker, A.D.M. "Negative Utilitarianism," *Mind* 83:331 (1974): 424–428.
- Wall, Steven. *Liberalism, Perfection and Restraint*. Cambridge: Cambridge University Press, 1998.
- Wolf, Susan. "Happiness and Meaning: Two Aspects of the Good Life," *Social Philosophy and Policy* 14 (1997): 207–225.
- Woodard, Christopher. "Hybrid Theories," in *The Routledge Handbook of Philosophy of Well-Being* (ed. Guy Fletcher). New York: Routledge, 2015.