Hedonism and Happiness in Theory and Practice

by

Daniel Michael Weijers

A thesis submitted to Victoria University of Wellington in fulfilment of the
requirements for the degree of Doctor of Philosophy

Victoria University of Wellington

2012

—for Linda, Eli, and Alex—

Contents

Acknowledgments	7
Abstract	9
Part 1	11
Chapter 1: Hedonism: An Encyclopaedic Overview	15
Chapter 2: The Experience Machine Objection to Hedonism	49
Chapter 3: The Case for Status Quo Bias in Experience Machine Scenarios	83
Chapter 4: We Can Test the Experience Machine	117
Chapter 5: Reducing Status Quo Bias and other Confounding Factors in Experience Machine Scenarios	135
Part 2	161
Chapter 6: Theories of the Meaning of Life: Optimistic Naturalism and Infinite Happiness	165
Chapter 7: Happiness and Climate Change: How Should the Responsibilities of Adaptation and Mitigation be Shared?	199
Conclusion	227
Postscript for Policymakers: The Science of Happiness for Policy- makers	231
References	263

Acknowledgements

I have many people to thank for supporting, encouraging, and inspiring me throughout the last three years. I have thoroughly enjoyed writing this thesis in large part because of them. I take great pleasure in being able to pay tribute to them here.

First and foremost, I wish to thank my supervisor, Nick Agar, for being my philosophical mentor and so much more. In addition to continuous useful feedback on my writing and ideas, Nick has generously offered me guidance, reassurance, and friendship. With Nick's help, I have become a better philosopher in all respects.

Ramon Das and Simon Keller have also helped hone my philosophical skills with invaluable advice on manuscripts for my thesis and other projects.

I'm also indebted to the entire Philosophy programme at Victoria University of Wellington, and particularly to the staff who have gone out of their way to provide me with useful career opportunities and advice: Nick Agar, Simon Keller, Ken Perszyk, Ramon Das, Kim Sterelny, Stuart Brock, and Ed Mares. And to Pennie Gapes for being so knowledgeable and reliable in her role as school manager.

I'm also grateful to the Victoria University of Wellington students that I have studied and socialised with. A great deal of my philosophical learning is thanks to engagement with these wonderful people in corridors, in seminars, and over beers. Most of all, I'd like to thank: Ewan Kingston, John Matthewson, David Gilbert, Chiara Elettra, Hatha McDivitt, Frieder Lempp, Tony Scott, Lucas Kengmana, Steve Riley, and Shri Sharma.

I have been very lucky to collaborate with several outstanding people while writing academic papers, setting up the *International Journal of Wellbeing*, organising a conference on Well-being and Public Policy, and

organising an edited book on well-being and public policy: Aaron Jarden, Philip S. Morrison, Ramon Das, David Eng, John F. Helliwell, and Sam Wren-Lewis. I have thoroughly enjoyed working with you all.

Certainly worth a mention; this thesis would not have been possible without the scholarship and grant funding provided by the Faculty of Humanities and Social Sciences at Victoria University of Wellington.

I also want to take this opportunity to mention four thinkers whose writing inspired me to investigate hedonism and happiness in more detail: Dan Haybron, Roger Crisp, Richard Layard, and Fred Feldman. If I had not read their work several years ago, I have no idea where I would be now.

To my parents, Anne and Mick, I want to thank you for raising me and helping me become the person I am today. I now realise that raising a child is a difficult and often underappreciated task, so please accept this heartfelt thank you for all the effort you have both put in over the last 30 years. I couldn't have had a better start in life.

To my brother Joe, aunt Beth, and father Mick, thanks for your useful comments on my often outlandish ideas.

Finally, thanks and more thanks to Linda, Eli, and Bud. This thesis is for you as much as it is for me—it's for our family. Thanks for supporting me over these three years and thanks for making me whole... You know I'd only get into the experience machine if you guys did too.

Abstract

In this thesis, I investigate several different questions about happiness and hedonism in theory and practice and offer several arguments and theories. In addition to making progress in these happiness-related areas of inquiry, this thesis aims to demonstrate the complexity and variety of happiness-related problems and the broad range of real-world problems that considerations of happiness can help to resolve. Furthermore, nearly every chapter of this thesis demonstrates how interdisciplinary analyses can bring new movement to problems that have become insulated within one academic discipline.

This thesis is divided into two main parts. Chapters 1 through 5 constitute Part 1, and Chapters 6 through 8 constitute Part 2.

Part 1 of this thesis is focused on theory and questions about *what we should believe*. In particular, Part 1 is concerned with Prudential Hedonism, a theory of what is good for a person, which claims (roughly) that a preponderance of pleasure over pain (sometimes referred to as happiness) is what is ultimately good for people.

After providing a broad overview of hedonism, and especially Prudential Hedonism, in Chapter 1, the remainder of Part 1 focuses on one main question from philosophical debates about well-being: does the experience machine thought experiment give us good reason to believe that internalist accounts of Prudential Hedonism are all false? The main conclusion that I argue for in Part 1 is that no, the experience machine thought experiment does not give us good reason to believe that internalist accounts of Prudential Hedonism are all false.

Part 2 of this thesis is focused on practice, and particularly on how considerations of happiness can inform certain practices and help us to understand *what we should do* in certain circumstances. Unlike Part 1, which

has a smooth narrative flow from chapter to chapter, Part 2 contains three relatively unrelated chapters, each of which investigates a different question without relying on the conclusions of any previous chapters.

Chapter 6 argues that an optimistic view about scientific and technological progress allows for two interesting new theories for the meaning of life debate, and discusses what people with certain kinds of belief might want to do to achieve true meaning in life. One of these theories posits that causing there to be infinite happiness can be a way to achieve a truly meaningful life.

Chapter 7 demonstrates how considerations of human happiness can justify why a particular set of distributive principles are the fairest way to apportion the burdens associated with adapting to, and mitigating, the potentially devastating effects of rapid climactic change. Based on these considerations, Chapter 7 includes fairly specific policy recommendations about what governments should do about climate change.

This thesis also includes a Postscript for Policymakers. Compared to Chapters 2 to 7, the Postscript for Policymakers takes a much higher-level approach; it seeks to provide general answers to two very broad questions. Given its broader scope and different intended audience, the Postscript for Policymakers does not include in-depth discussion of all likely objections. The two questions addressed in the Postscript for Policymakers are: should policymakers use findings from the science of happiness to guide their policy decisions, and how can they best do this? The Postscript for Policymakers concludes that findings from the science of happiness should be used to guide policymaking (with several qualifications), and it provides recommendations for how best to do this.

Part 1

Part 1 of this thesis focuses on theory and questions about *what we should believe*.

In philosophy, the two main theoretical debates about happiness are how to define happiness and whether being happy is all, part, or none of what makes our lives go well for us. Part 1 of this thesis engages in this second debate. The claim that being happy is all of what ultimately makes our lives go well for us is most often defended by Prudential Hedonists. Prudential Hedonists believe that happiness—usually defined as a preponderance of pleasure over pain, and where both pleasure and pain are defined broadly—is all of what ultimately makes our lives go well for us.

Philosophical, and especially popular, writing about happiness is full of ambiguity about what the term hedonism actually means. This ambiguity is perpetuated by conflicting descriptions in philosophical dictionaries and introductory textbooks. Simon Blackburn (2005, p. 161) and Gregory Pence (2005, p. 25) both severely limit hedonism by describing it as a specifically normative theory—which claims that pursuing our own pleasure should be the aim of all our actions—without mentioning any other types of hedonism. This normative account of hedonism is better known as Hedonistic Egoism and stands opposed to Hedonistic Utilitarianism (Moore 2011), which states that the morally good action is the one that is likely to maximise net pleasure for everyone (Brink 2006, p. 381). In a similarly misleading way, Vesey and Foulkes (1990) take the opposite approach and only discuss Hedonistic Utilitarianism and Motivational Hedonism. These omissions are unfortunate because it is clear that all of the main versions of normative hedonism rely on Prudential Hedonism as a justification. Such narrow descriptions of hedonism could lead an uninitiated reader to assume that hedonism is just a normative

theory, which is simply not the case. Since the precise meanings of the terms hedonism and, particularly, Prudential Hedonism are required to properly engage in any debate about whether being happy is all or only part of what makes our lives go well for us and since ambiguity about what these terms mean is so rife, Chapter 1 of this thesis is an encyclopaedic overview of these terms as they are used in philosophy.

Mentioned in the encyclopaedic overview of Chapter 1, is the experience machine objection to hedonism—the most prominent of all objections to Prudential Hedonism. The popularity of this particular objection, along with the widespread conception that it is successful enough to single-handedly render Prudential Hedonism implausible, motivates the focus of the remainder of Part 1.

Chapter 2 explains the experience machine objection to hedonism, and includes two formal versions of the objection to capture both its common use and its best use in the literature. All of the article-length attempts to refute the experience machine objection to hedonism from the last 15 years are then reviewed. Most of the articles use several different reasons to argue that the experience machine objection to hedonism is not convincing, so these reasons are grouped according to the premise of the argument that they attempt to refute. Why most of the refutations are misguided, or unlikely to convince those who believe that the experience machine objection to hedonism is successful, is then explained. The chapter concludes by identifying one potentially fruitful refutation of the experience machine objection to hedonism.

Chapter 3 takes on the task of developing this refutation. The attempted refutation rests heavily on how philosophical judgements about thought experiments are created, and the extent to which they can be affected by biases because of this process. So, much of this chapter is spent discussing

how our intuitive cognition works and how it can affect our judgements about thought experiments. A few philosophers have already argued that the status quo bias—an inappropriate preference for things to remain the same—is to blame for the perceived success of the experience machine objection to hedonism. Their arguments are analysed in detail in this chapter. Two of these philosophers, Felipe De Brigard (2010) and Basil Smith (2011), have even conducted experiments on this question and their methods and results are also discussed. Chapter 3 concludes that more empirical experiments are required for us to have good reason to believe that the status quo bias is really the main cause of the success of the experience machine objection to hedonism.

Chapter 4 responds to Basil Smith's (2011) charge that empirical surveys about experience machine thought experiments cannot actually tell us anything useful. Chapter 4 concludes that, while experiments on experience machine thought experiments can be set up in ways that make them more or less useful, there is no good reason to believe that well-constructed experiments on experience machine scenarios are useless.

Building on the insights from Chapter 4, Chapter 5 discusses several experiments on a few novel experience machine scenarios. The results of the experiments indicate that status quo bias and other confounding factors are very likely to be important causes of the apparent success of the experience machine objection to hedonism. The results also demonstrate that the experience machine objection to hedonism is not effective if a relatively bias-free version of the experience machine thought experiment is used. Chapter 5 concludes with some guidance for how introductory ethics classes should be taught in light of the findings in Part 1.

Taken together, the chapters in Part 1 constitute a background-rich and highly-detailed discussion of what we should believe about one particular

argument against a specific group of theories—the experience machine objection to hedonism.

Chapter 1

Hedonism: An Encyclopaedic Overview

Chapter Summary

This chapter provides a broad overview of the term ‘hedonism’ in general and of how philosophers use the term ‘Prudential Hedonism’ in particular. After Prudential Hedonism is disambiguated from various other hedonisms, its history, current major varieties, and main criticisms are discussed. This chapter does not attempt to argue for or against any conceptions of hedonism or accounts of Prudential Hedonism. Rather, this chapter intends to reflect the current shared knowledge about hedonism and Prudential Hedonism to provide context for the remainder of Part 1.

1. Introduction¹

The term ‘hedonism’, from the Greek ἡδονή (*hēdonē*) for pleasure, refers to several related theories about what is good for us, how we should behave, and what motivates us to behave in the ways that we do. All hedonistic theories identify pleasure and pain as the only important elements of whatever phenomena they are designed to describe. If hedonistic theories identified pleasure and pain as merely two important elements, instead of the *only* important elements of what they are describing, then they would not be nearly as unpopular as they all are. However, the claim that pleasure and pain are the *only* things of ultimate importance is what makes hedonism distinctive and philosophically interesting.

¹ This chapter provided the basis for Weijers (2011a).

Philosophical Hedonists tend to focus on hedonistic theories of value, and especially of well-being—the good life for the one living it. As a theory of value, hedonism states that all and only pleasure is intrinsically valuable and all and only pain is intrinsically dis-valuable. Hedonists usually define pleasure and pain broadly, such that both the pleasure of reading a good book and the pain of reading a bad thesis are included. Thus, a gentle massage and recalling a fond memory are both considered to cause pleasure and stubbing a toe and hearing about the death of a loved one are both considered to cause pain. With pleasure and pain so defined, hedonism is intuitively appealing as a theory about what is valuable for us. Indeed, its appeal is evidenced by the fact that nearly all historical and contemporary treatments of well-being allocate at least some space to discuss hedonism. Unfortunately for hedonism, these discussions rarely endorse it and some even deplore its focus on pleasure.

This chapter begins by clarifying the different types of hedonistic theories and the labels they are often given. Then, hedonism's ancient origins and its subsequent development are reviewed. The majority of this chapter is then concerned with describing the important theoretical divisions within Prudential Hedonism and discussing the major criticisms of these divisions.

2. Disambiguation

2.1 Folk Hedonism

When the term 'hedonism' is used in modern literature, or by non-philosophers in their everyday talk, its meaning is quite different from the one it takes when used by philosophers. Non-philosophers tend to think of a hedonist as a person who seeks out pleasure for themselves without any particular regard for their own future well-being, or for the well-being of others. According to non-philosophers, then, a stereotypical hedonist is

someone who never misses an opportunity to indulge of the pleasures of sex, drugs, and rock 'n' roll, even if the indulgences are likely to lead to relationship problems, health problems, regrets, or sadness for themselves or for others. Philosophers commonly refer to this everyday understanding of hedonism as 'Folk Hedonism'. Folk Hedonism is a rough combination of Motivational Hedonism, Hedonistic Egoism, and a reckless lack of foresight.

2.2 Value Hedonism and Prudential Hedonism

When philosophers discuss hedonism, they are most likely to be referring to hedonism about value, and especially the slightly more specific theory, hedonism about well-being. Hedonism as a theory about value (best referred to as Value Hedonism) holds that all and only pleasure is intrinsically valuable and all and only pain is intrinsically dis-valuable, or more simply, "pleasure is the only prudential good and pain is the only prudential bad" (Moore & Crisp 1996, p. 599). The term 'intrinsically' is an important part of the definition and is best understood in contrast to the term 'instrumentally'. Something is intrinsically valuable if it is valuable for its own sake. Pleasure is thought to be intrinsically valuable because, even if it did not lead to any other benefit, it would still be good to experience. Money is an example of an instrumental good; its value for us comes from what we can do with it (what we can buy with it). The fact that a copious amount of money has no value if no one ever sells anything reveals that money lacks intrinsic value. Value Hedonism reduces everything of value to pleasure. For example, a Value Hedonist would explain the instrumental value of money by describing how the things we can buy with money, such as food, shelter, and status-signifying goods, bring us pleasure or help us to avoid pain.

Hedonism as a theory about well-being (best referred to as Prudential Hedonism) is more specific than Value Hedonism because it stipulates what

the value is *for*. Prudential Hedonism holds that all and only pleasure intrinsically makes people's lives go better *for them* and that all and only pain intrinsically makes their lives go worse *for them*. Some philosophers replace 'people' with 'persons and sentient nonpersons', so as to apply Prudential Hedonism more widely. A good example of this comes from Peter Singer's (1990) work on animals and ethics. Singer questions why some humans can see the intrinsic disvalue in human pain, but do not also accept that it is bad for sentient non-human animals to experience pain.

When Prudential Hedonists claim that happiness is what they value most, they intend happiness to be understood as a preponderance of pleasure over pain.² An important distinction between Prudential Hedonists and Folk Hedonists is that Prudential Hedonists usually understand that pursuing pleasure and avoiding pain in the very short-term is not always the best strategy for achieving the optimal long-term balance of pleasure over pain.

Prudential Hedonism is an integral part of several derivative types of hedonistic theory, all of which have featured prominently in philosophical debates of the past. Prudential Hedonism is also an important theory in the debate about what well-being consists of, a debate which is generally seen as epistemically prior to many other important debates in moral philosophy, including how we should live.³ Since a proper understanding of Prudential Hedonism is required to understand most other types of hedonism, and several other important philosophical debates, the majority of this chapter is

² Most philosophers allow this understanding of happiness as one of many possible conceptions, but Dan Haybron (2001) has argued against it.

³ For an interesting discussion of why an understanding of what well-being consists in needn't be epistemically prior to the question of how we should live, see Tiberius (2008, part 1).

dedicated to discussing Prudential Hedonism. First, however, the main derivative types of hedonism are briefly discussed.

2.3 *Motivational Hedonism*

Motivational Hedonism (more commonly referred to by the less descriptive label, Psychological Hedonism) is the theory that the desire to encounter pleasure and avoid pain guides all of our behaviour. Most philosophical accounts of Motivational Hedonism include both conscious and unconscious desires for pleasure, but emphasize the latter.⁴ Epicurus,⁵ Jeremy Bentham (1789), and John Stuart Mill (1861) have all argued for varieties of Motivational Hedonism. Bentham used the idea to support his theory of Hedonistic Utilitarianism (discussed below). Weak versions of Motivational Hedonism hold that the desire to seek pleasure and avoid pain often or always has *some* influence on our behaviour. Weak versions are generally considered to be uncontroversially true and not especially useful for philosophy.

Philosophers have been more interested in strong accounts of Motivational Hedonism, which hold that *all* behaviour is governed by the desire to encounter pleasure and avoid pain, and *only* this desire. Strong accounts of Motivational Hedonism have been used to support some of the normative types of hedonism and to argue against non-hedonistic normative theories.

⁴ In contrast, psychologists and management theorists have often understood hedonism to include the volitional pursuit of please and avoidance of pain *but not* the unconscious pursuit of pleasure and avoidance of pain (which they consider to be an integral part of 'instinct theory'). See for example James (1950, vol. 1, pp. 143–4; vol. 2, pp. 549–59), Murray (1964), Porter, Bigley, and Steers (2003), and Vroom (1964).

⁵ See for example Diogenes (1925a) and Mitsis (1988).

Plato's Ring of Gyges example, in *The Republic* (Plato 1974), can be read as a notable objection to Motivational Hedonism. Plato's Socrates is discussing with Glaucon how men would react if they were to possess a ring that gave its wearer immense powers, including invisibility. Glaucon believes that a strong version of Motivational Egoism is true, but Socrates does not. Motivational Egoism is the theory that all the actions of each person are guided solely by the desire to improve one's own life. Since Motivational Hedonism is a sub-type of Motivational Egoism, it is also brought into doubt by successful objections to Motivational Egoism. Glaucon asserts that, emboldened with the power provided by the Ring of Gyges, everyone would succumb to the inherent and ubiquitous desire to pursue their own ends at the expense of others. Socrates disagrees, arguing that good people would be able to overcome this desire because of their strong love of justice, fostered through philosophising.

Strong accounts of Motivational Hedonism currently garner very little support for similar reasons. There are many examples of seemingly pain-seeking acts performed out of a sense of duty—from the soldier who jumps on a grenade to save his comrades to that time you rescued a trapped dog only to be (predictably) bitten in the process. Introspective evidence also weighs against strong accounts of Motivational Hedonism; many of the decisions we make seem to be based on motives other than seeking pleasure and avoiding pain. Given these reasons, most philosophers consider the burden of proof to be squarely on the shoulders of anyone wishing to argue for a strong account of Motivational Hedonism.⁶

⁶ Moore (2011) offers further challenges for motivational hedonism.

2.4 Normative Hedonism

Value Hedonism, occasionally with assistance from Motivational Hedonism, has been used to argue for specific theories of right action (theories that explain which actions are morally permissible or impermissible and why). The theory that happiness should be pursued (that pleasure should be pursued and pain should be avoided) is referred to as Normative Hedonism and sometimes Ethical Hedonism. There are two major types of Normative Hedonism: Hedonistic Egoism and Hedonistic Utilitarianism. Both types commonly use happiness (defined as pleasure minus pain) as the sole criterion for determining the moral rightness or wrongness of an action. Important variations within each of these two main types specify either the actual resulting happiness (after the act) or the predicted resulting happiness (before the act) as the moral criterion. Although both major types of Normative Hedonism have been accused of being repugnant, Hedonistic Egoism is usually considered the most offensive.

2.5 Hedonistic Egoism

Hedonistic Egoism is the theory that we should, morally speaking, do whatever is most in our own best interests. Because it is a form of hedonism, Hedonistic Egoism holds that what is most in our own best interests is whatever makes us happiest – that is whatever provides us with the most net pleasure after pain is subtracted. The most repugnant feature of this theory is that one never has to ascribe any value whatsoever to the consequences for anyone other than oneself. For example, a Hedonistic Egoist who did not feel saddened by theft would be morally required to steal, even from needy orphans, if he thought he could get away with it. Would-be defenders of Hedonistic Egoism might point out that performing acts of theft, murder, treachery and the like would not make them happier overall because of the

guilt, the fear of being caught, and the chance of being caught and punished. The would-be defenders should surrender, however, if it is pointed out that a Hedonistic Egoist is morally obliged by their own theory to pursue an unusual kind of practical education—a brief and possibly painful training period that reduces their moral emotions of sympathy and guilt. Such an education might be achieved by desensitising over-exposure to, and performance of, torture of innocents. If Hedonistic Egoists underwent such an education, their reduced capacity for sympathy and guilt would allow them to take advantage of any opportunities to perform pleasurable, but normally-guilt-inducing, actions, such as stealing from the poor.

Hedonistic Egoism is very unpopular amongst philosophers, not just for this reason, but also because it suffers from all of the objections that apply to Prudential Hedonism (discussed below).

2.6 Hedonistic Utilitarianism

Hedonistic Utilitarianism is the theory that the right action is the one that produces (or is most likely to produce) the greatest net happiness for all concerned. Hedonistic Utilitarianism is often considered fairer than Hedonistic Egoism because the happiness of everyone involved (everyone who is affected or is likely to be affected) is taken into account and given equal weight. Hedonistic Utilitarians, then, would advocate *not* stealing from needy orphans because to do so would usually leave the orphan far less happy and the (probably better-off) thief only slightly happier (assuming he felt no guilt). Despite treating all individuals equally, Hedonistic Utilitarianism is still seen as objectionable by some because it assigns no intrinsic moral value to justice, friendship, truth, or any of the many other goods that are often thought to be irreducibly valuable. For example, a Hedonistic Utilitarian would be morally obliged to publicly execute an

innocent friend of theirs if doing so was the only way to promote the greatest happiness overall. Although unlikely, such a situation might arise if a child was murdered in a small town and the lack of suspects was causing large-scale inter-ethnic violence. Some philosophers would argue that executing an innocent friend is immoral precisely because it ignores the intrinsic values of justice, friendship, and possibly truth.

Hedonistic Utilitarianism is rarely endorsed by philosophers, but mainly because of its reliance on Prudential Hedonism as opposed to its utilitarian element. Non-hedonistic versions of utilitarianism are about as popular as the other leading theories of right action, especially when it is the actions of institutions that are being considered.

3. The Origins of Hedonism

3.1 Cārvāka

Perhaps the earliest written record of hedonism comes from the Cārvāka, an Indian philosophical tradition based on the Barhaspatya sutras. The Cārvāka persisted for two thousand years (from about 600 BCE). Most notably, the Cārvāka advocated skepticism and Hedonistic Egoism—that the right action is the one that brings the actor the most net pleasure. The Cārvāka acknowledged that some pain often accompanied, or was later caused by, sensual pleasure, but that the pleasure was worth it.⁷

3.2 Aristippus and the Cyrenaics

The Cyrenaics, founded by Aristippus (c. 435–356 BCE), were also skeptics and Hedonistic Egoists (O’Keefe 2005a). Although the paucity of original texts

⁷ See Chattopadhyaya (1990) and Sarma (2011, pp. 3–13) for more details on the hedonism of the Cārvāka.

makes it difficult to confidently state all of the justifications for the Cyrenaics' positions, their overall stance is clear enough.⁸ The Cyrenaics believed pleasure was the ultimate good and everyone should pursue all immediate pleasures for themselves. They considered bodily pleasures better than mental pleasures, presumably because they were more vivid or trustworthy. The Cyrenaics also recommended pursuing immediate pleasures and avoiding immediate pains with scant or no regard for future consequences. Their reasoning for this is even less clear, but is most plausibly linked to their skeptical views—perhaps that what we can be most sure of in this uncertain existence is our current bodily pleasures.

3.3 *Epicurus*

Epicurus (c. 341–271 BCE), founder of Epicureanism, developed a Normative Hedonism in stark contrast to that of Aristippus. The Epicureanism of Epicurus is also quite the opposite to the common usage of Epicureanism; while we might like to go on a luxurious 'Epicurean' holiday packed with fine dining and moderately excessive wining, Epicurus would warn us that we are only setting ourselves up for future pain. For Epicurus, happiness was the complete absence of bodily and especially mental pains, including fear of the Gods and desires for anything other than the bare necessities of life (Inwood & Gerson 1994; Laertius 1925b; Mitsis 1998). Even with only the limited excesses of ancient Greece on offer, Epicurus advised his followers to avoid towns, and especially marketplaces, in order to limit the resulting desires for unnecessary things. Once we experience unnecessary pleasures, such as those from sex and rich food, we will then suffer from painful and hard to satisfy desires for more and better of the same. No matter how wealthy we might be, Epicurus would argue, our desires will eventually outstrip our means and

⁸ The details presented here are based on Laertius (1925a) and O'Keefe (2002; 2005b).

interfere with our ability to live tranquil, happy lives. Epicureanism is generally egoistic, in that it encourages everyone to pursue happiness for themselves. However, Epicureans would be unlikely to commit any of the selfish acts we might expect from non-epicurean egoists because Epicureans train themselves to desire only the very basics. And, desiring only the very basics gives Epicureans very little reason to do anything to interfere with the affairs of others.

3.4 The Oyster example

With the exception of a brief period discussed below, Hedonism has been generally unpopular ever since its ancient beginnings (Crisp 2006a, pp. 619–620; Feldman 2004, p. 7; Silverstein 2000, p. 279). Although criticisms of the ancient forms of hedonism were many and varied, one in particular was heavily cited. In *Philebus*, Plato's Socrates and one of his many foils, Protarchus in this instance, are discussing the role of pleasure in the good life (Plato 1937, part ii, p. 353). Socrates asks Protarchus to imagine a life without much pleasure but full of the higher cognitive processes, such as knowledge, forethought and consciousness and to compare it with a life that is the opposite. Socrates describes this opposite life as having perfect pleasure but the mental life of an oyster, pointing out that the subject of such a life would not be able to appreciate any of the pleasure within it. The harrowing thought of living the pleasurable but unthinking life of an oyster causes Protarchus to abandon his hedonistic argument. The oyster example can be easily avoided by defining pleasure as being a conscious experience, so any sensation that we are not consciously aware of (such as the oyster's "pleasure") is not pleasure.⁹

⁹ Roger Crisp (2006a, pp. 630–635) suggests stronger versions of the Oyster example and argues that even these stronger versions do not give us reason to reject Hedonism.

4. The Development of Hedonism

4.1 *Bentham*

Normative and Motivational Hedonism were both at their most popular during the heyday of Empiricism in the 18th and 19th Centuries. Indeed, this is the only period during which any kind of hedonism could be considered popular at all. During this period, two Hedonistic Utilitarians, Jeremy Bentham (1748–1832) and his protégé John Stuart Mill (1806–1873), were particularly influential. Their theories are similar in many ways, but provide notably distinct definitions of pleasure.

Bentham (1789) argued for several types of hedonism, including those now referred to as Prudential Hedonism, Hedonistic Utilitarianism, and Motivational Hedonism. He claimed that happiness was the ultimate good and that happiness was pleasure and the absence of pain. He acknowledged the hedonistic and egoistic nature of peoples' motivation, but argued that the maximization of collective happiness was the correct criterion for moral behaviour. Bentham's (1789) greatest happiness principle states that actions are immoral if they do not appear to maximise the happiness of all the people likely to be affected. Only the action that appears to maximise the happiness of all the people likely to be affected is the morally right action.

Bentham devised the greatest happiness principle to justify the legal reforms he had proposed (Sweet 2008). He believed that while he could not conclusively prove that the principle was the correct criterion for morally right action, it should still be accepted because it was both fair and better than existing criteria for evaluating actions and legislation. Bentham (1789) thought that his Hedonic Calculus could be applied to a situation to see what should, morally speaking, be done. Hedonic Calculus is a method of 'counting' the amount of pleasure and pain that would likely be caused by different actions. It required a methodology for measuring pleasure, which in turn required an

understanding of the nature of pleasure and specifically what aspects of pleasure were valuable for us.

Bentham's (1789, pp. 151–154) Hedonic Calculus identifies several aspects of pleasure that should be attended to when working out what the most pleasurable action will be, including certainty, propinquity, extent, intensity, and duration. The Hedonic Calculus also makes use of two future-pleasure-or-pain-related aspects of actions—fecundity and purity. Certainty refers to the likelihood that the pleasure or pain will occur. Propinquity refers to how long away (in terms of time) the pleasure or pain is. Fecundity refers to the likelihood of the pleasure or pain leading to more of the same sensation. Purity refers to the likelihood of the pleasure or pain leading to some of the opposite sensation. Extent refers to the number of people the pleasure or pain is likely to affect. Intensity refers to the felt strength of the pleasure or pain. Finally, duration refers to how long the pleasure or pain are felt for. It should be noted that only intensity and duration have intrinsic value for an individual. Certainty, propinquity, fecundity, and purity are all instrumentally valuable for an individual because they affect the likelihood of an individual feeling future pleasure and pain. Extent is also not directly valuable for an individual's well-being because it refers to the likelihood of other people experiencing pleasure or pain.

Bentham's inclusion of certainty, propinquity, fecundity, and purity in the Hedonic Calculus helps to differentiate his hedonism from Folk Hedonism. Folk Hedonists rarely consider how likely their actions are to lead to future pleasure or pain, focussing instead on the pursuit of immediate pleasure and the avoidance of immediate pain. So while Folk Hedonists would be unlikely to study for an exam, anyone using Bentham's Hedonic Calculus would consider the future happiness benefits to themselves (and possibly others) of passing the exam and then promptly begin studying.

Most importantly for Bentham's Hedonic Calculus, the pleasure from different sources is always measured against these criteria in the same way, that is to say that no additional value is afforded to pleasures from particularly moral, clean, or culturally-sophisticated sources. For example, Bentham (1825, p. 206) held that pleasure from the parlour game push-pin was just as valuable for us as pleasure from music and poetry.¹⁰ Since Bentham's theory of Prudential Hedonism focuses on the quantity of the pleasure, rather than the source-derived quality of it, it is best described as a type of Quantitative Hedonism.

4.2 Mill

Bentham's indifferent stance on the source of pleasures led to others disparaging his hedonism as the philosophy of swine. Even Bentham's student, John Stuart Mill, questioned whether we should believe some of the consequences of his Quantitative Hedonism. Mill questioned whether a satisfied pig leads a better life than a dissatisfied human, or whether a satisfied fool leads a better life than a dissatisfied Socrates (Mill 1861, p. 9).¹¹

Like Bentham, Mill endorsed the varieties of hedonism now referred to as Prudential Hedonism, Hedonistic Utilitarianism, and Motivational Hedonism. Mill also thought happiness, defined as pleasure and the

¹⁰ Bentham is often misquoted on this point, including by John Stuart Mill (1838): "He says, somewhere in his works, that, 'quantity of pleasure being equal, push-pin is as good as poetry'." What Bentham actually said is: "Prejudice apart, the game of push-pin is of equal value with the arts and sciences of music and poetry. If the game of push-pin furnish more pleasure, it is more valuable than either." (Bentham 1825, p. 206).

¹¹ Mill is also often misquoted. He said "It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied. And if the fool, or the pig, are of a different opinion, it is because they only know their own side of the question. The other party to the comparison knows both sides." (Mill 1861, p. 9).

avoidance of pain, was the highest good. Where Mill's hedonism differs from Bentham's is in his understanding of the nature of pleasure. Mill (1861, pp. 8–11) argued that pleasures could vary in quality, being either higher or lower. Mill employed the distinction between higher and lower pleasures in an attempt to avoid the criticism that his hedonism was, like Bentham's, just another philosophy of swine (Crisp 2006a, p. 619). Lower pleasures are those associated with the body, which we share with other animals, such as pleasure from quenching thirst or having sex. Higher pleasures are those associated with the mind, which were thought to be unique to humans, such as pleasure from listening to opera, acting virtuously, or philosophising. Mill (1861, pp. 8–11) justified this distinction by arguing that those who have experienced both types of pleasure realise that higher pleasures are much more valuable. He dismissed challenges to this claim by asserting that those who disagreed lacked either the experience of higher pleasures or the capacity for such experiences. For Mill, higher pleasures were not different to lower pleasures in mere degree; they were different in kind. Since Mill's theory of Prudential Hedonism focuses on the quality of the pleasure, rather than the amount of it, it is best described as a type of Qualitative Hedonism.¹²

4.3 Moore

George Edward Moore (1873–1958) was instrumental in bringing hedonism's brief heyday to an end. Moore's criticisms of hedonism in general, and Mill's hedonism in particular (Moore 1903, chap. 3), were frequently cited as good reasons to reject hedonism during his lifetime (e.g. Wild 1927, p. 11). Indeed, since Moore, hedonism has been viewed by most philosophers as being an initially intuitive and interesting family of theories, but also one that is flawed on closer inspection (Crisp 2006a, pp. 619–620). Moore was a pluralist about

¹² See Smith and Sosa (1969) and Donner (2006) for more on Mill's Hedonistic Utilitarianism.

value and argued persuasively against the Value Hedonists' central claim—that all and only pleasure is the bearer of intrinsic value.¹³ Moore's most damaging objection against Hedonism was his heap of filth example (1903, p. 84). Moore himself thought the heap of filth example thoroughly refuted what he saw as the only potentially viable form of Value Hedonism—that conscious pleasure is the only thing of intrinsic value. Moore used the heap of filth example to argue that Value Hedonism is false because pleasure is not the only thing of value.

In the heap of filth example, Moore (1903, p. 84) asks the reader to imagine two worlds, one of which is exceedingly beautiful and the other a disgusting heap of filth. Moore then instructs the reader to imagine that no one would ever experience either world and asks if it is better for the beautiful world to exist than the filthy one. Moore (1903, p. 84) assumed that his rational contemporaries would agree that it would be better if the beautiful world existed, and very few objections were published at the time (Mettrick 1928, pp. 389, 397). Relying on this assumed agreement, Moore infers that the beautiful world is more valuable than the heap of filth and, therefore, that beauty must be valuable. Moore then concluded that all of the potentially viable theories of Value Hedonism (those that value only conscious pleasures) must be false because something, namely beauty, is valuable even when no conscious pleasure can be derived from it.

Moore's heap of filth example has rarely been used to object to Prudential Hedonism since the 1970's because it is not directly relevant to Prudential Hedonism (it evaluates worlds and not lives). Moore's other objections to Hedonism also went out of favor around the same time. The demise of these arguments was partly due to mounting objections against them,¹⁴ but mainly

¹³ See Shaw (1948, chap. 1) for more detail.

¹⁴ See Blake (1926), Mettrick (1928), and Savery (1934), for example.

because arguments more suited to the task had been developed (Crisp 2006a, p. 620). Of particular note are the arguments against Hedonistic Utilitarianism and especially Prudential Hedonism. These arguments are discussed after the contemporary varieties of hedonism are introduced below.

5. Contemporary Varieties of Hedonism

5.1 The Main Divisions

Several contemporary varieties of hedonism have been defended, although usually by just a handful of philosophers at any one time. Other varieties of hedonism are also theoretically available but have received little or no discussion. Contemporary varieties of Prudential Hedonism are grouped here based on how they define pleasure and pain.¹⁵ In addition to providing different notions of what pleasure and pain are, contemporary varieties of Prudential Hedonism also disagree about which aspect or aspects of pleasure are valuable for well-being (and dis-valuable for pain).

The most well-known disagreement about which aspects of pleasure are valuable occurs between Quantitative and Qualitative Hedonists. Quantitative Hedonists argue that how valuable pleasure is for well-being depends only on the *amount* of pleasure, and so they are only concerned with dimensions of pleasure such as duration and intensity (Frankena 1973, pp. 84–85). Quantitative Hedonism is often accused of over-valuing animalistic, simple, and debauched pleasures.

Qualitative Hedonists argue that, in addition to the dimensions related to the amount of pleasure, one or more dimensions of quality can impact how pleasure affects well-being (Sobel 2002, p. 241). The quality dimensions might

¹⁵ See Feldman (1997) and Katz (2009) for more detail on different definitions of pleasure in philosophy and (to varying extents) the implications for hedonism.

be based on how cognitive or ‘animalistic’ the pleasure is (as it was for Mill), the moral status of the source of the pleasure, or some other non-amount-related dimension. Qualitative Hedonism is criticised by some for smuggling values other than pleasure into well-being by misleadingly labelling them as dimensions of pleasure (Frankena 1973, pp. 84–85). How these qualities are chosen for inclusion can also be criticised for being ad hoc because inclusion of these dimensions of pleasure is often suggested in response to objections that Quantitative Hedonism cannot easily deal with. That is to say, the inclusion of these dimensions could be accused of being an exercise in plastering over holes, rather than deducing corollary conclusions from existing theoretical premises. Quantitative Hedonists would argue that any dimension of quality can be better explained in terms of dimensions of quantity. For example, they might claim that moral pleasures are no higher in quality than immoral pleasures, but that moral pleasures are instrumentally more valuable because they are likely to lead to more moments of pleasure or fewer moments of pain in the future.

Hedonists also have differing views about how the value of pleasure compares with the value of pain. This is not a practical disagreement about how best to measure pleasure and pain, but rather a theoretical disagreement about comparative value. An example of such a disagreement is the debate about whether pain is worse for us than an equivalent amount of pleasure is good for us. The default position in philosophy is that one unit of pleasure (sometimes referred to as a Hedon) is equivalent but opposite in value to one unit of pain (sometimes referred to as a Dolor) (Feldman 2004, p. 26). Several Hedonistic Utilitarians have argued that reduction of pain should be seen as more important than increasing pleasure, sometimes for the Epicurean reason that pain seems worse for us than an equivalent amount of pleasure is good

for us.¹⁶ Imagine that a magical genie offered for you to play a game with him. The game consists of you flipping a fair coin. If the coin lands on heads, then you immediately feel a burst of very intense pleasure and if it lands on tails, then you immediately feel a burst of very intense pain. Is it in your best interests to play the game?

Another area of disagreement amongst some Hedonists is whether pleasure is entirely internal to a person or if it includes external elements. Internalism about pleasure is the thesis that, whatever pleasure is, it is always and only inside a person. Externalism about pleasure, on the other hand, is the thesis that, pleasure is more than just a state of an individual (i.e. that a necessary component of pleasure lies outside of the individual). Externalists about pleasure might, for example, describe pleasure as a function that mediates between our minds and the environment, such that every instance of pleasure has one or more integral environmental components. The vast majority of historic and contemporary versions of Prudential Hedonism consider pleasure to be an internal mental state.¹⁷

One of the least known disagreements about what aspects of pleasure make it valuable is the debate about whether we have to be conscious of pleasure for it to be valuable. The dominant position is that pleasure is a conscious mental state, or at least that any pleasure a person is not conscious of does not intrinsically improve their well-being (Bramble forthcoming). Torbjörn Tännsjö (1998, chap. 5), Dan Haybron (2008b), and Eric Schwitzgebel

¹⁶ Negative Utilitarianism was inspired by Karl Popper (1952) and coined by Ninian Smart (1958, pp. 542–543) as the normative ethical theory that demanded "the least amount of avoidable suffering for all" and is usually motivated by the notion that suffering (often understood as pain) is much worse than an equivalent amount of happiness (often understood as pleasure). It should be noted that Ninian Smart was *not* a proponent of Negative Utilitarianism.

¹⁷ See Sarch (2011) for the most recent discussion of this debate.

(2008) provide the strongest arguments for the possibility that some pleasures that we are not conscious of may still have prudential value.

5.2 Pleasure as Sensation

The most common definition of pleasure is that it is a sensation, something that we identify through our senses or that we feel. Psychologists claim that we have at least ten senses, including the familiar sight, hearing, smell, taste, and touch, but also movement, balance, and several sub-senses of touch, including heat, cold, pressure, and pain. New senses get added to the list when it is understood that some independent physical process underpins their functioning.¹⁸ The most widely-used examples of pleasurable sensations are the pleasures of eating, drinking, listening to music, and having sex (Davis 1981, p. 312). Use of these examples has done little to help Hedonism avoid its debauched reputation.

It is also commonly recognised that our senses are physical processes that usually involve a mental component, such as the tickling feeling when someone blows gently on the back of your neck.¹⁹ If a sensation is something we identify through our sense organs, however, it is not entirely clear how to account for abstract pleasures. This is because abstract pleasures, such as a feeling of accomplishment for a job well done, do not seem to be experienced through any of the senses in the standard lists. Hedonists might attempt to resolve this problem by arguing for the existence of an independent pleasure sense and by defining sensation as something that we simply feel (regardless of whether it has been mediated by sense organs).

¹⁸ See Macpherson (2011) for a much more detailed discussion.

¹⁹ Hirsch and Liebert's (1998) experiments on pain are an excellent demonstration of this.

Most Hedonists who describe pleasure as a sensation are Quantitative Hedonists and argue that the pleasure from different senses is essentially the same (Crisp 2006b, p. 109; e.g. Smuts 2011). Qualitative Hedonists, in comparison, could use the framework of the senses to help differentiate between qualities of pleasure. For example, a Qualitative Hedonist might argue that pleasurable sensations from touch and movement are always lower quality than those from the other senses.

5.3 Pleasure as Intrinsically Valuable Experience

Hedonists have also defined pleasure as ‘intrinsically valuable experience’, that is to say any experiences that we find intrinsically valuable either are, or include, instances of pleasure (e.g. Smuts 2011). According to this definition, the reason that listening to music and eating a fine meal are both intrinsically pleasurable is that those experiences include an element of pleasure (along with the other elements specific to each activity, such as the experience of the texture of the food and the melody of the music). By itself, this definition enables Hedonists to make an argument that is close to perfectly circular. Defining pleasure as intrinsically valuable experience, and well-being as all and only experiences that are intrinsically valuable, allows a Hedonist to all but stipulate that Prudential Hedonism is the correct theory of well-being. Where defining pleasure as intrinsically valuable experience is not circular is in its stipulation that only experiences matter for well-being. Some well-known objections to this idea are discussed below.

Another problem with defining pleasure as intrinsically valuable experience is that the definition does not tell us very much about what pleasure is or how it can be identified. For example, knowing that pleasure is intrinsically valuable experience would not help someone to work out if a particular experience were intrinsically or just instrumentally valuable.

Hedonists have attempted to respond to this problem by explaining how to discover whether an experience is intrinsically valuable.

One method is to ask yourself if you would like the experience to continue for its own sake, rather than because of what it might lead to (e.g. Brandt 1966, pp. 268–269). Wanting an experience to continue for its own sake reveals that you find it to be intrinsically valuable. A similar definition of pleasure describes it as a feeling that is apprehended as desirable (Sidgwick 1907, p. 127). Preference Hedonism is a form of Prudential Hedonism that defines pleasure in these ways. While Preference Hedonism does represent a coherent theory of well-being, defining intrinsically valuable experiences as those you want to perpetuate, or understand as desirable, makes the theory quite different to traditional hedonistic theories of well-being. Indeed, the fact that what a person wants, or considers desirable, is the main criterion for something having intrinsic value, makes these kinds of theories more similar to preference satisfaction theories of well-being (which have their own strengths and weaknesses). The central claim of preference satisfaction theories of well-being is that some variant of getting what one wants, or should want, under certain conditions is the only thing that intrinsically improves one's well-being.

Another method of fleshing out the definition of pleasure as intrinsically valuable experience is to describe how intrinsically valuable experiences actually feel. This method produces accounts of pleasure that are much closer to the concept described by traditional Prudential Hedonists but the actual description of what all pleasures feel like is often thought to be unconvincing in light of the extensive variety of experiences that we generally think of as 'pleasures' (discussed below).

It has also been argued that what makes an experience intrinsically valuable is that you like or enjoy it for its own sake (e.g. Smuts 2011).

Hedonists arguing for this definition of pleasure usually take pains to position their definition in between the realms of sensation and preference satisfaction. They argue that since we can like or enjoy some experiences without concurrently wanting them, or feeling any particular sensation, then liking is distinct from both sensation and preference satisfaction.²⁰ Liking and enjoyment are also difficult terms to define in more detail, but they are certainly easier to recognise than the rather opaque ‘intrinsically valuable experience’.

Merely defining pleasure as intrinsically valuable experience and intrinsically valuable experiences as those that we like or enjoy still lacks enough detail to be very useful for contemplating well-being. A potential method for making this theory more useful would be to draw on the cognitive sciences to investigate if there is a specific neurological function for liking or enjoying. Cognitive science has not reached the point where anything definitive can be said about this, but a few neuroscientists have provided experimental evidence that liking and wanting (at least in regards to food) are neurologically distinct processes in rats and have argued that it should be the same for humans (Berridge & Kringelbach 2011, p. 1).²¹ The same scientists have wondered if the same processes govern all of our liking and wanting, but this question remains untested and unresolved.

If these scientists’ wonderings are correct, however, it would provide considerable support for the hedonic tone theory of pleasure. The hedonic tone theory of pleasure claims that all experiences are given a neurological

²⁰ David Sobel (1999) considers this to be an untenable position, but, as discussed below, if scientific support for the hedonic tone theory becomes widespread, theories like Smuts’ might be considered much more reasonable.

²¹ See Kringelbach & Berridge (2010, pp. 17–19) for a collection of leading cognitive scientists’ opinions about whether there is a “common currency” for all pleasures.

tag, or gloss, that indicates whether it is liked or disliked and to what degree. The hedonic tone theory has been supported by Broad (1959, pp. 229–231), Duncker (1941), and Smuts (2011). Without firm scientific support, however, hedonic tone theory cannot convincingly refute the objection that there is no coherent and unifying definition of pleasure (discussed below).

Most Hedonists who describe pleasure as ‘intrinsically valuable experience’ believe that pleasure is internal and conscious. Hedonists who define pleasure in this way could also be either Quantitative or Qualitative Hedonists depending on whether they think that quality is a relevant dimension of how intrinsically valuable we find certain experiences.

5.4 Pleasure as Pro-Attitude

One of the most recent developments in hedonism is the rise of defining pleasure as a pro-attitude—a positive psychological stance toward some object. Any account of Prudential Hedonism that defines pleasure as a pro-attitude is referred to as Attitudinal Hedonism because it is a person’s attitude that dictates whether anything has intrinsic value. Positive psychological stances include approving of something, thinking it is good, and being pleased about it. The object of the positive psychological stance could be a physical object, such as a painting one is observing, but it could also be a belief, such as ‘my country is not at war’, or even a sensation. An example of a pro-attitude towards a sensation could be being pleased about the fact that the ice cream you are eating tastes so delicious.

Fred Feldman (2004), the leading proponent of Attitudinal Hedonism, argues that the sensation of pleasure has only instrumental value—it only brings you value if you also have a positive psychological stance toward that sensation. In addition to his basic Intrinsic Attitudinal Hedonism, which is a form of Quantitative Hedonism, Feldman (2004) has also developed many

variants that are types of Qualitative Hedonism. For example, Desert-Adjusted Intrinsic Attitudinal Hedonism, which reduces the intrinsic value a pro-attitude has for a person's well-being based on the quality of deservedness (i.e. on the extent to which the particular object deserves a pro-attitude or not) (Feldman 1997, part 3; 2004, pp. 120–122). Desert-Adjusted Intrinsic Attitudinal Hedonism might stipulate that sensations of pleasure arising from adulterous behaviour do not deserve approval, and so assign them no value.

Defining pleasure as a pro-attitude (while maintaining that all sensations of pleasure have no intrinsic value) makes Attitudinal Hedonism importantly dissimilar to traditional accounts of Prudential Hedonism. Indeed, defining pleasure as a pro-attitude runs the risk of creating a preference satisfaction account of well-being because being pleased about something (without feeling any pleasure) seems hard to distinguish from having a preference for that thing.

6. Contemporary Objections

6.1 Pleasure is Not the Only Source of Intrinsic Value

The most common argument against Prudential Hedonism is that pleasure is not the only thing that intrinsically contributes to well-being. Living in reality, finding meaning in life, producing noteworthy achievements, building and maintaining friendships, achieving perfection in certain domains, and living in accordance with religious or moral laws are just some of the other things thought to intrinsically add value to our lives (Sinnott-Armstrong 2011). When presented with these apparently valuable aspects of life, Hedonists usually attempt to explain their apparent value in terms of pleasure. A Hedonist would argue, for example, that friendship is not valuable in and of itself, rather it is valuable to the extent that it brings us pleasure. Furthermore,

to answer why we might help a friend even when doing so harms us, a Hedonist will argue that the prospect of future pleasure from receiving reciprocal favours from our friend, rather than the value of friendship itself, should motivate us to help in this way.

Those who object to Prudential Hedonism on the grounds that pleasure is not the only source of intrinsic value use two main strategies. In the first strategy, objectors make arguments that some specific value cannot be reduced to pleasure. In the second strategy, objectors cite very long lists of apparently intrinsically valuable aspects of life and then challenge Hedonists with the arduous task of trying to explain how the value of each of them can be explained solely by reference to the pursuit of pleasure and the avoidance of pain. This second strategy provides good reason to be a pluralist about value because the odds seem to be against any monistic theory of value, such as Prudential Hedonism. The first strategy, however, has the ability to show that Prudential Hedonism is false, rather than being just unlikely to be the best theory of well-being.

The most widely cited argument for pleasure not being the only source of intrinsic value is based on Robert Nozick's experience machine thought experiment (Nozick 1974, pp. 42–45).²² Nozick's experience machine thought experiment was designed to show that more than just our experiences matter to us because living in reality also matters to us (Feldman 2011). This argument has proven to be so convincing that nearly every single recent book on ethics that discusses hedonism cavalierly rejects it using only this argument or this one and one other (Crisp 2006a, pp. 619–620; Feldman 2004,

²² Although, the most influential, Nozick's experience machine thought experiment is not the first use of such a thought experiment to investigate the good life; Jack Smart, for example, used a less vivid depiction of an experience machine to make a similar point a year earlier (Smart 1973, pp. 19–22). See Chapter 2 for more detail on this.

p. 7; Silverstein 2000, p. 279). This argument will not be discussed in any more detail here since it is the topic of the next few chapters.

Suffice it say for now that even if Nozick's experience machine thought experiment is not as decisive a refutation of Prudential Hedonism as it is often thought to be, the wider argument (that living in reality is valuable for our well-being) is still problematic for Prudential Hedonists. That our actions have real consequences, that our friends are real, and that our experiences are genuine seem to matter for most of us regardless of considerations of pleasure. Unfortunately, we lack a trusted methodology for discerning if these things *should* matter to us. Perhaps the best method for identifying intrinsically valuable aspects of lives is to compare lives that are equal in pleasure and all other important ways, except that one aspect of one of the lives is improved. Using this methodology, however, seems likely to lead to an artificially pluralist conclusion about what has value. This is because any increase in a potentially valuable aspect of our lives is likely to be viewed as a free bonus. And, most people will probably choose the life with the free bonus just in case it has intrinsic value, not necessarily because they think it does have intrinsic value.

6.2 Some Pleasure is Not Valuable

The main traditional line of criticism against Prudential Hedonism is that not all pleasure is valuable for well-being, or at least that some pleasures are less valuable than others because of non-amount-related factors. Some versions of this criticism are much easier for Prudential Hedonists to deal with than others depending on where the allegedly dis-valuable aspect of the pleasure resides. If the dis-valuable aspect is experienced with the pleasure itself, then both Qualitative and Quantitative varieties of Prudential Hedonism have sufficient answers to these problems. If, however, the dis-valuable aspect of

the pleasure is never experienced, then all types of Prudential Hedonism struggle to explain why the allegedly dis-valuable aspect is irrelevant.

Examples of the easier criticisms to deal with are that Prudential Hedonism values, or at least overvalues, perverse and base pleasures. These kinds of criticisms tend to have had more sway in the past, and doubtless encouraged Mill to develop his Qualitative Hedonism. In response to the charge that Prudential Hedonism mistakenly values pleasure from sadistic torture, sating hunger, copulating, listening to opera, and philosophising all equally, Qualitative Hedonists can simply deny that it does. Since pleasure from sadistic torture will normally be experienced as containing the quality of sadism (just as the pleasure from listening to good opera is experienced as containing the quality of acoustic excellence), the Qualitative Hedonist can plausibly claim to be aware of the difference in quality and allocate less value to perverse or base pleasures accordingly.

Prudential Hedonists need not relinquish the Quantitative aspect of their theory in order to deal with these criticisms, however. Quantitative Hedonists, can simply point out that moral or cultural values are not necessarily relevant to well-being because the investigation of well-being aims to understand what the good life for the one living it is, and what *intrinsically* makes their life go better for them. A Quantitative Hedonist can simply respond that a sadist who gets sadistic pleasure from torturing someone does improve their own well-being (assuming that the sadist never feels any negative emotions or gets into any other trouble as a result). Similarly, a Quantitative Hedonist can argue that if someone genuinely gets a lot of pleasure from porcine company and wallowing in the mud, but finds opera thoroughly dull, then we have good reason to think that having to live in a pig sty would be better for her well-being than having to listen to opera.

Much more problematic for both Quantitative and Qualitative Hedonists, however, are the more modern versions of the criticism that not all pleasure is valuable. The modern versions of this criticism tend to use examples in which the dis-valuable aspect of the pleasure is never experienced by the person whose well-being is being evaluated. The best example of these modern criticisms is the deceived businessman thought experiment devised by Shelly Kagan (1998, pp. 34–36).²³ This thought experiment is widely thought to show that pleasures of a certain kind, namely false pleasures, are worth much less than true pleasures.

Kagan (1998, pp. 34–36) asks us to imagine the life of a very successful businessman who takes great pleasure in the respect of his colleagues, amity of his friends, and love of his wife and children until the day he dies. Then Kagan asks us to compare this life with one of equal length and the same amount of pleasure (experienced as coming from exactly the same sources), except that in each case the businessman is mistaken about how those around him really feel. This second (deceived) businessman experiences just as much pleasure from the respect of his colleagues and the love of his family as the first businessman. The only difference is that the second businessman has many false beliefs. Specifically, the deceived businessman's colleagues actually think he is useless, his wife doesn't really love him, and his children are only nice to him so that he will keep giving them money. Given that the deceived businessman never knew of any of these deceptions and his experiences were never negatively impacted by these deceptions indirectly, which life do you think is better?

²³ Kagan's thought experiment is based on Nagel's earlier presentation of the idea that being deceived is bad for us (Kagan 1994, p. 311; Nagel 1970, pp. 76–78; 1979, pp. 4–7). Kagan's clearest discussion of the deceived businessman thought experiment can be found in his *Normative Ethics* (1998, pp. 34–36).

Nearly everyone thinks that the deceived businessman has a worse life. This is a problem for Prudential Hedonists because the pleasure is quantitatively equal in each life, so they should be equally good for the men living them. Qualitative Hedonism does not seem to be able to avoid this criticism either because the falsity of the pleasures experienced by the deceived businessman is a dimension of the pleasure that he never becomes aware of. Theoretically, an externalist and qualitative version of Attitudinal Hedonism could include the falsity dimension of an instance of pleasure (even if the falsity dimension never impacts the consciousness of the person) to attempt to avoid this objection.²⁴ However, the resulting definition of pleasure bears little resemblance to what we commonly understand pleasure to be and also seems to be ad hoc in its inclusion of the truth dimension but not others. Besides, a dedicated Prudential Hedonist of any variety can always stubbornly stick to the claim that the lives of the two businessmen are of equal value because the deceptions never affect the deceived party's mental states. But this argument will do little to convince the vast majority to take Prudential Hedonism more seriously.

6.3 There is No Coherent and Unifying Definition of Pleasure

Another major line of criticism used against Prudential Hedonists is that they have yet to come up with a meaningful definition of pleasure that unifies the seemingly disparate array of conceivable pleasures while remaining recognisable as pleasure. Some definitions lack sufficient detail to be informative about what pleasure actually is, or why it is valuable, and those that do offer enough detail to be meaningful are faced with two difficult tasks.

²⁴ As Feldman has done with his Truth-Adjusted Intrinsic Attitudinal Hedonism (2004, pp. 109–114).

The first obstacle for a useful definition of pleasure for hedonism is to unify all of the diverse pleasures in a reasonable way. Phenomenologically, the pleasure from reading a good book is very different to the pleasure from bungee jumping, and both of these pleasures are very different to the pleasure of having sex.²⁵ This obstacle is unsurpassable for most versions of Quantitative Hedonism because it makes the value gained from different pleasures impossible to compare. Not being able to compare different types of pleasure results in being unable to discern if a life is better than another in most even vaguely realistic cases. Furthermore, its inability to compare lives means that Quantitative Hedonism could not be usefully used to guide behaviour since it cannot instruct us on which life to aim for.

Attempts to resolve the problem of unifying the different pleasures while remaining within a framework of Quantitative Hedonism, usually involve pointing out something that is constant in all of the disparate pleasures and defining that particular thing as pleasure. When pleasure is defined as a strict sensation, this strategy fails because introspection reveals that no such sensation exists.²⁶ Pleasure defined as the experience of liking or as a pro-attitude does much better at unifying all of the diverse pleasures. However, defining pleasure in these ways makes the task of filling in the details of the theory a fine balancing act. Liking or having a pro-attitude must be described in such a way that they are not solely a sensation or in a way that they would best fit in a preference satisfaction theory of well-being instead of a hedonistic one. Furthermore, they must perform this balancing act while still describing a scientifically plausible and conceptually coherent account of pleasure. Most

²⁵ Usually known as the heterogeneity problem, this objection to Prudential Hedonism has been discussed by many philosophers over the years, but most famously by Sidgwick (1907, p. 127).

²⁶ Although Ben Bramble (forthcoming) disputes this widely agreed-upon claim.

attempts to define pleasure as liking or pro-attitudes seem to disagree with either the folk conception of what pleasure is or any of the plausible scientific conceptions of how pleasure functions.

Most varieties of Qualitative Hedonism do better at dealing with the problem of diverse pleasures because they can evaluate different pleasures according to their distinct qualities. Qualitative Hedonists still need a coherent method for comparing the different pleasures with each other in order to be more than just an abstract theory of well-being, however. And, it is difficult to construct such a methodology in a way that avoids counter examples while still describing a scientifically plausible and conceptually coherent account of pleasure.

The second obstacle is creating a definition of pleasure that retains at least some of the core properties of the common understanding of the term 'pleasure'. As mentioned, many of the potential adjustments to the main definitions of pleasure are useful for avoiding one or more of the many objections against Prudential Hedonism. The problem with this strategy is that the more adjustments that are made, the more apparent it becomes that the definition of pleasure is not recognisable as the pleasure that gave Hedonism its distinctive intuitive plausibility in the first place. When an instance of pleasure is defined simply as when someone feels good, its intrinsic value for well-being is intuitively obvious. However, when the definition of pleasure is stretched, so as to more effectively argue that all valuable experiences are pleasurable, it becomes much less recognisable as the concept of pleasure we use in day-to-day life and its intrinsic value becomes much less intuitive.

7. The Future of Hedonism

All things considered, the future of hedonism seems relatively bleak. Hedonists should find the considerable number and strength of the arguments against Prudential Hedonism's central principle—that pleasure and only pleasure intrinsically contributes positively to well-being, and the opposite for pain—daunting, if not overwhelming. Hedonists have been creative in their definitions of pleasure so as to avoid these objections, but more often than not the accounts they end up defending are seen as not in line with traditional hedonistic concept of pleasure, not particularly realistic, or both.

Perhaps the only hope that Hedonists of all types can have for the future is that advances in cognitive science lead to a better understanding of how pleasure works in the brain and how biases affect our judgements about thought experiments.²⁷ If our improved understanding in these areas confirms a particular theory about what pleasure is, and also provides reasons to doubt some of the widespread judgements about the thought experiments that make the vast majority of philosophers reject hedonism, then hedonism might experience at least a partial revival. The good news for Hedonists is that at least some emerging theories and results from cognitive science do appear to support a key aspect of hedonism. Indeed, a few neuroscientists endorse a hedonic tone view of pleasure and pain—they think it's likely that all pleasures (and all pains) use the same underlying neural mechanisms (Berridge & Kringelbach 2011).

²⁷ A topic that is highly related to evaluating the experience machine objection to hedonism, which is the focus of the next few chapters.

8. Conclusion

This chapter disambiguated the term 'hedonism' and provided a broad overview of Prudential Hedonism, including its history, main variants, and the major criticisms directed at it. The disambiguations in this chapter (particularly between Prudential Hedonism and other types of hedonism, and between internalist and externalist accounts of hedonism) will help to identify the scope of the experience machine objection to hedonism. The experience machine objection to hedonism was the only major criticism that was not discussed in any detail in this chapter. Instead, this important objection is discussed in great detail in the remainder of Part 1 (Chapters 2 to 5).

Chapter 2

The Experience Machine Objection to Hedonism

Chapter Summary

Prudential Hedonism has been beset by many objections, the strength and number of which have led most modern philosophers to believe that it is implausible. One objection in particular, however, is nearly always cited when a philosopher wants to argue that Prudential Hedonism is implausible—the experience machine objection to hedonism. This chapter examines this objection in detail. First, the deductive and abductive versions of the experience machine objection to hedonism are explained. Then a taxonomy of the contemporary attempts to refute the abductive version of the experience machine objection to hedonism is created and the contemporary responses to each version of the argument are assessed. Consideration of responses shows that while the deductive version of the objection is implausible, the abductive version is fairly powerful. However, one type of response seems promising against the abductive version. This response argues that experience machine thought experiments elicit judgments that are too biased to be used as evidence for the experience machine objection to hedonism. It is argued that only this type of refutation seems likely to convince proponents of the abductive version of the experience machine objection to hedonism that the objection is much weaker than they believe it to be. Finally, it is suggested that more evidence is required before anything definitive can be said on the matter.

1. Introduction²⁸

As remarked several times in the literature, so many strong objections have been levelled at Prudential Hedonism that most modern philosophers believe it to be implausible.²⁹ One objection in particular, however, is nearly always cited when a philosopher wants to argue that internalist accounts of Prudential Hedonism are implausible—the experience machine objection to hedonism (Barber 2011, p. 257). Indeed, virtually everyone who has written about Prudential Hedonism since the late 1970s cites the experience machine thought experiment as a (and often the) decisive objection against it (Tiberius 2006, p. 496).³⁰

In 1957 popular author Ray Bradbury published a short story, *The Happiness Machine*, in his collection of short stories *Dandelion Wine*. Bradbury described the happiness machine as a fantastic contraption that can simulate all kinds of wonderful experiences, such as sensing the sights, sounds, and tastes of Paris. Despite the apparent appeal of such a machine,

²⁸ This chapter provided the basis for Weijers (2011b).

²⁹ See for example: Crisp (2006a, pp. 619–620) Feldman (2004, p. 7) Silverstein (2000, p. 279), and Weijers (2011a; forthcoming-a).

³⁰ Examples of authors who have stated or implied that the experience machine thought experiment is a knock-down refutation of internalist accounts of Prudential Hedonism (or all internalist mental state accounts of well-being) are easily found (e.g. Attfield 1987, p. 33; Baggini & Fosl 2007, pp. 74–76; Becker 1992, p. 25; Brink 1989, pp. 223–224; Brülde 2007, pp. 26–29, 33; S. Bok 2010, pp. 24–28; Darwall 1997, pp. 162, 178; Feldman 2002, p. 615; Finnis 1980, p. 33; 1983, pp. 37–42; Griffin 1986, pp. 9–10; Hausman 2010, p. 329; Haybron 2008a, p. 21; Hooker 2000, p. 39; Hurka 2011, pp. 68–70; Jollimore 2004, pp. 333–334; Kagan 1998, pp. 34–36; 2009, p. 253; Kazez 2007, pp. 51–54; Keller 2009, p. 657; Kraut 2007, pp. 124–126; Kymlicka 1990, pp. 13–14; Lopez 2007, p. 75; Nozick 1989, pp. 99–117; Railton 1984, pp. 148–149; Sobel 2002, p. 244; Tiberius 2004; p. 311, n. 4; Tiberius & Hall 2010, pp. 214–215; Thomson 1987, p. 41; van Wyk 1990, p. 109).

Bradbury's story highlighted the dangers of it, including that the experiences the machine creates only give the illusion of true happiness. Perhaps it was this intriguing story that led to both Jack Smart and Robert Nozick discussing the philosophical implications of such a machine. Indeed, Jack Smart may have been the first philosopher to point out that our disinclination to use such a machine creates problems for Prudential Hedonism (1973, pp. 19–22). Without a doubt, however, it was Robert Nozick's vivid description of an experience machine that popularised the idea amongst philosophers:

Suppose that there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time, you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences? If you are worried about missing out on desirable experiences, we can suppose that business enterprises have researched thoroughly the lives of many others. You can pick and choose from their large library or smorgasbord of such experiences, selecting your life's experiences for, say, the next two years. After two years have passed, you will have ten minutes or ten hours out of the tank, to select the experiences of your *next* two years. Of course, while in the tank you won't know that you're there; you'll think it's all actually happening. Others can also plug in to have the experiences they want, so there's no need to stay unplugged to serve them. (Ignore problems such as who will service the machines if everyone plugs in.) Would you plug in? *What else can matter to us, other*

than how our lives feel from the inside? Nor should you refrain because of the few moments of distress between the moment you've decided and the moment you're plugged. What's a few moments of distress compared to a lifetime of bliss (if that's what you choose), and why feel any distress at all if your decision *is* the best one? (Nozick 1974, pp. 42–3, his italics)

Although Nozick originally devised the experience machine thought experiment to make a point about how animals should be treated, it was quickly adopted by anyone who wanted to argue for the falsity of Prudential Hedonism (Weijers 2011b). The experience machine thought experiment is equally effective against any kind of theory that posits the internal aspects of our experiences as the only valuable things in a life, but Prudential Hedonism is often singled out because it is the most widely discussed exemplar of this type of theory.

As discussed in Chapter 1, accounts of Prudential Hedonism hold that all pleasure and only pleasure intrinsically contributes positively to well-being (and the opposite for pain). Internalist mental state theories of well-being hold that only the internal aspects of our beliefs, desires, feelings, and other mental states intrinsically affect our well-being. Internalist mental state theorists about well-being do not dispute that external events, such as winning the lottery, can impact our well-being. However, they would argue that winning the lottery only affects our well-being instrumentally (and only to the extent that it affects the internal aspects of our mental states). For example, according to hedonistic variants of internalist mental state theories about well-being, winning the lottery is usually a good thing, not because winning lots of money is good in and of itself, but because winning lots of money

tends to make people feel happier.³¹ Any use of the term 'Prudential Hedonism' (or its linguistic derivatives) from here on refers to all hedonistic variants of internalist mental state theories about well-being.

This chapter explains the experience machine objection to hedonism and why it has been so influential. A taxonomy of the contemporary attempts to refute the experience machine objection to hedonism is also created and the responses are evaluated along the way. One particular type of response argues that experience machine thought experiments elicit judgments that are too biased to be used as evidence for the experience machine objection to hedonism. It is argued that this type of refutation seems the most likely to convince proponents of the experience machine objection to hedonism that the objection is much weaker than they believe it to be. Finally, it is suggested that more evidence is required before anything definitive can be said on the matter.

2. The Argument

The vast majority of people who read Nozick's scenario think that they would choose to remain in reality.³² The experience machine objection to hedonism relies on this widespread judgment and uses it to infer that there is more to the good life for the one living it than how our experiences feel to us on the inside. A major strength of the experience machine objection to hedonism is how much it concedes to its opponents while still producing a resounding verdict against them. A life in an experience machine is not described as just slightly more pleasurable than a real life, but rather as a "lifetime of bliss" in which you can receive any and all of the best experiences possible (Nozick

³¹ At least until they become accustomed to their newfound wealth or squander it and revert back to feeling about as happy as they did before their windfall.

³² Empirical evidence for this claim is provided in Chapter 5.

1974, p. 43). The overwhelming influence of the experience machine objection to hedonism is doubtless caused by a combination of this great concession to proponents of pleasure and the dramatic impact the idea of an experience machine tends to have on our imaginations and memories. Furthermore, because the lack of direct connection with reality in an experience machine life is usually assumed to be the only relevant difference between the two options in Nozick's scenario, most people also infer that living in reality must make our lives go better for us regardless of whether it leads to increased enjoyment. It deserves to be emphasised that the main justification for both of these inferences is the widespread judgement that connecting to an experience machine in Nozick's scenario is a bad idea.

Proponents of the experience machine objection to hedonism very rarely construct it as a formal argument. Indeed, it is often discussed in a page or less before it is acknowledged as a complete refutation of Prudential Hedonism. Will Kymlicka gives it this treatment:

Now if pleasure were our greatest good, then we would all volunteer to be hooked for life to this machine... But surely very few people would volunteer. Far from being the best life we can lead, it hardly counts as leading a life at all... The hedonistic account of utility is wrong, for the things worth doing and having in life are not all reducible to one mental state like happiness. (Kymlicka 1990, p. 13)

The loose language many philosophers have used to explain the experience machine objection to hedonism has led to two distinct interpretations of it, one deductive and the other abductive in structure.³³

³³ As far as I am aware, only Alex Barber (2011), Torbjörn Tännsjö (2007), and Matthew Silverstein (2000) have drawn attention to the distinction between deductive and abductive

The deductive version of the experience machine objection to hedonism is formalised below.

DP1. In terms of the internal aspects of our experiences, an experience machine life would be much better than a life in reality. (Stipulated in thought experiment)

DP2. When instructed to ignore their responsibilities to others, the vast majority of reasonable people report preferring reality over a life in an experience machine. (Empirical claim)

DP3. If when instructed to ignore their responsibilities to others, the vast majority of reasonable people report preferring reality over a life in an experience machine, then reality matters intrinsically to the vast majority of reasonable people

DC1. Therefore, reality matters intrinsically to the vast majority of reasonable people. (*Modus ponens* DP2, DP3)

DP4. If something matters intrinsically to the vast majority of reasonable people, then that thing has intrinsic prudential value

DC2. Therefore, reality has intrinsic prudential value. (*Modus ponens* DC1, DP4)

DP5. If internalist Prudential Hedonism is true, then the internal aspects of pleasure and pain are the only things of intrinsic prudential value (or disvalue) in a life. (Stipulated definition)

DC3. Therefore, internalist Prudential Hedonism is false. (*Modus tollens*, DC2, DP5)

versions of the experience machine objection to hedonism, which is surprising given the distinction's importance for evaluating the widely-cited objection.

The abductive version of the experience machine objection to hedonism is formalised below.

- IP1. In terms of the internal aspects of our experiences, an experience machine life would be much better than a life in reality. (Stipulated in thought experiment)
- IP2. When instructed to ignore their responsibilities to others, the vast majority of reasonable people report preferring reality over a life in an experience machine. (Empirical claim)
- IP3. The best explanation for IP2 is that reality matters intrinsically to the vast majority of reasonable people
- IP4. Inference to the best explanation: If a hypothesis is the best explanation of an observation, then it is rational to believe that hypothesis is true. (Standard methodological premise)
 - IC1. Therefore, it is rational to believe that reality matters intrinsically to the vast majority of reasonable people.
(*Modus ponens* IP3, IP4)
- IP5. The best explanation for reality mattering intrinsically to the vast majority of reasonable people is that reality has intrinsic prudential value
- IP6. Inference to the best explanation. (Standard methodological premise)
 - IC2. Therefore, it is rational to believe that reality has intrinsic prudential value. (*Modus ponens* IP5, IP6)
- IP7. If internalist Prudential Hedonism is true, then the internal aspects of pleasure and pain are the only things of intrinsic prudential value (or disvalue) in a life. (Stipulated definition)

IC3. Therefore, it is rational to believe that internalist Prudential Hedonism is false. (*Modus tollens*, IC2, IP7)

Several points about the arguments should be noted. First, many published versions of this objection simply identify hedonism as their target, rather than internalist accounts of Prudential Hedonism (e.g. Kymlicka 1990, p. 13).³⁴ This is probably due to either ignorance of the different accounts of hedonism or simplification, perhaps for educational purposes. These simpler presentations of the target have likely influenced the philosophers who have argued that some variants of hedonism can avoid the experience machine objection because they are not traditional internalist accounts of pleasure or happiness.³⁵ Since it is not clear that the objection was ever really designed to refute anything other than internalist accounts of Prudential Hedonism, only the responses that attempt to defend internalist versions of hedonism will be discussed in this chapter.

Second, both arguments state that they are based on ‘reality’ mattering to people, but should be understood as being based on ‘something other than the internal aspects of pleasure and pain’ mattering to people. This later phrase is more likely to be true, better supported by the experience machine thought experiment, and about as good at giving reason to doubt hedonism. However, it is also particularly unwieldy; so to increase readability, ‘reality’ will be used instead.

³⁴ As discussed in Chapter 1, internalism about pleasure is the thesis that, whatever pleasure is, it is always and only inside a person. Externalism about pleasure, on the other hand, is the thesis that, pleasure is more than just a state of an individual (i.e. that a necessary component of pleasure lies outside of the individual).

³⁵ (e.g. Donner 1991; Feldman 2004; Heathwood 2007; Lopez 2007; Sumner 1996).

Third, if something ‘matters intrinsically’, then it seems to have intrinsic value. If reality matters intrinsically to me, then I believe that reality has value *qua* reality—that reality has value over and above the value of any other things that might come from it.

Finally, the deductive version of the experience machine objection to hedonism has a much stronger conclusion than the more nuanced abductive version. As would be expected, with this strong conclusion comes a strong evidential burden on the premises of the deductive version. The conclusion of the abductive version (‘it is rational to believe that internalist Prudential Hedonism is false’) is still strong enough to discredit Prudential Hedonism. Nevertheless, many philosophers have discussed the experience machine objection to hedonism as if it were a deductive refutation of hedonism, including Nozick himself (1989, pp. 99–117).

3. The Responses

Despite Lopez’s claim that the experience machine objection to hedonism has “never [been] seriously disputed” (2007, p. 75), many academics have published responses which they consider to be refutations of it. However, the fact that different responses continue to be published to this day probably supports Lopez’s claim that the experience machine objection to hedonism has never been *seriously* disputed. Furthermore, as we shall see, several of the responses only work against the deductive version of the experience machine objection to hedonism and so might not qualify as serious disputations. Responses to the deductive argument are considered first, followed by the responses to the abductive version. This compilation of responses is intended to be exhaustive of all the published paper-length attempts to refute the experience machine objection to hedonism during the last 15 years and

indicative of all of the types of critical responses to the experience machine objection to hedonism to date.

3.1 Responses to the Deductive Version

The responses discussed in this section are limited to those directed against the deductive argument. All of these related responses involve DP4 (repeated below for convenience).

DP4. If something matters intrinsically to the vast majority of reasonable people, then that thing has intrinsic prudential value

Harriet Baber (2008) argues that the experience machine objection to hedonism is unfair because it presupposes some form of preferentism, or desire-satisfaction account of well-being.³⁶ The common thread running through all preferentist, or desire-satisfaction accounts of well-being, is the principle that having (certain kinds of) our preferences satisfied is the only thing that intrinsically improves our well-being. Or, as Baber puts it, “according to [preferentism] what makes a state of affairs good for a person is her desiring it...” (Baber 2008, p. 134).

Baber describes the key premise of the experience machine objection to hedonism as follows: “If a reasonable and informed subject, *i*, would choose *S* over *S'*, then *S* would contribute more to *i*’s wellbeing than *S'*.” (Baber 2008, p. 133, her italics). In Baber’s interpretation of the key premise, we can see that reasonable and informed peoples’ choices (i.e. their preferences) dictate what contributes to well-being. This move should be considered problematic, but

³⁶ Torbjörn Tännsjö (2007, pp. 94–95) offers a similar response to the deductive version of the experience machine objection to hedonism, but he is much more careful to acknowledge that the inferential version of the argument is a better interpretation of it and that the inferential version does not suffer from this kind of response.

not for the reason provided by Baber. The experience machine objection to hedonism does not assume that preferentism is true because, as Nozick points out, our preferences about the experience machine do not directly dictate the values involved:

Notice that I am not saying simply that since we desire connection to actuality the experience machine is defective because it does not give us what we desire... for that would make “getting whatever you desire” the primary standard. Rather, I am saying that the connection to actuality is important whether or not we desire it—that is *why* we desire it—and the experience machine is inadequate because it doesn’t give us *that*.” (Nozick 1989, pp. 106–107, his italics)

Baber’s (2008, p. 133) version of the key premise skips an important step in the argument. Baber starts with the equivalent of ‘the vast majority of reasonable people report preferring reality over a life in an experience machine’ and skips over ‘reality matters intrinsically to the vast majority of reasonable people’ straight to ‘reality has intrinsic prudential value’. The first step is important because it shows how the preference that is revealed by contemplating the experience machine thought experiment is evidence for forming a judgement about what matters to us. That judgment (reality matters intrinsically) then leads to the judgement that it has intrinsic prudential value.

So, while Baber interprets ‘choosing’ as ‘having a preference for’ in her version of the key premise, ‘choosing’ seems better interpreted as ‘making a judgment about value’. Of course, that judgment might still be mistaken. But if the vast majority of reasonable and informed people make the same judgment, then we have reason to believe that it is not mistaken. Consider the

possibility that Nozick's experience machine scenario is found to elicit extremely biased judgments and that a new bias-free version of the experience machine thought experiment elicits widespread agreement from reasonable and informed people that a life in an experience machine is better than one in reality. This might give us reason to believe that some version of hedonism is true. The method used to reach this hedonism-endorsing conclusion is not preferentialism; it is the dominant method in moral philosophy. As Roger Crisp puts it: "Intuitions appropriately reflected upon are unavoidable in ethical theory." (Crisp 2006a, p. 636). After all, could there be any way to endorse hedonism above other theories of well-being without the judgments of reasonable people about whether some X has more intrinsic value than some Y?

Fred Feldman (2011) also criticises the move from 'matters' to 'has value', but on different grounds. Feldman presents a number of different versions of the experience machine objection to hedonism, but finds none of them particularly convincing. Essentially, Feldman thinks that what matters intrinsically to reasonable people is of little consequence to questions of intrinsic value because reasonable people lack certain traits that are required to really know what has intrinsic value (and therefore what should matter intrinsically to us). Most importantly, Feldman insists that the move from 'matters intrinsically to reasonable people' to 'has intrinsic value' only works if the reasonable people are also "axiologically insightful"—they already understand what has intrinsic value. Therefore, he concludes that his close analogue of DP4 is false.

Feldman (2011) rightly points out that changing the argument so that only what matters intrinsically to people who are both reasonable and axiologically insightful can actually tell us what has intrinsic value would produce a solid (deduction-supporting) bridge between 'matters intrinsically'

and 'has intrinsic value'. Feldman is also right that changing the argument in this way would create a new problem for it: now Hedonists could deny his version of DP2 (that reasonable and axiologically insightful people would choose reality over a life in an experience machine) and there would be no suitably axiologically insightful person who could disconfirm this denial.

Feldman's (2011) denial of DP4 reveals a flaw in the deductive version of the experience machine objection to hedonism. His refutation of the experience machine objection to hedonism is only effective against the deductive version, however. We don't have to assume that reasonable people know everything about axiology before we can infer that their agreement gives us a defeasible reason to believe that something has value. Again, how could we know anything about what has value if only the judgments of axiologically insightful people (who already understand what has intrinsic value) were relevant to questions of value? So Feldman's (2011) critique of the experience machine objection to hedonism gives us reason to believe that the more nuanced abductive version of the argument is the only viable one.

Sharon Hewitt (2009, p. 348) takes a related approach, arguing that: "even if our intuitions [about what matters intrinsically to us] are directly responsive to the existence of real relationships... this should not immediately lead us to conclude that these things are objectively intrinsically valuable."³⁷ Hewitt is arguing that the move from 'matters intrinsically' to 'has intrinsic value' cannot be deductive because there are reasons other than something having intrinsic value for us that make things matter intrinsically to us, including irrational ones. Or, as Barber (2011, p. 269, his italics) puts it: "It is not enough... [to] *think* that hedonism is mistaken, since we could be mistaken that it is mistaken." Jason Kawall (1999, p. 385) agrees for two

³⁷ This is not Hewitt's complete strategy to refute the experience machine objection to well-being. This argument is combined with others that are discussed below.

reasons. First, Kawall points out that: “While it is true that we value many things besides our mental lives, it could well be that the other values contribute to our personal well-being only through the effects they have on our mental lives” (1999, p. 385). Kawall also notes that “we can value more than our own well-being” which can lead to people sacrificing their well-being to further some other value, such as “the well-being of their children” or “the search for truth” (1999, p. 385). Hewitt (2009) and Kawall (1999) are entirely correct and, when combined with Feldman’s (2011) critique, their views give us strong reason to believe that DP4 is probably false.

Analysis of Baber’s (2008), Feldman’s (2011), Hewitt’s (2009) and Kawall’s (1999) responses to the experience machine objection to hedonism show that the move from ‘choosing’ or ‘mattering intrinsically’ to ‘has intrinsic prudential value’ is best understood as an inference to the best explanation, not as a deduction. Indeed, no careful philosopher should believe that widespread agreement between reasonable people on a matter of value *deductively proves* any particular conclusion. As Silverstein (2000, p. 299, his italics) puts it: “Most philosophers recognize this... and take the experience machine argument to be an *indirect* refutation of hedonism.”

Instead of trying to deductively prove what has value, philosophers should focus on making inferences to the best explanation. What then is the best explanation for something mattering intrinsically to the vast majority of reasonable and informed people? ‘That it has intrinsic value’ is a plausible answer to this question. Furthermore, if competing answers have been sought by reasonable people and none seem more credible, then there is good reason to believe ‘that it has intrinsic value’ is the correct answer. Of course, that belief should be amended if a more credible competing answer surfaces at any stage.

3.2 Responses to the Abductive Version

Some of the responses to the experience machine objection to hedonism discussed in this section may also apply to the deductive version, but they are included here because the deductive version has already been shown to be implausible. The result in this section is a taxonomy of all of the responses to the most plausible version of the experience machine objection to hedonism.

3.2.1 Responses to IP5

At least three authors have attempted to defend hedonism against the experience machine objection by denying IP5 (repeated below for convenience). These authors all argue that consideration of the process by which our preferences are formed gives us reason to doubt that reality having intrinsic prudential value best explains why it matters intrinsically to the vast majority of people.

IP5. The best explanation for reality mattering intrinsically to the vast majority of reasonable people is that reality has intrinsic prudential value

Matthew Silverstein argues against IP5 by appealing to the hedonistic basis of our seemingly anti-hedonistic preferences:

[O]ur experience machine intuitions reflect our desire to remain connected to the real world, to track reality. But according to the account of the relation between happiness and our desires outlined above, the desire to track reality owes its hold upon us to the role it has played in the creation of happiness. We acquire our powerful attachment to reality after finding again and again that deception almost always ends in suffering. We develop a desire to track reality because, in almost all cases, the connection

to reality is conducive to happiness. Our intuitive views about what is prudentially good, the views upon which the experience machine argument relies, owe their existence to happiness. (Silverstein 2000, p. 296)

According to Silverstein, our desires are created by our experience of what has brought us happiness in the past and our intrinsic desires (what matters intrinsically to us/what we desire for its own sake) are created by *repeated* experiences of what has brought us happiness in the past. This account of desire formation is based on the work of Richard Brandt (1979) and an idea of Peter Railton's (1989). Silverstein distances himself from Brandt's view slightly by stressing that his argument does not rely on Brandt's extreme claim that happiness-related "conditioning is the *only* fundamental process involved in the acquisition of desires" (Silverstein 2000, p. 293, his italics; Brandt 1979, p. 100). Instead, Silverstein claims that all desires are created by happiness and that happiness is the main influence on our desires. This allows for some desires to be affected by non-happiness-related factors, although Silverstein doesn't discuss any such factors.

If Silverstein's (2000) account of desire creation is true, then why doesn't everyone simply save time by desiring only happiness? Silverstein quotes both Sidgwick and Mill on what has become known as the paradox of hedonism (and also the paradox of happiness) to answer this question.³⁸ The

³⁸ Sidgwick: "[The] fundamental paradox of Hedonism, that the impulse towards pleasure, if too predominant, defeats its own aim." (1907, p. 48). Mill: "I now thought that this end [happiness] was only to be attained by not making it the direct end. Those only are happy (I thought) who have their minds fixed on some object other than their own happiness.... Aiming thus at something else, they find happiness by the way." (1969, pp. 85–86).

paradox of hedonism is that pursuing happiness directly is likely to bring about much less happiness than pursuing other goods. While not strictly a paradox, there is certainly a lot of *prima facie* tension between the propositions 'happiness is the greatest good' and 'happiness should not be pursued'. The paradox of hedonism is well-supported in the philosophical literature, but it seems to be most true of Folk Hedonism (discussed in Chapter 1), which usually entails the greedy pursuit of immediate pleasures at the risk of harm and unhappiness for themselves and others sometime after. Drug-taking (when the drug causes harmful side-effects) is a good example of how taking the direct route to happiness does not pay off in the long run. But even if we wanted to pursue happiness directly, it is not clear how we would do it. There is no genuine 'happiness shop'; so our attempts to attain happiness are always indirect to some extent (including drug-taking).

Perhaps Silverstein's (2000) example of why reality and truth matter intrinsically to us is the best example of the lessons to be learnt from the paradox of hedonism. Deception has brought pain so consistently, Silverstein argues, that reality and truth have come to matter intrinsically to us; we will experience much more happiness if we pursue truth and reality than if we try to pursue happiness as directly as possible. We might wish to live in a bubble of false beliefs because lots of facts about the world can make us unhappy (e.g. your partner is cheating on you). In the vast majority of cases, however, we will be happier in the long run if we face up to the truth and try to make the best of the situation (e.g. confront your partner and sort out your problems or break up with them and get a new more loyal partner). So, the lessons we should learn from contemplation of the paradox of hedonism are that we shouldn't always prefer what will make us happy in the moment and that certain other goods mattering intrinsically to us is the best way to achieve this.

Silverstein is using his account of intrinsic desire creation to deny IP5 by arguing that the best explanation for reality mattering intrinsically to the vast majority of people is not that reality has intrinsic prudential value. Rather, it is that preferring reality (avoiding deception) nearly always leads to happiness in the long run. Silverstein (2000, p. 297) also argues that all of our intrinsic desires are formed this way and concludes that: "The most plausible explanation is a hedonistic one: the reason all of our desires point towards happiness is that happiness is the only thing that is intrinsically prudentially valuable." So Silverstein has attempted to turn the experience machine objection to hedonism on its head by arguing that the widespread preference for reality over a life in an experience machine (along with all other preferences) actually provides evidence in support of Prudential Hedonism!

Unfortunately for Silverstein, it is not clear that many philosophers would agree that his 'all desires are created by happiness' view better explains reality mattering intrinsically to the vast majority of people than reality having intrinsic prudential value does. The vast majority of philosophers would probably accept that the circumstances surrounding our previous experiences of happiness will have affected what currently matters to us, but they will not accept that those experiences of happiness are always the *main* cause of what currently matters to us. They could argue that reality matters to us a small amount because it avoids the pain associated with deception, but also that it matters to us a lot because it makes experiences more meaningful (regardless of how much happiness they bring).

This response would leave Silverstein in a difficult position. If strong Motivational Hedonism is true (i.e. if all of our desires are completely governed by the pursuit of pleasure and avoidance of pain), then Silverstein could ignore this response. However, Silverstein (2000, p. 293, n. 42) does not claim that strong Motivational Hedonism is true, so he cannot pursue this line

of argument. And even if he did, there are good philosophical and introspective reasons to doubt strong Motivational Hedonism (as discussed in Chapter 1), at least until advanced cognitive science decides this issue for us (Silverstein 2000, p. 294, n. 47). This leaves Silverstein the problem of trying to motivate why the preference for reality in particular is likely to be largely or wholly governed by our underlying preference for happiness. Considering the prevalent pluralist beliefs most current philosophers have about prudential value (particularly about real achievements and meaning in life), Silverstein has not done enough to convince us that reality mainly matters intrinsically to us because of our past experiences of reality-related happiness.

Roger Crisp (2006a) and Sharon Hewitt (2009) also argue that IP5 is false because reality mattering intrinsically to the vast majority reasonable people is better explained by evolutionary and psychological explanations. Both Crisp and Hewitt discuss the paradox of hedonism and then use specific examples to argue that many of our judgments about which goods matter intrinsically to us could have developed because they aided our ancestors' procreative fitness and provided us with pleasure throughout our personal development. Crisp (2006a) mainly discusses how our preference to accomplish real achievements could have developed into a powerful intrinsic desire under the guidance of selfish genes and a pleasure-seeking brain. Hewitt (2009) focuses on how our preference to establish real interpersonal relationships could have developed in a similar way. Since both accomplishing real achievements and establishing real interpersonal relationships require us to live in reality, and on the assumption that evolutionary and psychological mechanisms best explain why accomplishing real achievements and establishing real interpersonal relationships matter to us, Crisp (2006a) and Hewitt (2009) both conclude that evolutionary and psychological mechanisms best explain why reality matters intrinsically to us.

Crisp's (2006a) and Hewitt's (2009) arguments will encounter the same initial problem as Silverstein's; other philosophers are unlikely to be convinced that the evolutionary and psychological mechanisms explain why reality matters intrinsically to us *better* than reality having intrinsic prudential value does (e.g. Fletcher 2007). Again, without the support that the truth of strong Motivational Hedonism would offer, it seems like proponents of the experience machine objection to hedonism can reply that Crisp's (2006a) and Hewitt's (2009) explanations are plausible, but that connection with reality being valuable for its own sake is *more* plausible and, most importantly, that this is the reason it matters intrinsically to them and why they prefer reality over a life in an experience machine.

So even when taken together, Silverstein's (2000), Crisp's (2006a), and Hewitt's (2009) denials of IP5 are unlikely to convince many non-hedonists that reality mattering intrinsically is mainly caused by it having intrinsic prudential value. This is especially the case because non-Hedonists have both introspective and philosophical evidence (from the experience machine thought experiment amongst other sources) that connection to reality has prudential value over and above any pleasure or pain that it might lead to. Indeed, it seems like defenders of Prudential Hedonism should put this particular argument on the shelf until cognitive science can provide more detailed information on how our judgments and preferences are created.

3.2.2 Responses to IP3

A slightly more promising group of responses to the experience machine objection to hedonism deny IP3 (repeated below for convenience). These denials all identify features of the experience machine thought experiment that might elicit responses to it that are biased or otherwise corrupted by irrelevant factors. The arguments all deny IP3 on the basis that the best

explanation for why the vast majority of reasonable people report preferring reality over a life in an experience machine is that at least one feature of the thought experiment (that is irrelevant to the purpose of the thought experiment or to assessing well-being) is what really matters to the vast majority of reasonable people.

- IP2. When instructed to ignore their responsibilities to others, the vast majority of reasonable people report preferring reality over a life in an experience machine. (Empirical claim)
- IP3. The best explanation for IP2 is that reality matters intrinsically to the vast majority of reasonable people

Most proponents of the experience machine objection to hedonism understand the experience machine thought experiment to isolate a prudential value comparison between reality and how our experiences feel to us on the inside. But the ability of exotic philosophical thought experiments, like the experience machine, to isolate what they intend to is derided by some psychologists and behavioural economists who research how judgements and preferences are formed. Cass Sunstein had this to say:

I believe that some philosophical analysis, based on exotic moral dilemmas, is inadvertently and even comically replicating the early work of Kahneman and Tversky by uncovering situations in which intuitions, normally quite sensible, turn out to misfire. The irony is that where Kahneman and Tversky meant to devise cases that would demonstrate the misfiring, some philosophers develop exotic cases with the thought that the intuitions are likely to be reliable and should form the building blocks for sound moral judgments. An understanding of the operation of heuristics offers reason to doubt the reliability of those

intuitions, even when they are very firm (cf. the emphasis on moral learning from real-world situations in Churchland 1996). (Sunstein 2005, p. 541)

Bronsteen, Buccafusco, and Masur (2010, p. 1609) worry that our intuitive reactions to the experience machine thought experiment might be based on more than just an isolated prudential value comparison between reality and how our experiences feel to us on the inside: “When we ask whether someone attached to the machine has greater welfare, we must look behind whatever visceral aversion to the machine we might have and assess (i) whether that aversion relates to welfare and (ii) whether the aversion springs from rejecting the rules of the hypothetical example.” Bronsteen and colleagues’ worry is shared to varying degrees by many philosophers.

Indeed, many potential causes of negative visceral sensations, especially fear, have been identified in the experience machine thought experiment. D. W. Haslett (1990) expressed concern that the experience machine thought experiment is so unrealistic that it demands too much of the readers’ imaginations and that the judgements it elicits should be expected to be unreliable. Wayne Sumner (1996, p. 95) asks: “How do we know that the technology is foolproof? What happens if there is a power failure?” Bronsteen, Buccafusco, and Masur note that readers of the experience machine thought experiment “might not be convinced the machine will actually work; or they might fear that while on the machine, they will be vulnerable to harm from those in the real world” (2010, p. 1609). Torbjörn Tännsjö agrees, stating that some people might not choose an experience machine life “because of an (unreasonable) fear that those in charge of the machine... would take advantage of them in some nasty way” (2007, p. 93). Goldsworthy (1992, p. 18) emphasises these points, claiming that a reasonable person might choose

not to plug in because of fear of “catastrophic, unimaginably horrible consequences of malfunction or abuse”. Hewitt adds that anyone connecting to an experience machine “must trust those outside of the machine to look out for his interests as well as he could himself if he were living in contact with the external world” (2009, p. 338). Barber (2011, p. 267) makes an analogy with some people’s reluctance to fly: “Some refuse to fly but will drive even if they know this is far more dangerous, and boarding an airplane is nothing compared to volunteering oneself up to a different—and delusional—plane of reality.”

Adam Kolber (1994, pp. 13–14) notes that the fears of machine underperformance or failure are exacerbated by the troubling irrevocability of the experience machine and our general fear of the unfamiliar: “We are hardly comfortable enough with our own world to risk life under totally foreign circumstances.” (1994, p. 13). Both Weijers (forthcoming-a) and Mendola (2006) also emphasise how worries about the machine feed into a general fear of the unknown, with Mendola (2006, p. 450) claiming that it is the “unfamiliar gadgetry which invokes our fear of the unfamiliar”. These fears lead Crisp to dismiss the question about whether people would or should choose to connect to an experience machine because their choices are likely to be affected by “differing attitudes to risk” (2006a, p. 635). All of these worries about the machine are highly relevant to our well-being, but they should not enter into a direct comparison between reality and how our experiences feel to us on the inside. Therefore, if these worries affect a significant number of people’s preference in the experience machine thought experiment, IP3 starts to come under pressure.

Only a few of the authors mentioned argue that these irrelevant fears might amount to a refutation of IP3. This is because, although these fears plausibly affect people’s preferences in the experience machine thought

experiment, they don't obviously provide a better explanation for them than reality mattering intrinsically to us. Even Silverstein (2000), who argues against the experience machine objection to hedonism, argues that this response to IP3 will not work. Silverstein claims that with "a bit of mental dexterity" (i.e. tweaking the scenario and having faith in the outcomes stipulated in the scenario) he thinks that we can allay any doubts about our intuitive experience machine-related fears (2000, p. 284). And without these fears, Silverstein believes that "we remain unwilling to accept a lifetime on the experience machine" (2000, p. 285).³⁹ For this reason, many of the philosophers who argue that these fears might provide a good explanation for the widespread preference for reality over a life in an experience machine also provide a further objection to IP3.

That further objection to IP3 is that people can have preferences for things because they promote several different kinds of value (not just prudential value). Silverstein (2000, p. 290) puts it like this: "One way to lessen the force of the experience machine intuitions is to demonstrate that they are really about something other than well-being." This objection argues that people might report preferring reality over a life in an experience machine because they prefer to promote aesthetic, moral, or other non-prudential values, which they could not achieve while connected to an experience machine. If the widespread preference for reality over an experience machine life is best explained by the widespread desire to promote moral values, then IP3 is false and the experience machine thought experiment does a very bad job of isolating a prudential value comparison between reality and how our experiences feel to us on the inside. If IP3 is false for this reason, then the experience machine thought experiment doesn't tell us much at all about

³⁹ This claim is empirically tested in Chapter 5.

prudential value and, therefore, it can't provide the basis for a reasonable objection to Prudential Hedonism.

Silverstein (2000, p. 291) argues that the low aesthetic value of an experience machine life might be heavily impacting people's preferences in the experience machine thought experiment: "When we entertain Nozick's thought experiment, we find life on the machine unattractive: it is aesthetically displeasing." Most philosophers who argue for non-prudential values affecting our preferences in the experience machine thought experiment, however, focus on moral values. Kawall gives the example of a soldier who throws himself on a grenade as evidence that people do sacrifice their own well-being for moral reasons (1999, pp. 385–386). Mendola notes that "things other than our own well-being... matter to us. For instance, our lives have effects on other people which are quite significant, and which we care about" (2006, p. 450). Kolber suggests that some people might not report preferring a life in the experience machine because it's self-indulgent and self-indulgence is considered immoral (1994, p. 14). Bronsteen, Buccafusco, and Masur note that readers of the experience machine thought experiment "might not want to forego the opportunity to use their lives to improve the lives of others" (2010, p. 1609).

Nozick (1974, p. 43) anticipated this objection and attempted to protect the experience machine thought experiment against it by stating: "Others can also plug in to have the experiences they want, so there's no need to stay unplugged to serve them. (Ignore problems such as who will service the machines if everyone plugs in.)" But Feldman provides two effective examples of why this approach might not work:

A person might still worry about fulfilling his moral obligations even if he were convinced that others would be plugging in. For example, consider someone who solemnly promised his mother

that he would rescue her if her Experience Machine should happen to malfunction. Believing that others would be in Experience Machines (or would have the option of plugging in) would not relieve him of his feeling of obligation to keep out of the machine so as to be available for rescue operations, should they be necessary. And even more obviously, suppose he knows that others have been given the opportunity to plug in but have chosen to remain unplugged. His feeling of obligation to them would be unaffected. (Feldman, 2011, n. 17)

Again Silverstein (2000, p. 291) argues that these concerns can be stipulated away, but it is unlikely that they can be *completely* stipulated away. Based on how judgements about thought experiments are actually formed, both Hewitt (2009) and Weijers (forthcoming-a) have argued that simply stipulating that readers needn't worry about something that they value greatly doesn't always work.⁴⁰ Therefore, there seems to be good reason to think that some people might prefer reality over an experience machine life because a non-prudential value, probably a morally-related one, matters a lot to them. Having said this, without further evidence about how many people report preferring reality over an experience machine life because a non-prudential value really matters to them, it seems like reality mattering intrinsically to the vast majority of people might still be the best explanation for the widespread preference for reality.

When all of these potential causes of people choosing reality over the experience machine are taken together, however, they begin to pose a potential threat to IP3. The difficulty is in assessing how much effect these causes have individually and cumulatively. For this reason, the philosophers

⁴⁰ Explained in detail in Chapter 3.

who argue that these alternate causes might provide the best explanation for the widespread preference for reality over a life in an experience machine also try to create experience machine scenarios that isolate and eliminate them to see if IP2 still holds.

3.2.3 Responses to IP2

Another group of responses to the experience machine objection to hedonism deny IP2 (reproduced below for convenience). The contemporary denials of IP2 are often subtle and indirect, but at least one is incredibly straightforward.

IP2. When instructed to ignore their responsibilities to others, the vast majority of reasonable people report preferring reality over a life in an experience machine. (Empirical claim)

Torbjörn Tännsjö takes a direct approach by claiming that he would not choose reality over an experience machine (2007, p. 95). He also argues that lots of people take drugs, which is basically the pharmacological equivalent of connecting to an experience machine (Tännsjö 1998, p. 112). Combined with his worry about people's fear of abuse while connected to an experience machine, this evidence forms the basis for his claim that "it is far from clear that ... the claim that we would not plug in, is *true*" (Tännsjö 2007, p. 93, his italics). This approach will do very little to convince anyone who has presented the experience machine thought experiment to students (or any other group of people) that IP2 is true because they have had first-hand experience of the vast majority of people reporting that they would prefer reality over a life in the experience machine.⁴¹

⁴¹ The one exception here seems to be Barber (2011, p. 263, n. 7), who claims that his undergraduate philosophy students were fairly evenly split ("52% non-enterers to 48%

A much more promising approach to denying IP2 is to construct an alternate version of the experience machine thought experiment that does *not* elicit a widespread preference for reality over a life in an experience machine while remaining essentially the same in all relevant ways. Based on the worry that the potential causes of people's preference for reality over an experience machine life, such as irrational fear, are irrelevant to assessing prudential value, several philosophers have developed new experience machine scenarios. These new scenarios attempt to eliminate from consideration all of the factors that are irrelevant to an isolated prudential value comparison between reality and how our experiences feel to us on the inside. If other versions of the experience machine thought experiment can equally or better isolate this prudential value comparison, and people's judgments about them significantly diverge, then either IP2 or IP3 might be false. Kolber explains the strategy behind this approach:

[A]ll thought experiments that consider the same issue as the [experience machine thought experiment] must yield the same results... [Otherwise] Nozick's argument has failed. [This is because] it seems unlikely that a reflective, unbiased person could give different answers to two versions of essentially the same question. (Kolber 1994, p. 13)

Kolber (1994, p. 15), De Brigard (2010, pp. 47–49), and Weijers (forthcoming-a) have all created new experience machine scenarios that attempt to reduce interference from irrelevant factors. What is most notable about all of these new scenarios is that they are designed to minimise the impact of unfamiliarity with, and fear of, experience machines by framing being

enterers"), although he implies that his data is not as credible as data from more formal surveys.

connected to a machine as the status quo.⁴² The scenarios are discussed in much more detail in the next chapter, but here is an excerpt from one of De Brigard's reversed scenarios:

"I am afraid I have some disturbing news to communicate to you" says Mr. Smith. "There has been a terrible mistake. Your brain has been plugged by error in to an experience machine created by super duper neurophysiologists. All the unpleasantness you may have felt during your life is just an experiential preface conducive toward a greater pleasure... we'd like to give you a choice: you can either remain connected to this machine (and we'll remove the memories of this conversation taking place) or you can go back to your real life. By the way, you may want to know that your real life is not at all as your simulated life. In reality you are a prisoner in a maximum security prison in West Virginia." What would you choose? (De Brigard 2010, p. 4)

Kolber's (1994, p. 15), De Brigard's (2010, pp. 47–49), and Weijers's (forthcoming-a) scenarios all appear to isolate a prudential value comparison between reality and how our experiences feel to us on the inside about as well as Nozick's original scenario. However, the judgements they tend to elicit seem to be quite different.⁴³ In all of these reversed scenarios, especially De Brigard's Negative scenario, it seems that most people might prefer a life connected to an experience machine over a life in reality. This is a questionable empirical claim, but if it's true then either IP2 is false because the vast majority of people don't actually prefer reality (at least in all cases) or IP3

⁴² A move similar to the reversal test proposed by Bostrom and Ord (2006).

⁴³ Empirical evidence surrounding this claim is presented in Chapters 3 and 5.

is highly questionable because people's preferences in experience machine cases might be explained better by the status quo (what is most familiar to us) than by reality mattering intrinsically. However, exactly how reasonable people would decide in these reversed cases is not the only potential problem for Kolber, De Brigard, and Weijers; their scenarios might have introduced more biases and other confounding factors than were present in Nozick's original scenario.

Nevertheless, for anyone who endorses the experience machine objection to hedonism and found themselves preferring a life connected to an experience machine over a life in reality in one of these new scenarios, the onus should be on them to point out the confounding feature of the new scenario(s). This apportionment of the benefit of the doubt seems fair because proponents of the experience machine objection to hedonism have already been using a suspect thought experiment (Nozick's experience machine scenario) as evidence in an argument, making it inconsistent for them to deny the same practice for others.

A safer response for proponents of the experience machine objection to hedonism is the direct approach taken by Tännsjö above; to simply deny that they would choose a life connected to an experience machine over a life in reality in the new scenarios. While this approach is safer, it is not enough to prevent this kind of denial of IP2 because these authors have begun to test their thought experiments and it is no longer clear what the vast majority of people believe about experience machine scenarios. Having said this, questions have been raised about the validity of some of these specific tests and of testing the experience machine in general (e.g. Smith 2011). This means that attempts to deny IP2 by creating a new scenario that produces a different result from Nozick's scenario requires a sound empirical result before they stand a good chance of convincing stubborn proponents of the experience

machine objection to hedonism that IP2 or IP3 is false. For these reasons, the merits of these tests are discussed in the subsequent chapters of this section.

4. Conclusion

This chapter has introduced the experience machine objection to hedonism, explained how it is widely thought to refute Prudential Hedonism, and presented deductive and inductive versions of the argument supporting it. These two versions were used to create a taxonomy of responses to the experience machine objection to hedonism that is intended to be exhaustive of all the paper-length attempts to refute the experience machine objection to hedonism from the last 15 years and indicative of all of the types of critical responses to it.

The discussion of the responses to the deductive version of the experience machine objection to hedonism made it clear that it was implausible. Most importantly, the deductive version ignored the fairly obvious fact that reality might matter intrinsically to us for reasons other than reality being intrinsically prudentially valuable.

This chapter has also shown that the abductive version of the experience machine objection to hedonism might eventually be seen as unlikely to be true because of the pressure on IP3. That pressure is coming from the numerous possible causes of people's preference for reality over a life in an experience machine. Furthermore, when this pressure on IP3 is combined with the threat to IP2 coming from alternate experience machine scenarios, serious doubts about the experience machine objection to hedonism begin to arise. Most prominent of these threats to IP2 and IP3 is the worry that people's preferences for reality over a life in the experience machine in Nozick's scenario are heavily affected by the framing of the status quo, as demonstrated by the reversed experience machine scenarios of Kolber (1994),

De Brigard (2010), and Weijers (forthcoming-a). To further investigate the fate of the abductive version of the experience machine objection to hedonism, the potential effects of biases and other confounding factors on our judgments about experience machine scenarios are investigated in the next chapter. Further empirical investigation later in this section will also give us good reason to think that the abductive version of the experience machine objection to hedonism is unlikely to be true.

Chapter 3

The Case for Status Quo Bias in Experience Machine Scenarios

Chapter Summary

This chapter considers whether status quo bias—the irrational preference for things to remain the same—affects all of the main experience machine scenarios discussed by philosophers. First, the role of intuitions in making judgments about thought experiments is discussed. Second, the existing experience machine scenarios are analysed to investigate the role of status quo bias in our judgements about experience machine scenarios. This analysis involves an in-depth examination of the support De Brigard's (2010) empirical results offer to the case for status quo bias affecting judgments about experience machine scenarios. It is argued that while the commonplace choice to remain in reality when offered a life in the experience machine can credibly be partially explained by status quo bias, it is not yet obvious that the status quo bias is the main or even a major cause of this choice.

1. Introduction⁴⁴

This chapter considers whether status quo bias—the irrational preference for things to remain the same—affects all of the main experience machine scenarios discussed by philosophers. Thought experiments have long been the friend of philosophers, allowing many problems to be addressed without requiring the 'muddying of knees' often entailed by field work. More recently, however, warnings have been issued about thought experiments

⁴⁴ This chapter is the basis for Weijers (forthcoming-a) and, to a lesser extent, Weijers (2011c).

propensity to mislead for a variety of reasons (e.g. Dennett 1980; Hofstadter & Dennett 1981; Unger 1996; Woodward & Allman 2007). Experimental philosophy has also provided many reasons to doubt the philosophical usefulness of many specific thought experiments over the last ten years (Knobe 2007; Nadelhoffer & Nahmias 2007). This chapter continues these emerging traditions by investigating the role of bias (and especially status quo bias) in our judgments about experience machine thought experiments.

First, the role of intuitions in making judgments about thought experiments is discussed. Second, the existing experience machine thought experiments are analysed to investigate the role of status quo bias in our judgements about experience machine scenarios. This analysis involves an in-depth examination of the support De Brigard's (2010) empirical results offer to the case for status quo bias affecting experience machine scenarios. It is argued that while the commonplace choice to remain in reality when offered a life in the experience machine can credibly be partially explained by status quo bias, it is not yet obvious that the status quo bias is the main or even a major cause of this choice.

2. Intuitions and Intuition Pumps

Dennett coined the phrase 'intuition pump' to describe a thought experiment that (by design or not) elicits a response with a strong intuitive component (Dennett 1980). Typically, judgments about thought experiments will have a deliberative component and an intuitive component (Woodward & Allman 2007). When thought experiments create misleading responses in philosophers and other reasonable people who generally pride themselves on their rational thinking, the intuitive component is more likely to be blamed. Indeed, Bostrom and Ord note that our overall judgments can be "crucially and unavoidably" influenced by the psychological biases that our intuitive

cognition is prone to (Bostrom & Ord 2006, p. 657). Systematic biases also occur in deliberative thinking; however, the combination of two factors characteristic of intuitive judgments makes them more likely to mislead. First, it is very difficult to know when a judgment is misguided and, second, it is even more challenging to ascertain whether the intuitive component of a judgment is tracking relevant, or merely distracting, information. The emerging fields of behavioural economics, moral psychology, and experimental philosophy have been producing results that are helping us to understand the extent to which intuitive cognition impacts our judgments and what features of certain thought experiments our intuitive cognition is likely to track. However, this task is made very difficult by the nature of intuitive cognition.

Woodward and Allman provide a neurobiological account of intuitive cognition and contrast it with the other kind of cognition used in judgments about thought experiments—deliberative cognition (Woodward & Allman 2007). They describe an intuition as the visceral sensation that results from a very fast, unconscious and probabilistic processing of many variables in parallel—a definition that is widely accepted in the cognitive sciences (Lieberman 2000; Myers 2004; Woodward & Allman 2007, p. 13). In contrast, deliberative thought is a much slower cognitive process, which consciously uses inductive and deductive reasoning on very limited numbers of variables at a time (Woodward & Allman 2007, p. 13; Bruner 1960). While both modes of cognition have their strengths and weaknesses, and both are susceptible to systematic biases, intuitive judgments (judgments that are heavily influenced by intuitive cognition) have the significant disadvantage of us not being able to know if they have been influenced by various psychological biases

(Gilovich, Griffin, & Kahneman 2002).⁴⁵ This disadvantage arises because of the process by which intuitions are created.

When novel stimuli are encountered, the brain runs probabilistic inference simulations based on the matches between all of its current stimuli and past experiences (Woodward & Allman 2007). Part of the simulation process involves the reward centre, which produces the message (the visceral sensation or feeling) that we become consciously aware of (Craig 2004; Critchley *et al.* 2004). We are effectively always performing unconscious pattern recognition by monitoring our current environment and comparing it to our archive of experiences. The result of this processing is a range of predictions about what might happen next. These probabilistically determined predictions are then evaluated using a similar process (comparing them to the value of similar previous actual events) and an overall evaluation is generated. Inferring the value of predicted outcomes in this way can allow for psychological biases to systematically influence our intuitions and, thereby, our overall judgments. At most risk from this process are judgments that are heavily influenced by intuitive cognition—intuitive judgements.

How does this process affect our judgments about thought experiments? Systematic biases can sneak in during this kind of pattern-recognition process because the features of the current thought experiment (or past experiences to which they are being compared) that have the most weight in the probabilistic processing might not be the features deemed morally relevant by the readers of the thought experiment. It might be natural to assume that a seasoned philosopher could apply their well-honed rational mind to their intuitive judgements and eliminate any biases before coming to a final judgment.

⁴⁵ People might also make biased deliberative judgments without realising, but philosophers and anyone else trained in logic would be embarrassed to be caught out making such mistakes.

However, even when highly educated (and presumably rational) people are made fully aware of these biases, they still make judgments that bear all of the hallmarks of being adversely affected by them. Consider the self-serving bias: the phenomenon that explains why practically everyone (even social psychologists who are well aware of this bias) view themselves as more moral and better at driving than the average person (Myers 2004, p. 95; van Lange, Taris & Vonk 1997). About half of these presumably intelligent people are fundamentally mistaken about some of their own characteristics because of a bias they should have corrected for.

To further complicate the matter, it is difficult to know whether or not a judgment has a strong intuitive component and if that component has been distorted by any biases. Because the causes of our intuitions are processed subconsciously, they are not open to introspection (Lieberman 2000; Woodward & Allman 2007). That is, we might experience a visceral reaction when we come into contact with a new stimulus, such as a thought experiment, but we only really know what the sensation feels like; we know little, if anything at all, about what caused it and why. The best method for establishing the likely causes of an intuition is by reconstruction. By carefully considering all of the possible environmental cues and how they might match a subject's past experiences, one (or more) of those cues may stand out as an obvious candidate for explaining how the intuition was initially constructed. This is the process usually followed (although rarely explicated) by philosophers when they interpret what evidence a particular thought experiment might provide. Using this process, we can reasonably assume that the uneasy feeling we suddenly notice while imagining a thought experiment is probably caused by our imagining of the thought experiment (unless we have just eaten at that dodgy diner around the corner). What is much less

reasonable to assume, however, is which aspects of the thought experiment are causing the intuitive judgment.

Philosophers often assume they have isolated the relevant moral factors in the construction of their thought experiments (Sunstein 2005), but this assumption shows either a misunderstanding of the internal workings of intuitive cognition or a lack of awareness that our intuitive cognition can affect our overall judgments. The pared-down, and often unrealistic, thought experiments used by philosophers usually stipulate only very few aspects of the situation under assessment. Indeed, the decision to create thought experiments in this way is often justified by claiming that less background information prevents irrelevant factors from being considered. But, this is not necessarily the case (Hewitt 2009). Recall that intuitive cognition operates by comparing the new stimuli with existing experiences to try to predict what might happen next. Minimalistic thought experiments are matched to the closest real experiences. Those real experiences will not be minimalistic, however; they will contain many features absent from the thought experiment, including emotionally salient (and therefore powerful) ones. Subconscious simulations are then run based on the real experiences that the thought experiment was most closely matched with. Since this process takes into account features that were relevant to the real experiences, but not included in thought experiment, the resulting intuitions will be partly based on irrelevant information.

Making matters even worse, however, is the fact that most philosophical thought experiments stipulate features that are so unrealistic that we have not experienced anything like them—indeed we are likely to have experienced the very opposite of them. When these clashes occur, our intuitions are likely to be based on information that is not just irrelevant, but contrary to the point

of the experiment itself. This misinformed intuition then corrupts the overall judgment that we reach about such thought experiments.

For example, a thought experiment that was designed to fairly evaluate a particular kind of life might describe that life as being computer-generated to make the scenario more plausible. The stipulation that the life in question would be computer-generated is supposed to be irrelevant, but it is actually quite likely to affect our intuitive judgment of that life. Many of our experiences with complex computerized machinery have involved disappointing underperformance and catastrophic crashing. (Is there an academic who has not lost important work due to her computer crashing?) So, when our intuitive cognition is matching the stipulations of the thought experiment with our past experiences, some of our myriad experiences of computer failure might be the most similar to the stipulations of the thought experiment *as a whole* despite contradicting some individual stipulations. So our intuition about choosing the life that just happens to be computer-generated could be influenced by the misapplication of our otherwise rational fear of computer failure.

It is these features of intuitive cognition that enable structural biases to affect our judgments about thought experiments in ways that they do not affect our deliberative cognition. With this understanding of how judgments about thought experiments are formed, we now return to the experience machine objection to hedonism to assess the likelihood that reasonable people's judgments about it are affected by bias or other irrelevant factors that we might not be aware of.

3. Intuitions and the Experience Machine Objection to Hedonism

As discussed in the previous chapter, the vast majority of people who read Nozick's (1974) experience machine scenario think that they would choose to

remain in reality. Based on this widespread judgment, the experience machine objection to hedonism infers that there is more to the good life for the one living it than how our experiences feel to us on the inside. Furthermore, because the lack of a direct connection with reality in an experience machine life is assumed to be the only relevant difference between the two options in Nozick's scenario, most people also infer that living in reality must make our lives go better for us regardless of whether it leads to increased enjoyment. It deserves to be emphasised that the main justification for both of these inferences is the widespread judgement that connecting to an experience machine in Nozick's scenario is a bad idea.

Just how widespread is the judgment that connecting to an experience machine in Nozick's scenario would be worse for us than continuing our normal life in reality? Most introductory ethics lecturers know that the vast majority of students presented with Nozick's scenario claim not to want to connect to the experience machine. Similarly, most philosophers know that none, or nearly none, of their colleagues would choose to connect to an experience machine. Furthermore, initial empirical data from the International Wellbeing Study indicates that these judgements may generalise to non-philosophers. In the Further Assessment Study component for the first intake of the International Wellbeing Study, only 12% (19/156) of the participants⁴⁶ chose to connect to an experience machine when presented with a simplified version of Nozick's experience machine scenario. The International Wellbeing Study Further Assessment Study experience machine scenario (IWSFAS scenario) reads as follows:

⁴⁶ The sample for the Further Assessment Study component for the first intake of the International Wellbeing Study (www.wellbeingstudy.com, Jarden *et al.*) is far from representative. It could roughly be described as a self-selecting group of English speakers from several countries around the world who are interested in well-being and like filling out questionnaires about well-being.

Imagine that scientists figured out a way to stimulate a person's brain so that they experience constant and permanent pleasure every moment of their life. It's perfectly safe, no chance of malfunction, and not harmful to a person's health. Would you choose to be plugged into such a machine?

It is likely that this widespread judgement about Nozick's scenario has a strong intuitive component. This can be inferred because, upon first exposure, it is not always obvious why the thought of connecting to an experience machine produces the negative feeling that it usually does.⁴⁷ Regardless of the actual causes of intuitions, when the vast majority of philosophers share an intuition, it can be used as a premise in philosophical arguments. Good philosophers do not consider widely agreed upon intuitions to be *unquestionable* premises, however. David Sobel, for example, uses the widespread intuition that a real life is better than a life connected to an experience machine as a premise in his argument against Quantitative Hedonism, while acknowledging that discrediting the intuition would refute his argument (2002, p. 244). Sobel explains that the credibility of intuitions elicited from contemplation of thought experiments can be undermined by "telling a convincing story about the genesis of such intuitions that would explain why we have them while revealing them to be misleading" (2002, p. 244). Apparently unbeknownst to Sobel, a version of this story had already been told by Adam Kolber (1994), who identified the status quo bias as the

⁴⁷ Since deliberative judgments are open to introspection, and intuitive judgments are not, judgments that appear to be formed *because of* a reason are more likely to have a large deliberative component and judgments that appear to be formed without any immediately obvious reason are more likely to have a large intuitive component.

main cause of the widespread negative intuition about connecting to an experience machine in Nozick's scenario.

4. Kolber and Status Quo Bias

In Kolber's (1994) story about our intuitive responses to Nozick's scenario, status quo bias is the lead villain. For now, status quo bias is best defined as an inappropriate preference for things to remain the same (Bostrom & Ord 2006), but a more detailed discussion of this complex psychological phenomenon follows below. Kolber argues that our intuitive judgments about thought experiments should only be considered good evidence for an argument if other thought experiments considering the same issue elicit similar intuitive judgments (1994, p. 13). Kolber presents a thought experiment that reverses one aspect of Nozick's scenario, while keeping the other aspects consistent and focussing on the same issue:

[I]magine that you are currently hooked up to an experience machine. Nothing *seems* any different under these circumstances than it seems in the real world. The paper you are now reading wouldn't exist in the traditional sense but would only seem to. It is, in fact, a possibility that you *are* currently hooked up to such a machine since there is nothing in the world (or module) that could prove otherwise to you. Suppose you had reason to believe that you were hooked up to such a machine. Would you then care any less about your parents and friends (that is, the people you call your "parents" and "friends")? In deciding whether or not to get off the experience machine, you might want to know what life would be like in the real world. This is fair enough. If Nozick lets you know what life is like on the experience machine, we can likewise say some things about life

in the “real” world. We can imagine lots of possibilities, but this much is sure: Your life off of the experience machine will contain worse mental states than your life on it. The easiest way to consider this is to imagine life in the real world as less enjoyable (in an experiential sense) than your life currently is. In addition, if you get off the machine, the totality of subjective mental states will be worse. We can also stipulate that everyone else’s mental states will either stay the same or get worse by your getting off the machine. (Kolber 1994, p. 15, *his italics*)

Kolber asserts that “more people would want to stay on the machine [in this reversed scenario] than would agree to connect in [Nozick’s scenario]” (1994, p. 15). Kolber goes on to argue that the different intuitive judgments elicited by the two scenarios reveal that a bias is likely to be affecting our judgements about what matters to us in experience machine scenarios. Kolber mentions several potential sources of bias, but concludes that continuing the status quo seems to be what really matters to us when we consider experience machine scenarios (1994, pp. 15–16). Status quo bias is implicated in this result because the only difference between the two scenarios seems to be whether reality or the experience machine is framed as the status quo. Since the status quo is arbitrary and since the experience of the status quo could be maintained perfectly by an experience machine, Kolber advises against using the experience machine to investigate what matters or what should matter to us regarding well-being (1994, p. 16).

5. Weijers and Status Quo Bias

Building on Kolber’s idea of creating a reversed experience machine scenario to examine the role of status quo bias in our judgments about experience

machines, Weijers (forthcoming-a) presents the trip to reality thought experiment. The Trip to Reality holds constant the realness of experiences inside and outside of the machine, while changing a few other factors that are usually considered irrelevant by proponents of the experience machine objection to hedonism.

Imagine that you leave your family for a weekend to attend a conference on the experience machine thought experiment. While you are there, someone informs you that you are actually in an experience machine. (Which means that you haven't *really* left your family behind to go to this conference because all of your experiences have been machine-generated). She offers you a red and a blue pill. She explains that taking the blue pill will take you to reality and taking the red pill has the dual effects of bringing you back from reality and completely wiping any memories of having being in reality or being offered funny pills. Being a curious philosopher you swallow the blue pill. It turns out that reality is quite different to the world you have been experiencing inside the machine. Some of the most noticeable differences are that your experiences are more mundane and less enjoyable than before. You also discover that nearly all of your friends and family are either in experience machines or do not exist in reality! Your father is there, so you spend time with him. But, a few conversations reveal that he is not really the person you know as 'Dad'. It is time to make the choice. Will you take the red pill so that you can go back to your previous life (with no idea that it is not in fact real)? Or will you throw the red pill away and try to make the best life you can in the

more real, but less enjoyable, surrounds of reality? (Weijers forthcoming-a)

In the Trip to Reality thought experiment, the thought of getting into an experience machine does not elicit the same intuitive judgment that the equivalent act does in Nozick's scenario. In my experience of presenting the two scenarios, dramatically more people choose a life in an experience machine when considering the Trip to Reality thought experiment than when considering Nozick's experience machine scenario. If my experience is representative, then status quo bias provides a plausible and fairly elegant explanation for this difference.

6. Status Quo Bias

A group of overlapping psychological heuristics, best referred to as status quo bias, might provide the best main causal explanation for both the intuitive judgment that a life in Nozick's experience machine is worse than real life and the intuitive judgment that reality is worse than a life in an experience machine in the Trip to Reality thought experiment. But a better understanding of status quo bias is required to assess this claim. Status quo bias is closely linked to a family of psychological heuristics, including loss aversion (valuing losses more than equivalent gains in uncertain circumstances) and the endowment effect (overvaluing what we have and know). Note that status quo bias and status quo heuristic will be used differently here. Status quo heuristic will be taken to mean the cognitive mechanism that influences our choices towards the status quo, while the status quo bias will refer to instances of the status quo heuristic that are inappropriate for some reason—a misfiring of the status quo heuristic.

The overlapping psychological heuristics that make up status quo heuristic are importantly linked by our valuing prospective gains only about half as much as we value avoiding equivalent prospective losses of things we already have or know—the status quo.⁴⁸ Simply put, status quo bias is an inappropriate preference to keep things the way they are.⁴⁹ Such preferences are often considered inappropriate or irrational because they assign value to certain things over and above any utility value they might have (in the broadest possible sense).⁵⁰

The endowment effect—an aspect of the status quo heuristic—has been used to explain why only about 10% of undergraduate students, who were rewarded with either a mug or a chocolate bar for filling out a survey, took up

⁴⁸ See for example Gilbert (2006), Gilovich, Griffin, and Kahneman (2002), Kahneman and Tversky (2000), Kahneman, Knetsch, and Thaler (1991), Samuelson and Zeckhauser (1988), Tversky and Kahneman (1991).

⁴⁹ Psychologists' definitions of status quo bias are less useful because they are often long-winded and technical. The long-windedness is likely a result of the uncertainty about what related psychological heuristics and biases should be considered a part of status quo bias. Aspects of prospect theory, loss aversion, the endowment effect, and other psychological phenomena have all been proposed as possible constituents or causes of status quo bias. See for example Druckman (2001), Gilbert (2006), Gilovich, Griffin and Kahneman (2002), Kahneman, Knetsch and Thaler (1991), Kahneman and Tversky (2000), Samuelson and Zeckhauser (1988), and Tversky and Kahneman (1991). An alternate definition for status quo bias to the one I use here has been published in the philosophical literature: "people come to prefer what they are likely rather than unlikely to get" (Dorsey 2010, p. 535). While this definition is fairly compatible with the one used here, it seems to be unduly weighted in favour of future concerns.

⁵⁰ Note that even items that are monetarily worthless can have immense utility value. Take, for instance, your child's first pair of baby shoes. Their resale value is approximately nil, but every time you see them, you recall pleasant memories that bring you joy. An irrational preference for these shoes would be to prefer them to something else that would give you more joy (and other types of utility), all other things being equal.

the cost-free opportunity to swap their reward for the other type⁵¹ (Knetsch 1989; Knetsch & Sinden 1984). An irrational preference for the status quo has been posited as the explanation for Hartman, Doane and Woo's field study of power consumers (Hartman, Doane, & Woo 1991). The consumers were sorted into two groups, one for consumers with *more* reliable and expensive power services and one for consumers with *less* reliable and expensive power services. When provided with six reliability-to-cost mixes, with one option indicated as their status quo, the vast majority in both groups expressed a preference for the mix indicated as the status quo for their group (60% and 58%) while only a tiny fraction wanted the reliability-to-cost mix that the other group had (both 6%) (Hartman, Doane, & Woo 1991, p. 149).

The huge number of studies establishing the ubiquity and influence of the status quo heuristic might be resisted if they were all experiments concerning swapping chocolate bars and mugs, or even experiments concerning important financial decisions. The fact of the matter is, however, that even *post facto* analysis studies of real-life important decisions (such as choosing a mutual fund) also demonstrate the pervasive effects of the status quo heuristic (e.g. Kempf & Ruenzi 2006; Samuelson & Zeckhauser 1988). Even establishing torture as the status quo (by saying that it has been common practice for 40 years) made participants in one study significantly more likely to report supporting the practice of torture (Crandall *et al.* 2009).

Possible explanations, other than the status quo heuristic, do exist for these examples, but the literature in support of the status quo heuristic is extensive

⁵¹ Cost-free is perhaps not totally accurate here, since participants had to raise a sheet of coloured paper that was on their desk to indicate that they would like to switch rewards. However, this cost would be so small that it would only affect the behaviour of the laziest students... most of which would probably still have been at home in bed during the experiment.

and, while the mechanisms underpinning the status quo heuristic, and other related psychological heuristics, might not yet be fully understood, psychologists are generally convinced that the effects of the status quo heuristic on our judgments are significant, widespread, and often irrational (Gilovich, Griffin & Kahneman 2002; Kahneman & Tversky 2000). Therefore, the claim that responses to experience machine thought experiments are heavily influenced by status quo bias is greatly strengthened when the wealth of support for the status quo heuristic in the social sciences literature is considered.

7. The Case for Status Quo Bias Explaining Our Judgments about the Trip to Reality and Nozick's Scenario

In the experience machine thought experiment the choice to get into the machine involves giving up something very important and (in nearly all cases) valuable that you have and are familiar with (your current life) for something that is supposedly more valuable but fairly unknown to you (a life in the machine). It is clearly risky to consider swapping your current life for another one if you are unsure of what the new life will be like. I propose that both this caution regarding the unknown and an irrational over-valuing of the familiar is affecting intuitive judgments regarding the experience machine thought experiment. Evidence for this can be found by a comparison between Nozick's scenario and the Trip to Reality thought experiment, in which being in the machine is framed as our current (familiar) life. This comparison is similar to what Bostrom and Ord call the reversal test, which specifically assesses whether status quo bias is an important causal factor in the resulting judgment (Bostrom & Ord 2006). The reversal test turns the scenario around to frame the supposedly operant variable as the status quo rather than as a change in circumstances. In the Trip to Reality thought experiment, being in

an experience machine is described as our current real life, which means the status quo for us is being hooked up to a machine. Comparing both scenarios, we might discover that the majority of judgments are in favour of maintaining the status quo *regardless of how real the future experiences would be*.

We must pause at this point to consider if there are any differences between Nozick's scenario and the Trip to Reality thought experiment, other than the framing of the status quo. It turns out that there are other differences that seem to affect our intuitive judgments in these cases, but that they are all related to the status quo. The most notable difference is that the risks involved in each case seem markedly different. In the original experience machine thought experiment, our intuitive cognition would have deemed a machine life as risky, despite the stipulation in the thought experiment that the machine works perfectly. This intuition of risk likely arises from all of our previous experience with computerized machines crashing at least once, if not regularly, and often not providing the quality of performance that they promise.⁵² In the Trip to Reality thought experiment, the risks of machine failure and machine underperformance are less likely to affect our intuitive judgment about a life in a machine because that scenario would be matched to our non-crashing real-life experiences during the intuitive processing of the thought experiment.

It might be argued that these considerations imply that a preference for the status quo is not at all irrational; why would it be irrational to prefer to avoid potential risk? Of course, taking risk into consideration would not be irrational if you were offered to connect to an experience machine in real life. It would be rational to demand the highest possible level of evidence that there was no risk—something that is rarely possible in real life. In Nozick's

⁵² Note that this is my attempt at reconstructing the intuition and so, for the reasons discussed above, it should be viewed as an imprecise process.

scenario, however, the risk of machine underperformance or failure was ruled out by stipulation. Therefore, considering these risks during judgment formation about the thought experiment is irrational—it gives weight to irrelevant factors.

Another factor that seems likely to affect our judgments in the original experience machine thought experiment is Nozick's stipulation that we are to ignore family considerations (Nozick 1974, p. 43). Nozick explains that we do not have to worry about this because they could all plug into a machine as well (Nozick 1974, p. 43). It might be argued that the choice not to plug into a machine is therefore based on the rational preference not to force such a risky decision on your whole family, or if risk could somehow be eliminated, the rational preference not to force such an important decision on your whole family. This is a defect in Nozick's formulation of the thought experiment—it brings an important real-life factor into consideration that should be omitted when arguing about well-being, or what the good life is *for* the person living it. A person's machine life could be experientially identical, in respect to experiences with family, to a real life and need not include the experience of forcing them to do anything. If Nozick had simply stipulated that considerations of obligations to others should be ignored, then considering them would become irrational—it would give weight to irrelevant factors.⁵³ It is less obvious that the preference to maintain our current family relationships as they are is related to status quo bias, but it does reflect the most important aspects of an irrational preference for the way things are. The preference would be irrational in this case because it is for the real family connections over the different (but experientially the same or better) family connections

⁵³ In view of the discussion above, however, it should be clear that our intuitive cognition would probably not be able to reliably comply with the stipulation to ignore our family.

that would result in a machine life. Recall that when in the machine we would not know that we were no longer experiencing reality.

In a similar vein, it might be argued that the choice to remain in the machine in the Trip to Reality thought experiment is at least partially the result of a rational, as opposed to irrational, preference for the close relationships and other goods that we have and know in our machine life—perhaps the status quo heuristic at work, but not status quo bias. Indeed, this seems likely, and this further undermines the experience machine objection to hedonism. The experience machine objection to hedonism was devised to show that reality (and not just how our experiences feel on the inside) should matter to us and, since internalist accounts of Prudential Hedonism do not intrinsically value reality, they must be false. Consideration of the Trip to Reality thought experiment reveals, however, that it is in fact rational to prefer the life we experience over a *real* life, which is great news for internalist accounts of Prudential Hedonism.

All things considered, there are several potential differences between the experience machine and Trip to Reality thought experiments, and the reasons why they might elicit the judgments that they do. All of the important differences are related to the status quo, however—they are all related to what we have, what we know, and how we value those things. Therefore, the best explanation for why most people prefer their real life to a life plugged into Nozick's experience machine is not obviously that real experiences are important for our well-being. Rather, the best explanation might be that people's judgments are heavily influenced by an irrational preference for what they already have and know—by status quo bias. In order to be more confident about exactly how good an explanation status quo bias is for the widespread preference for reality over an experience machine life, we can turn to some initial empirical evidence.

8. De Brigard and Status Quo Bias

Kolber's (1994) & Weijers (forthcoming-a) arguments could be easily dismissed if it is *not* the case that more people would choose to stay in the machine on their versions than on Nozick's version. However, experimental data gathered by De Brigard (2010) seem to show that reversed experience machine thought experiments *do* elicit very different intuitions from Nozickian ones. De Brigard asked three groups of 24 Chapel Hill students whether they would choose an experience machine life or a real life in different scenarios. All of the scenarios were reversed experience machine thought experiments, but each described reality differently. De Brigard's Neutral scenario read as follows:

It is Saturday morning and you are planning to stay in bed for at least another hour when all of the sudden you hear the doorbell. Grudgingly, you step out of bed to go open the door. At the other side there is a tall man, with a black jacket and sunglasses, who introduces himself as Mr. Smith. He claims to have vital information that concerns you directly. Mildly troubled but still curious, you let him in. "I am afraid I have some disturbing news to communicate to you" says Mr. Smith. "There has been a terrible mistake. Your brain has been plugged by error in to an experience machine created by super duper neurophysiologists. All the unpleasantness you may have felt during your life is just an experiential preface conducive toward a greater pleasure (e.g., like when you had to wait in that long line to get tickets for that concert, remember?). Unfortunately, we just realized that we made a mistake. You were not supposed to be connected; someone else was. We apologize. That's why we'd like to give you a choice: you can either remain connected to this machine

(and we'll remove the memories of this conversation taking place) or you can go back to your real life." What would you choose? [Please circle only one option]

☐

Remain connected

☐

Go back to reality

Please explain your answer briefly: (De Brigard 2010, p. 47)

De Brigard's Positive and Negative scenarios both added two sentences to the end of his Neutral scenario (after "your real life"). In De Brigard's Negative scenario, the following is added: "By the way, you may want to know that your real life is not at all as your simulated life. In reality you are a prisoner in a maximum security prison in West Virginia." (De Brigard 2010, p. 47). And, in De Brigard's Positive scenario, this is added: "By the way, you may want to know that your real life is not at all as your simulated life. In reality you are a multimillionaire artist living in Monaco." (De Brigard 2010, p. 47). The results of these three scenarios are shown in Figure 1 below.

Figure 1: Reported Choices between an Experience Machine Life and One of Three Descriptions of Reality in De Brigard's Scenarios

Anyone who thinks that living in reality is very important to us might find the results for De Brigard's three scenarios quite surprising. Less than 13% (3/24) of the participants responding to De Brigard's Negative scenario chose reality, just over 54% (13/24) of the participants chose to return to reality in De Brigard's Neutral scenario, and 50% (12/24) of the participants responding to De Brigard's Positive scenario chose reality (De Brigard 2010, pp. 47–48).

De Brigard's Negative scenario is the closest to Kolber's (1994, p. 15) suggestion and, as Kolber predicted, the result is very different from what we would expect (and I have found) from testing Nozick's scenario.⁵⁴ Unfortunately, De Brigard did not test Nozick's scenario on any of his sample groups, so we can't easily get an idea of exactly how large the difference is between the responses to Nozick's scenario and to his reversals of it. The best we can do is to compare the results of De Brigard's Negative scenario with the International Wellbeing Study Further Assessment Study (IWSFAS) scenario. Less than 13% (3/24) of the participants responding to the De Brigard's Negative scenario and about 88% (137/156) of participants responding to the IWSFAS scenario reported preferring to live in reality. This difference is certainly large, but it's impossible to say how much of the difference is caused by varying characteristics between the sample groups.

De Brigard's Negative scenario also appears to be the most similar to Nozick's scenario in terms of the relative difference between the two key aspects that experience machine thought experiments are widely thought to compare (reality and how our experiences feel to us on the inside). In Nozick's scenario, the choice is between average but real experiences and great but unreal experiences. In De Brigard's Negative scenario, the choice is between bad but real experiences and average but unreal experiences. De Brigard believes that whether reality or an experience machine was framed as

⁵⁴ See Chapter 5.

the status quo explains a considerable amount of this difference between the responses to his scenarios and the responses to (non-reversed) Nozickian experience machine scenarios (2010, pp. 50–51).

De Brigard provided further evidence for the impact of status quo by conducting a follow up experiment; De Brigard's Neutral Status Quo Emphasised (Neutral SQE) scenario. De Brigard's Neutral SQE scenario is based on his Neutral scenario, replacing some of the text after 'give you a choice'. The changes are signified by italics in the following: "you can either remain connected to this machine (and we'll remove the memories of this conversation taking place) or you can *disconnect. However, you may want to know that your life outside is not at all like the life you have experienced so far.* What would you choose? Remain connected[/] *Disconnect*" (De Brigard 2010, p. 49). Just over 54% (13/24) of the participants chose to return to reality in De Brigard's Neutral scenario compared to 41% (33/80) of the participants in De Brigard's Neutral SQE scenario (De Brigard 2010, pp. 48–49). De Brigard argues that the 13% difference between his two neutral scenarios probably came about because the subtle change in the wording of the second scenario tapped into the status quo bias; i.e. that stipulating that "your life outside is not at all like the life you have experienced so far" in the Neutral SQE scenario made reality less familiar and therefore less desirable because of our irrational preference for things to stay the same (2010, pp. 48–49).⁵⁵ The results of these two scenarios are shown in Figure 2 below.

⁵⁵ Despite the small number of respondents in De Brigard's Neutral scenario group, applying a two-tailed Fischer's exact test to the data (with the hypothesis to expect a difference between the two groups) shows that we can be approximately 91% confident that the 13% difference in responses to these two scenarios was not caused by chance ($p = 0.094$).

Figure 2: Reported Choices between an Experience Machine Life and One of Two Descriptions of Reality in De Brigard's Scenarios

De Brigard contends that the best explanation for the results of his experiments is that experience machine thought experiments elicit intuitions that are tainted by status quo bias (2010, pp. 50–51). Basil Smith (2011), however, is not so sure.

9. Smith and Status Quo Bias

Basil Smith (2011) points out several weaknesses of De Brigard's experiments and conducts his own experiment to support his argument. Smith's criticisms of De Brigard's experiments include, the sample groups being too small, a lack of information about the procedure of the experiment, and an important disanalogy between De Brigard's scenarios and Nozick's scenario.⁵⁶

The sample groups in De Brigard's experiments certainly are small; three of the four groups contained only 24 participants (De Brigard 2010, p. 46). The small size of these groups means that we can be less confident in the results, especially if there are other potentially confounding factors involved. But, as

⁵⁶ Smith also expresses concern with the representativeness of De Brigard's all-student sample, but since this is also a potential issue for my experiments, it is addressed in the next chapter.

Smith points out, it's difficult to assess the likelihood of the influence of other potentially confounding factors because De Brigard provides little information about the procedure and background conditions of his experiments (Smith 2011, p. 35).

Even more importantly, Smith identifies a subtle but significant disanalogy between De Brigard's scenarios and Nozick's scenario that seems like it would encourage the participants in De Brigard's scenarios to choose an experience machine life over reality (Smith 2011, pp. 37–39). In Nozick's scenario, neither remaining in reality nor connecting to an experience machine would necessarily mean experiencing starting life anew (i.e. starting life with completely new relationships and employment and domestic circumstances). Continuing a real life clearly doesn't involve starting life anew and, since a life in the experience could be very similar to a person's real life (except with a much higher incidence of amazing experiences), an experience machine needn't *feel* like starting life anew either. In De Brigard's scenarios, however, there is an imbalance between the two choices; remaining connected to an experience machine does not involve starting life anew, but returning to reality *does*. In De Brigard's Positive and Negative scenarios, life starting anew is stipulated by the inclusion of "... your real life is not at all as your simulated life" (2010, p. 47). In De Brigard's Neutral SQE scenario, life starting anew is stipulated by the inclusion of "... your life outside is not at all like the life you have experienced so far" (2010, p. 49). Although there is no similar stipulation in De Brigard's Neutral scenario, starting life anew is at least partially implied in this scenario as well (2010, p. 47).

De Brigard might have two responses to this criticism. First, he might suggest that Nozick's scenario also implies that choosing the experience machine life requires starting life anew in terms of the above-mentioned personal circumstances, thereby making his scenarios analogous with

Nozick's after all. It is true that Nozick's silence on this point may lead some people to believe they would have to start their life anew in the experience machine.⁵⁷ However, there still remains a large difference between the slight ambiguity in Nozick's scenario and the implication and stipulations in De Brigard's scenarios on this issue.

De Brigard might also argue that starting life anew is an integral facet of losing the status quo. Such a response would be inadequate, however; an experience machine life in Nozick's scenario does involve a change in the status quo, but it does not necessitate a change to the same magnitude. While Nozick's scenario involves a change in a person's connection to reality, De Brigard's scenarios involve a change in a person's connection to reality and a change in their social, domestic, and employment circumstances. Importantly, such a change in these personal circumstances is also likely to result in a major interruption in that person's pursuit of their life plans and goals.

Furthermore, Smith's own experiments highlight the importance people tend to place on not dramatically changing these personal circumstances (Smith 2011, pp. 40–44). Smith conducted nine experiments on a variety of different groups, including members of a philosophy club from London, students from a community college in Virginia, a hiking group from Southern California, and attendees of a Sunday service at Westminster Abbey (2011, p.40). The individual groups were small (between 13 and 30 participants) and were probably homogenous in many respects. Taken together, however, the 184 participants represented both genders and most ages well. The groups as a whole are also likely to have covered a wide range of personal characteristics, although not necessarily in a way that is representative of any

⁵⁷ Indeed, my test of Nozick's scenario revealed that four respondents (out of the 99 who chose reality) reported reasons in line with cutting off their current relationships as a justification for their choice. See Chapter 5 for more details.

large population. Smith's scenarios were based on De Brigard's Neutral scenario and will be referred to as Smith's Pretend Neutral scenario and Smith's Neutral scenario. Half of each of the smaller groups was presented with Smith's Pretend Neutral scenario and the other half was presented with Smith's Neutral scenario (Smith 2011, p. 42). Smith's Pretend Neutral scenario reads as follows:

It is a Saturday morning and you are planning to stay in bed for at least another hour when all of the sudden you hear the doorbell. Grudgingly, you step out of bed to go to the door. At the door is a tall man, with a black jacket and sunglasses, who introduces himself as Mr. Smith. He claims to have vital information that concerns you directly. Mildly troubled but still curious, you let him in. "I have some news to communicate to you" says Mr. Smith. "You are a client of my corporation. Decades ago, you employed our super-duper neurophysiologists to plug your brain into an experience machine. All the experiences you have had so far are nothing but the product of a computer program designed to provide you with pleasurable experiences. All the unpleasantness you may have felt during your life is just an experiential preface conducive toward greater pleasure (e.g. like when you had to wait in that long line to get tickets for that concert, remember?). Of course, you do not remember employing us. Regardless, every decade, we offer our clients the option of either remaining connected to this machine (and yes, we'll remove your memories of this conversation), or you can go back to your real life."

Please understand that, if you choose to return to the real world, your entire life up to this point has been *fabricated*. So, if you choose to return, you will be starting life anew, without any of the relationships, employment, or domestic circumstances of this life. Since all of this was fabricated, *none of it ever existed*. Therefore, if you do so choose, we can give you counselling, and can offer you a comparable future. But you will begin anew. (Smith 2011, pp. 41–42)

Smith refers to this scenario as being “Pretend Neutral” to make the point that De Brigard’s Neutral and (especially) Neutral SQE scenarios don’t describe reality as neutral at all because they involve the negative feature of requiring participants to start life anew (2011, p. 42). Smith’s Pretend Neutral scenario is designed to be analogous to De Brigard’s Neutral SQE scenario so that he can compare it to another scenario that does not require that participants start their life anew if they choose reality (Smith’s Neutral scenario). This comparison will indicate how much effect the prospect of starting life anew has on participants when choosing between reality and an experience machine life. If the comparison reveals that the prospect of starting life anew has a significant effect on participants’ choices, then De Brigard’s inclusion of the requirement that those choosing reality must start their life anew means his results have much less relevance to critiquing Nozick’s scenario.

Despite the major sampling differences between De Brigard’s Neutral SQE scenario and Smith’s Pretend Neutral scenario, these roughly analogous scenarios produced fairly similar results. About 71% of the respondents to Smith’s Pretend Neutral scenario (Smith 2011, p. 43) and about 59% of the respondents to De Brigard’s Neutral SQE scenario chose to remain connected to an experience machine (De Brigard 2010, p. 49). As discussed, the similarity

of these two results makes the outcome of Smith's Neutral scenario relevant when assessing the importance of De Brigard's results for critiquing Nozick's scenario.

Smith's Neutral scenario starts with the exact same paragraph that his Pretend Neutral scenario begins with, but replaces the second paragraph with one that reads as follows:

Please understand that, if you choose to return to the real world, your entire life up to this point has been *fabricated*. However, this will not affect you, and you will not know about it. Scientists have sifted through your memories (i.e. our computer files), copied them, have cloned all your relations, created your exact job, and have *rebuilt* your domestic circumstances. If you choose to return to reality, we will erase your memory of this conversation, and your life will continue as before. (Smith 2011, pp. 41–42)

About 27% of the respondents to Smith's Pretend Neutral scenario chose to remain connected to an experience machine (Smith 2011, p. 43). When compared to the 71% of the respondents who chose to remain connected to an experience machine in Smith's Pretend Neutral scenario, the difference is large (about 45%) and highly statistically significant.⁵⁸ The results of these two scenarios are shown in Figure 3 below.

⁵⁸ Using a two-tailed Fischer's exact test, we observe that we can be more than 99.9% confident that the difference between these two scenarios is not the product of chance ($p = 0.000$ to 3 d.p.).

Figure 3: Reported Choices between an Experience Machine Life and One of Two Descriptions of Reality in Smith's Scenarios

The results of Smith's experiments indicate that, at least for his eclectic sample, whether an option in a choice involves starting life anew can have a very large impact on that choice. Specifically, Smith's results indicate that having to acquaint oneself with new social, domestic, and employment circumstances seems to be considered as a very large disincentive when choosing between lives. This essentially endorses an aspect of Kawall's insight that people will always tend to reject offers for 'better' lives if that means giving up their "current lives and commitments" (Kawall 1999, p. 383). This is somewhat surprising because De Brigard cites this very passage from Kawall in defense of his argument (De Brigard 2010, p. 51). Nevertheless, De Brigard's construction of his scenarios ties reality with starting life anew and therefore makes them disanalogous to Nozick's scenario in what seems to be a very important way.

Smith's scenarios are also disanalogous to Nozick's scenario. Smith's *Pretend Neutral* scenario is disanalogous for the same reason that De Brigard's *Neutral SQE* scenario is—it includes the stipulation that life must start anew in reality. Smith's *Neutral* scenario is disanalogous, but for a different reason—it implies that there is no difference between the quality of

experiences in reality and in an experience machine. This implication arises from very little mention of what experience machines do, no specific mention of the relative difference in the quality of experiences between reality and in an experience machine, and the stipulation in the last line of the scenario that “If you choose to return to reality, we will erase your memory of this conversation, and *your life will continue as before.*” (Smith 2011, p. 42, emphasis added).

Considering all of the problems Smith has identified with De Brigard’s experiments, and his own intriguing experimental results, what should we conclude about the possible effects of status quo bias on responses to experience machine scenarios? First, it is clear that there are enough important differences between both De Brigard’s and Nozick’s, and Smith’s and Nozick’s, scenarios that none of them are related enough to provide direct support or criticism of Nozick’s scenario.

But it should also be clear that both De Brigard’s and Smith’s experiments suggest that status quo bias has at least some effect on at least some respondents’ choices between lives in reality and lives in experience machines. In Smith’s Neutral scenario, nearly 27% of respondents chose to remain connected to an experience machine even though it was ambiguous as to whether their lives would be different in any way if they switched to reality. It seems reasonable to believe that a considerable proportion of this 27% were affected by a desire not to change their status quo.

Despite all of the problems with De Brigard’s experiments, the difference between his Negative scenario and the IWSFAS scenario is very large (about 75%). Even if we assume that the difference between Smith’s scenarios (about 45%) was wholly caused by whether respondents had to start life anew if they chose reality, much of the difference between the results of De Brigard’s Negative scenario and the IWSFAS scenario remains in need of explanation.

Again it seems reasonable to believe that a considerable proportion of this remaining amount (about 30%) could be explained by reference to status quo bias. Of course, these differences could be explained by several biases or other factors, including sampling differences or even chance, but the established pervasiveness of status quo bias (discussed above) offers considerable support for the claim that the effects of status quo bias on our responses to experience machine thought experiments considerably reduces the philosophical value of those responses.

10. Conclusion and Implications

While it's not clear exactly how much influence status quo bias and other confounding factors have on our judgments about Nozick's scenario, it seems very likely that they are having at least some affect. Therefore, there is a strong case that the normative significance of the intuitions elicited by Nozickian experience machine thought experiments is undermined. Furthermore, since reversed experience machine thought experiments, such as De Brigard's and Smith's, also frame certain aspects of one of the choices as the status quo, the normative significance of the intuitions elicited by these kinds of experience machine thought experiment is also undermined.

Furthermore, that the experience machine thought experiment appears to have misled the widespread and firmly held judgments of so many reasonable people stands testament to the fact that we should all be less confident about using thought experiments *alone* as evidence for our arguments. This is especially the case when those thought experiments are unrealistic, since our intuitive cognition is much more likely to process irrelevant factors in such cases (Churchland 1996). The status quo bias is of course just one of many psychological biases that can have major effects on any of our judgments with a significant intuitive component. So, heed the

warning; many thought experiments in philosophy should be revisited by applying what we are currently learning about intuitions to ensure that more of our firmly held judgments are not, unbeknownst to us, marred by the effects of psychological biases.

But, most importantly, what does this all mean for the experience machine objection to hedonism? Is status quo bias the best explanation for why the vast majority of reasonable people report preferring reality over a life in an experience machine? And, would the vast majority of people choose reality over an experience machine life in a scenario that was free from the pernicious effects of status quo bias and other confounding factors?

It is clear from the foregoing analysis that status quo and related biases are very likely to be affecting reasonable people's judgments about experience machine scenarios to some extent. But it is *not* clear that status quo bias is the *best* explanation for why the vast majority of reasonable people report preferring reality over a life in an experience machine, even when combined with the other potentially confounding factors mentioned in the previous chapter. The nature of intuitive cognition simply makes it too difficult to speculate about the magnitude of the effects of these factors. Perhaps the best way to get a better idea about whether certain factors have a *significant* influence on our judgments about experience machine cases is to test whether very minor differences in the scenarios significantly affect the outcomes from the same sample of respondents. Identifying and eliminating confounding factors in this way might tell us what the best explanation is for why the vast majority of reasonable people report preferring reality over a life in an experience machine in Nozick's scenario. Furthermore, eliminating confounding factors brings us closer to an unbiased experience machine scenario that might not produce widespread agreement about whether reality is a better choice than a life in an experience machine. If this is the case, then

one of the strongest aspects of the experience machine objection to hedonism (that nearly everyone agrees on which the best life is) will become untenable, putting the whole argument in jeopardy.

To further investigate these possibilities, Chapter 5 attempts to isolate and eliminate status quo bias and other confounding factors from the experience machine thought experiment through the testing method described above. But before this task can be carried out with good conscience, Basil Smith's (2011) argument that the experience machine cannot be properly tested at all must be assessed. This is the task of the next chapter.

Chapter 4

We Can Test the Experience Machine

Chapter Summary

Critiques of experimental philosophy abound. The most potentially damning of these criticisms are arguments against certain kinds of experiments being able to provide useful information about specific philosophical questions. For example, in his provocative “Can We Test the Experience Machine?” (2011), Basil Smith argues that we should recognise a limit on experimental philosophy and that experiments on the experience machine fall on the wrong side of that limit. In this chapter, I will argue that even if we take Smith’s limit seriously it does not prevent us from usefully testing most experience machine thought experiments, including De Brigard’s inverted experience machine scenarios. I will also argue that taking Smith’s limit seriously has far-reaching consequences for traditional (non-experimental) philosophy. Then, I will argue that Smith’s limit is too rigid to be taken seriously anyway. Finally, I address Smith’s objection to student samples being useful for testing philosophical thought experiments.

1. Introduction⁵⁹

The advent of experimental philosophy has intensified debate about the role of intuitions and judgments about thought experiments in philosophy. The rise of experimental philosophy has also led to a broad range of criticisms about the aims and techniques of experimental philosophers. Since criticisms can lead to improvements, these debates appear likely to improve both

⁵⁹ This chapter provided the basis for Weijers (forthcoming-b).

experimental and traditional philosophy. This chapter enters these debates, but only in a very limited way. This chapter focuses on the only argument in these debates that specifically targets the experience machine.

In his provocative “Can We Test the Experience Machine?” (2011), Basil Smith argues that we should recognise a limit on experimental philosophy that acknowledges the uselessness of testing certain kinds of thought experiments, including the experience machine. Smith particularly targets De Brigard’s (2010) experiments on experience machine scenarios. This chapter analyses Smith’s proposed limit and concludes that Smith’s arguments should do little to deter most kinds of tests on experience machine thought experiments.

In this chapter, I will argue that even if we take Smith’s limit seriously, it does not prevent us from usefully testing most experience machine thought experiments, including De Brigard’s inverted experience machine scenarios. I will also argue that taking Smith’s limit seriously has far-reaching consequences for traditional (non-experimental) philosophy. Then, I will present counterexamples to Smith’s limit and argue that it is too rigid to be taken seriously anyway. Finally, I address Smith’s objection to student samples being useful for testing philosophical thought experiments.

2. Smith’s Limit

Smith describes his proposed limit on experimental philosophy as follows:

*[C]ertain philosophical thought experiments (e.g. the experience machine, the trolley problem, etc.) ask subjects to make decisions from the position of *confronted agents*, or those who have entered a specific state of mind. But when taking surveys (answering either ‘yes’ or ‘no’ or giving a score on a Likert scale), subjects are *not* in that state of mind, nor can they*

imagine it. Therefore, when experimental philosophers claim they have tested certain philosophical experiments (and thereby the intuitions they evoke) we have reason to believe they have not. (Smith 2011, p. 30, his italics)

Essentially, Smith recommends that experimental philosophers should stop asking questions of the form: ‘how *would* you react in this situation?’ (henceforth ‘*would* questions’), especially when the situation is hypothetical, intense, and unfamiliar. I will refer to this as ‘Smith’s limit’. Smith argues for his limit on the grounds that survey answers to *would* questions do not provide useful information because they require respondents to perform a very difficult task—accurately predicting their reactions to hypothetical, intense, and unfamiliar situations.

Smith gives two reasons for why it is so difficult for survey respondents to accurately predict their reactions to hypothetical, intense, and unfamiliar thought experiments. Both of the reasons are related to how difficult it is for survey respondents to adopt the role of an appropriately confronted agent. A confronted agent is someone who has context-specific subjective experiences, including affective responses such as feelings of confusion, incredulity, fear, and uncertainty. First, the respondents would have none of the affective responses they would have experienced if the scenario were real (Smith 2011, p. 39). Second, respondents would have given their ideal responses (how they think they *should* react in the given scenario) as opposed to how they think that they would actually react if they really found themselves in the hypothesised scenario (Smith 2011, p. 39). Since the affective responses we experience can heavily influence our reactions, and since how we think we should react is often different to how we would actually react, Smith

concludes that “we cannot compare the responses subjects give on a survey with their reactions to the real event.” (Smith 2011, p. 39, his italics).

Smith also provides two examples of types of studies that require respondents to adopt the role of confronted agents *to some extent*; studies that involve respondents anticipating their own futures and ones that involve respondents attempting to identify with the moral decisions of others (Smith 2011, p. 46). Smith concludes his article by clarifying that his limit on experimental philosophy affects studies “*just to the degree*” that they require respondents to adopt the role of confronted agents (Smith 2011, p. 46, his italics). Presumably this means that studies with questions requiring respondents to undertake just a small amount of *would*-based pondering (relative to the total pondering required to form a judgment about the relevant scenario) should not automatically be considered failed tests, just somewhat undermined tests.

3. Taking Smith’s Limit Seriously

Smith acknowledges that his limit “does not apply to most experimental studies” (Smith 2011, p. 46). But does it apply to studies on experience machine thought experiments? Or, as Smith puts it in the title of his paper: can we test the experience machine? Although never directly stated, we can infer that Smith’s answer to this question would be: ‘we can only test experience machine thought experiments that do not require respondents to adopt the role of a confronted agent’. Smith is very clear about his belief that De Brigard’s inverted experience machine thought experiments cannot be tested:

[T]he inverted experience machine, as well as other similar such experiments (e.g. justified theft dilemmas, questions of torture, the trolley problem, etc.) have a unique set of characteristics that

make it impossible to gather the right subjects to test [i.e. subjects that can adopt the role of an appropriately confronted agent]. Therefore, in practice, these thought experiments are impossible to test. (Smith 2011, p. 37)

Smith is relatively quiet, however, on whether other variants of the experience machine thought experiment can be tested. I will argue that his limit does not prevent us from usefully testing most experience machine thought experiments, including De Brigard's inverted experience machine scenarios.

The most famous experience machine thought experiment was proposed by Robert Nozick. Nozick describes experience machines as being able to "give you any experience you desired... a lifetime of bliss" all without you realising that your experiences were machine-generated (Nozick 1974, pp. 42–43). Nozick then asks his readers: "Would you plug in" to an experience machine for the rest of your life? (Nozick 1974, pp. 43). This is clearly a *would* question—one that, on Smith's view, only an appropriately confronted agent could credibly provide an answer to. Therefore, Smith would object to any attempt to test this scenario.

De Brigard (2010, p. 47) asks "What would you choose?" at the end of all of his scenarios. Therefore, applying Smith's limit to De Brigard's inverted experience machine thought experiments reveals the same judgment; Smith would object to any attempt to test this scenario too.

Smith's (2011, p. 37) claim that De Brigard's tests are useless, and his implication that any test of Nozick's experience machine scenario would also be useless, is based on the assumption that these tests have the purpose of discovering what people would choose if they were really offered a choice between reality and a life in an experience machine. It's far from clear,

however, that when De Brigard and Nozick were creating their thought experiments discovering the truth about what people *would* do was their ultimate aim.

The stated purpose of De Brigard's tests was to investigate how informing participants that they "have been living inside an experience machine [their] entire life... would affect, per se, their judgements on their own happiness or well-being" through studying their responses to his scenarios (De Brigard 2010, p. 46). It is true that De Brigard mentions the connection between the experience machine and Psychological (Motivational) Hedonism: "Many philosophers... consider the alleged effectiveness of this thought-experiment a rather convincing proof against the tenability of psychological hedonism. (De Brigard 2010, p. 45). However, this is the only mention of it in the whole article. De Brigard makes it more clear in his conclusion (2010, pp. 53–55) that his primary research question is whether our preferences in experience machine scenarios are best explained by our valuing of reality or the status quo (i.e. the truth of IP3).⁶⁰

The stated purpose of Nozick's experience machine scenario is to demonstrate that "something matters to us in addition to experience" (Nozick 1974, p. 44). Furthermore, when Nozick discusses his experience machine thought experiment in more detail, he stresses that he takes the experience machine to provide evidence about what has prudential value, not just what motivates us: "the connection to actuality is important whether or not we desire it—that is *why* we desire it—and the experience machine is inadequate because it doesn't give us *that*" (Nozick 1989, pp. 106–107, his italics).

⁶⁰ Restated for your convenience here. IP3: The best explanation for IP2 is that reality matters intrinsically to the vast majority of reasonable people. IP2: When instructed to ignore their responsibilities to others, the vast majority of reasonable people report preferring reality over a life in an experience machine.

For both of these stated purposes, then, it seems that knowing what people think is in their best interest to do, would be more useful than knowing what they would actually choose. This is because (as Smith has pointed out) when actually choosing, people are confronted agents and their decisions are affected by feelings of confusion, incredulity, fear, and uncertainty. In most cases, we should expect these influences to cause confronted agents' revealed choices to be less rational than they would have otherwise been.⁶¹ However, if we could really know what people think they *should* choose, then we would have a much better understanding of our well-being-related judgements, which is what it seems De Brigard and Nozick were really after.

In fact, assuming that Smith is correct about the reasons why survey participants cannot credibly respond to *would* questions, then De Brigard's and Nozick's 'what would you choose?' questions are probably eliciting responses more closely aligned with what the respondents think they *should* choose if they were given that option. If the feelings of confusion, incredulity, fear, and uncertainty that are required to properly adopt the role of a confronted agent for De Brigard's and Nozick's scenarios are likely to cause less rational choices, then the removal of them is likely to lead to *more* rational choices (i.e. choices that better reflect what the respondents think they should choose).⁶² Therefore, while Smith (2011, p. 39) bemoans the fact that responses to De Brigard's scenarios are likely to be the participants' 'ideal choices', as

⁶¹ See, for example, Hodgson (1988, pp. 10, 11, 57) on the many revealed 'decisions' based on unconscious reactions to context and psychological states of stress, fear, excitement, social pressure, and so on.

⁶² I am not making the false claim that rational decisions are those made with as little emotional processing as possible, just that decisions made under normal emotional conditions are likely to be more rational than those made under extreme emotional conditions.

opposed to what they would actually choose if the scenario were real, this actually seems like a strength of De Brigard's experiments.

Indeed, using questions of the 'how *would* you react in this situation?' variety is common, even when the researchers are not directly interested in what respondents would actually do if the scenarios were real. For example, Petrinovich and colleagues ran experiments on typical philosophical thought experiments (including trolley problems) using explicit *would* questions, but made it clear that they were most interested in better understanding people's underlying moral intuitions (Petrinovich, O'Neill, & Jorgensen 1993, p. 476).⁶³ Furthermore, Alex Barber (2011, p. 263, n. 6, his italics) has argued that "I read [Nozick's experience machine objection to hedonism] as being interested in your views about what you (self-interestedly) *should* do, but what you *would* do is our best guide to these views."⁶⁴ 'Best guide' is surely too strong here, but this quote shows that philosophers are also aware of the importance of the

⁶³ "It also might be objected that this research concerns only intuitions and that it is not possible to determine whether the intuitions revealed here would be translated into action. Of course, this objection is justified. However, the intent of this research is not to understand or predict what people would do but to probe the structure of their systems of moral intuitions. The impetus for the research was provided by the views of many moral philosophers regarding how people construct the world of morality. We argue that these dilemmas probe the core of moral beliefs. It would seem that an understanding of how people resolve these fantasy dilemmas might be a good basis on which to begin to understand action but that translation is not part of the present undertaking, although it would be a fascinating question for future research." (Petrinovich, O'Neill, & Jorgensen 1993, p. 476).

⁶⁴ Barber goes on to congratulate Nozick for revealing to Hedonists that their theoretical convictions might contradict their professed hypothetical actions since it might be able to get Hedonists to admit that in their "heart of hearts" they are not really Hedonists (2001, p. 266). Barber summarises: "Our hypothetical actions speak louder than our theoretical words" (2001, p. 266).

distinction between what we would do and what we should do in regards to the experience machine.

So what does this mean for experimental studies of experience machine scenarios that directly ask *would* questions, but do not seek to discover how people would actually react if the scenarios were real (a category that seems to include De Brigard's experiments)? Does Smith's limit deem them useless?

Even if the questions were reworded to (or reinterpreted by non-confronted respondents as) 'what *should* you choose?' questions, De Brigard's and Nozick's scenarios still require respondents to contemplate some *would* questions. On the new wordings (or interpretations), respondents no longer have to predict if they would actually choose an experience machine life, but they still have to ask themselves what a life in an experience machine *would* be like. Imagining what an experience machine life would be like is fairly easy for someone who has previously been confronted with one. Unfortunately, no one has come across an experience machine yet, so there is no group of people who could answer the question 'what should you choose?' as an appropriately confronted agent. This is one of the main points that Smith was trying to make (captured in the following claim):

[T]he inverted experience machine... [has] a unique set of characteristics that make it impossible to gather the right subjects to test. Therefore, in practice, [it is] impossible to test.
(Smith 2011, p. 37)

Nevertheless, it seems like participants' responses to the question 'what should you choose?' require less knowledge and emotional upheaval to be credible than their responses to the question 'what *would* you choose?' So, Smith might conclude that any studies on reworded versions of Nozick's or

De Brigard's experience machine thought experiments should be considered as flawed, but still useful tests, as opposed to being worthless non-tests.

And what of De Brigard's actual experiments and Nozick's actual scenario? Given their probable purposes, and the likelihood of their 'what would you choose?' questions being reinterpreted by readers and respondents as 'what should you choose?' questions, then it seems that Smith's limit does not rule out their usefulness. Indeed, it seems like only a very uncommon kind of experimental study on an experience machine scenario would be ruled out by Smith's limit. Even more importantly for current purposes, even if De Brigard's actual intent was to investigate the relevance of the experience machine to the plausibility of Motivational Hedonism, *I* have usefully employed his tests to investigate the relevance of the experience machine to the plausibility of *Prudential Hedonism*. Furthermore, the same arguments that I have run here could be used to redeem most of the other experimental studies that Smith claims to be failed tests (trolley problems etc.).

All of the results of my application of Smith's limit differ to those from his application of it. I have argued that appropriately applying Smith's limit identifies a few experimental studies as worthless and many as flawed (but still useful). Smith may disagree with my analysis of which particular studies belong in each of these categories, but we seem to agree on his justifications for being concerned about *would* questions. Non-confronted agents cannot imagine having the intense emotions that *really* being in extreme and unfamiliar situations elicit. And since intense emotions have can have strong and unpredictable effects on our reactions, non-confronted agents should not be expected to be able to accurately predict how they would react in extreme and unfamiliar situations. This is why responses to experience machine scenarios should not be taken at face value—no one could be sure about how

they would react to being in an experience machine. So, Smith and I agree that questions requiring survey respondents to predict how they would react in future or hypothetical scenarios are undermined, but perhaps we disagree about the extent to which asking *would* questions affects an experiment's usefulness.

4. How Taking Smith's Limit Seriously Could Affect Traditional Philosophy

When considering the extent to which asking *would* questions will affect an experiment's usefulness, it is worth also thinking about a very closely analogous case. Notice that when philosophers, and other readers of philosophy, ponder thought experiments they too are required to predict how they would react in future or hypothetical scenarios, or attempt to identify with the moral decisions of others. Notice also that philosophers and philosophy students don't usually appear to be experiencing feelings of confusion, incredulity, fear, and uncertainty when they read or discuss Nozick's experience machine or any other thought experiment.⁶⁵ Smith makes no comment about his limit on experimental philosophy applying to philosophy in general, but his justifications for the limit seem to apply to nearly all of philosophy. Many arguments in philosophy use thought experiments that require the reader to either predict how they would react in future or hypothetical scenarios or attempt to identify with the moral decisions of others. Therefore many arguments in philosophy would be deemed either flawed or useless by Smith's limit.

⁶⁵ I think that the reality here is that imagining a thought experiment *does* cause affective responses in the imaginer, but that these responses are much more muted than they would have been if the scenario were really experienced. The difference in responses to my Stranger and Self scenarios, discussed in the next chapter, seem to provide evidence for this claim.

I repeat Smith's justification for his limit here, with a few words changed (indicated by square brackets), to show how easily it applies to much more than just experimental philosophy:

[C]ertain philosophical thought experiments (e.g. the experience machine, the trolley problem, etc.) ask subjects to make decisions from the position of *confronted agents*, or those who have entered a specific state of mind. But when [reading thought experiments, readers] are *not* in that state of mind, nor can they imagine it. Therefore, when [people] claim they have [learnt from] certain philosophical [thought] experiments (and thereby the intuitions they evoke) we have reason to believe they have not. (Smith 2011, p. 30, his italics)

Of course, there are other reasons why the results of experimental studies may be less reliable than the opinion of philosophers about a particular thought experiment.⁶⁶ But Smith discusses his limit on experimental philosophy separately, claiming that it is strong enough *by itself* to show that certain thought experiments cannot be tested at all (2011, p. 37). Given that the justifications for Smith's limit imply that all thought experiments are undermined to the extent that they require us to adopt the role of confronted agents, I am curious about what Smith believes. He might believe: that philosophers, and others who read thought experiments, are better at adopting the role of confronted agents, or that large swathes of philosophy are flawed or useless, or that Nozick's and De Brigard's scenarios (whether experimented on or simply read by someone) are only slightly undermined by his concern about adopting the role of an appropriately confronted agent.

⁶⁶ Smith (2011) discusses many potential problems with experimental studies, all of which can be minimised or avoided by sound experimental design.

In the context of the debate about the effectiveness of the experience machine objection to hedonism, it seems that most of these possibilities would be welcomed by Hedonists. If both reading experience machine thought experiments and responding to experience machine scenarios in a survey elicit severely flawed or useless judgments, then IP3 is likely to be false; i.e. the best explanation for why the vast majority of reasonable people report preferring reality over a life in an experience machine is probably *not* that reality matters intrinsically to the vast majority of reasonable people. Indeed, we wouldn't be able to say what the best explanation for the widespread reported preference for reality is because those reports are all useless (or at least severely undermined).

Furthermore, if both reading experience machine thought experiments and responding to experience machine scenarios in a survey elicit only slightly undermined judgments, then De Brigard's (2010) and others' challenge to IP2 (based on alternate experience machine thought experiments) should not be significantly affected; i.e. since Smith's limit would not significantly reduce our confidence in De Brigard's (2010) experimental result (that less than 13% of respondents to his Negative scenario chose reality over an experience machine life), we have reason to believe that the vast majority of reasonable people do *not* always choose reality over an experience machine life. Furthermore, the slight reduction in confidence that Smith's limit causes us to acknowledge with De Brigard's results should apply equally to the original interpretation of IP2 based on people reading (but not being tested on) Nozick's scenario. Therefore, even if we take Smith's limit seriously, we seem to have no significant reason to believe that the arguments of De Brigard (2010) and others, against the experience machine objection to hedonism, are any weaker.

5. Smith's Limit is too Rigid to be Taken Seriously

It is not clear that we have good reason to take Smith's limit clearly, however. Given the justifications Smith provides for his limit, the limit itself is too rigid. Just how useless the inclusion of *would* questions will make an experiment seems to be a question of degree. Indeed, this seems to be the case even when the *would* questions are explicit and involve participants attempting to predict their reactions to hypothetical, intense, and unfamiliar situations.

For example, imagine a survey designed to investigate the risks people would take for financial reward. A group of people would be asked (via survey) if they would accept an offer of \$1million to attempt a dangerous high-altitude balancing stunt. I contend that they would probably not exhibit any signs of feeling confused, incredulous, fearful, or uncertain, but their answers would likely be fairly predictive of their actual choice if they were really presented with such an offer. Therefore, even though this group would be being asked to predict their reactions to a hypothetical, intense, and unfamiliar situation, *and* they failed to adopt the role of an appropriately confronted agent, their responses would be far from useless. This example shows that Smith's limit is too rigid because it is not sensitive to the fact that all hypothetical, intense, and unfamiliar situations are not created equal; we are likely to be better at predicting our reactions to some situations even if we can't adopt the role of a confronted agent.

This criticism suggests that Smith's limit shouldn't be a limit at all. Rather, it should be something that experimental philosophers, and researchers using experimental data, should consider when evaluating the significance of the data for their particular research question. Researchers should ask themselves how useful the survey respondents' non-confronted responses are likely to be. If the research question is 'how would members of a certain group react to their home town being invaded by evil aliens?', then the non-confronted

responses will be of limited value. But, if the research question is ‘how *should* members of a certain group react to their home town being invaded by evil aliens?’, then the non-confronted responses will be worth something. Furthermore, it should be noted that non-confronted responses to this question are likely to be *more* useful than confronted responses to it because non-confronted responses will be less affected by irrational emotional influences. Therefore, Smith’s limit should be understood as one quality control check of data, among many, that may lead to a reduction in the degree of significance the data has for the research question.

This new interpretation of Smith’s limit supports the possibility from above that both reading experience machine thought experiments and responding to experience machine scenarios in a survey elicit only slightly undermined judgments when the research question is ‘does reality have intrinsic prudential value?’ And, as mentioned, since this slight undermining applies to both the defenders and attackers of IP2 (that we prefer reality over an experience machine life), Smith’s newly interpreted limit does not significantly affect the debate about the experience machine objection to hedonism.

6. Student Samples

Smith (2011) levels one more interesting criticism at experimental philosophers, which applies to the experiments of De Brigard and the experiments I conduct in the next chapter. Smith criticises the use of student samples, arguing that they are not representative enough to provide accurate results for experimental philosophy, including for testing the experience machine. Smith argues that many factors about participants, such as their age, class, religious beliefs, gender, and culture, will affect whether they are likely to think that connecting to an experience machine is a good idea (2011, pp. 35,

40). I expect that Smith is entirely correct, but that doesn't entail that middle-class, Western, undergraduate students are *always* an unrepresentative sample. In this part of my thesis, I am investigating the merits of using the experience machine objection to hedonism to make an argument in philosophy and arguments in philosophy often intend to be able to convince nearly all reasonable people who might come across them. According to IP2 of the argument, we want to know about the judgments of 'reasonable people'. Are students representative of *all* reasonable people?

We can probably assume that a sufficiently large sample of students will be made up of mainly reasonable people; however it is clear that student samples are not ideally representative of all reasonable people. But they might not be all that far off. As discussed in the previous chapter, Smith's (2011) test of his Pretend Neutral scenario on a two-country, multi-setting, and multi-demographic group of participants produced a fairly similar result to De Brigard's test of his roughly analogous Neutral Status Quo Emphasised scenario on a mono-country, mono-setting, and all-student group (71% and 59% respectively). Furthermore, most of the 12% difference between these two results might be explained by the extra emphasis that Smith put on having to start life anew in reality in his Pretend Neutral scenario.

What does this mean in relation to the debate about the experience machine objection to hedonism? It certainly means that we should consider results from student samples to be slightly less significant than results from more representative samples of all reasonable people. But it should also be noted that all-student sample groups are probably more representative of all reasonable people than trained philosophers are. The average undergraduate student class is clearly a more diverse group than most gatherings of trained philosophers in terms of their values, interests, personalities, and so on. Therefore, any proponent of the experience machine objection to hedonism

arguing something along the lines of: 'De Brigard's experimental results are not very significant because they are based on a fairly unrepresentative sample', faces a devastating response. De Brigard could simply deny IP2 because philosophers' evidence for it comes from their presentations of experience machine scenarios to philosophers and philosophy students (i.e. De Brigard could say that we do not know whether the vast majority of reasonable people report preferring reality over a life in an experience machine because *no one* has ever bothered to ask a representative sample of reasonable people!) So again, Smith's criticism seems to have only a small effect on how we should evaluate empirical results, but it also has an equal or greater effect on the practice of traditional philosophers.

Furthermore, the worry about the representativeness of student samples can be somewhat allayed by testing Nozick's scenario as well as the alternate experience machine scenarios on the same all-student sample group (without any individual completing more than one survey). This way the differences in responses to the different scenarios can still be very informative, since the specific idiosyncrasies of the particular student sample will be held constant across the different scenarios. For this reason, this approach is followed in the next chapter.

7. Conclusion

None of Smith's (2011) arguments give us a good reason why we cannot pursue the testing of experience machine scenarios proposed at the end of the previous chapter. However, we should learn from Smith's proposed limit that our judgments about exotic thought experiments are not as reliable as they are often thought to be and that the experience machine is a good example of an exotic scenario. Acknowledging this fact does give us reason to be cautious about the results of our tests, but it gives proponents of the experience

machine objection to hedonism equal reason to be cautious. Indeed, Smith's concerns seem to demand just as much caution about reality mattering intrinsically to us being the best explanation for why the vast majority of reasonable people choose reality over a life in an experience machine in Nozick's scenario. The use of student samples was also shown not to be a significant problem when conducted in the right way and compared to the traditional philosophical approach. The next chapter begins my testing (with due caution) of experience machine scenarios.

Chapter 5

Reducing Status Quo Bias and other Confounding Factors in Experience Machine Scenarios

Chapter Summary

As discussed at length in this section so far, Robert Nozick's experience machine thought experiment (Nozick's scenario) is widely thought to provide a knockdown argument against all internalist mental state theories of well-being. Recently, however, it has been argued that Nozick's scenario should not be used in this way because it elicits judgements marred by status quo bias and other irrelevant factors. These arguments all involve reversed experience machine thought experiments, but these scenarios also elicit judgements marred by status quo bias and other irrelevant factors. In this chapter, several experiments that were conducted in order to create and test a relatively bias-free experience machine scenario are reported on. It is argued that this relatively bias-free scenario should be used instead of all previous scenarios when evaluating internalist mental state theories of well-being. Unlike the existing experience machine scenarios, when this new scenario is used to assess internalist mental state theories of well-being it does *not* provide strong evidence to refute or endorse them. The implications for the experience machine objection to hedonism are discussed.

1. Introduction⁶⁷

As discussed at length in this section so far, Robert Nozick's (1974) experience machine thought experiment (Nozick's scenario) is widely thought to provide a knockdown argument against all internalist mental state theories of well-being. As discussed in Chapter 2, however, several philosophers have recently argued that the experience machine thought experiment elicits judgments that are biased in ways that are irrelevant to an isolated prudential value comparison between reality and how our experiences feel to us on the inside.

As discussed in Chapter 3, status quo bias has been identified as a likely candidate for significantly marring judgments about experience machine scenarios. These arguments used to single out status quo bias all involve reversed experience machine thought experiments, but these scenarios also elicit judgements marred by status quo bias and other irrelevant factors. Indeed, as Alex Barber (2011, p. 268) points out: "How successfully these or other tweaks can deal with all confounding factors is unclear." When combined with the point from Chapter 4 that philosophers' evidence in favour of IP2⁶⁸ is drawn from a particularly un-representative sample, we are left in a situation where we really can't be sure whether IP2 is true (i.e. we can't be sure if it's true that the vast majority of reasonable people report preferring reality over a life in an experience machine). We can't be sure of the truth of IP2 for Nozick's scenario because it has never been formally tested (and informal tests have mainly been requests for a show of hands during

⁶⁷ A paper based on this chapter has a 'minor revisions required' judgment from *Philosophical Psychology*.

⁶⁸ IP2: When instructed to ignore their responsibilities to others, the vast majority of reasonable people report preferring reality over a life in an experience machine. (Empirical claim)

philosophy lectures). But this is the least of our problems for assessing the experience machine objection to hedonism because we are also unsure of how much the judgments elicited by Nozick's scenario have been affected by status quo and other irrelevant biases (as discussed in Chapter 2). The key problem to assessing the experience machine objection to hedonism, therefore, is that we can't be sure of the truth of IP2 *for a relatively bias-free experience machine thought experiment*.⁶⁹

In this chapter, I report on several experiments that were conducted in order to test what features of Nozick's scenario are likely to confound the usefulness of the judgments it elicits for assessing the relative prudential value of reality and how our experiences feel to us on the inside. This process involves refining each iteration of experience machine scenario so that the final scenario is relatively bias-free compared to the rest. It is argued that this relatively bias-free scenario should be used instead of all previous scenarios when evaluating internalist accounts of Prudential Hedonism and all other internalist mental state theories of well-being. Unlike the existing experience machine scenarios, when this new scenario is used to assess internalist mental state theories of well-being it does not provide strong evidence to refute or endorse them. The implications for the experience machine objection to hedonism are then discussed, with the conclusion that, on balance, IP2 is probably false (i.e. it is *not* true that the *vast majority* of reasonable people report preferring reality over a life in an experience machine).

⁶⁹ Creating clear criteria for what counts as 'relatively bias free' is no easy task. Clearly 'completely bias free' experiments would be ideal, but they are probably impossible to create in real life. Indeed, the main problem is that it is probably impossible to know exactly how biased any particular experiment's results are. For this reason, a 'relatively bias free' experiment should be understood as an experiment that has *less* bias than the main previous experiments on the problem in question.

2. Analysing Nozick's Scenario

To investigate the potential influences on judgments about Nozick's scenario (discussed in Chapter 2), I presented 125 first year philosophy students with a verbatim version of Nozick's scenario and asked them if they would connect to an experience machine and why.⁷⁰ The sample is from two undergraduate philosophy courses at Victoria University of Wellington. Nozick's scenario and the two questions were written on sheets that were handed out by the tutors of the course. The survey sheets were handed out to the two classes on the same week. This scenario was tested at the same time as my Stranger No Status Quo scenario (discussed later). All participants in all of my experiments only ever took part in one survey and were only ever exposed to one scenario. The two scenarios were randomly distributed amongst the students. The experiment was conducted in April 2011 during the second week of my lectures on unrelated topics (the morality of advertising and philosophy of religion).

Only 21% (26/125) of the students indicated that they would connect to an experience machine in Nozick's scenario. For the 79% of students who didn't want to connect, 'reality, truth, or something related' was the most common justification. However, this justification only accounted for 35% (35/99) of the reasons this group gave for choosing reality over the experience machine.

Surprisingly, 44% (44/99) of the main justifications given by the students who didn't want to connect to an experience machine indicated that they experienced imaginative resistance to the stipulated or implied features of Nozick's scenario. The main justifications displaying imaginative resistance were grouped as follows: 'bad experiences are required to appreciate good experiences or to develop properly' (19%, 19/99); 'you would have no

⁷⁰ See pp. 68–69 of this thesis for the wording of Nozick's scenario as it appeared in my experiment.

autonomy or control in the machine',⁷¹ (14%, 14/99); 'I can't because I have responsibilities to others' (4%, 4/99); 'the machine might break down or not produce great experiences in the future' (3%, 3/99).

Some main justifications also revealed overactive imagination, for example, 'the machine seems scary or unnatural' (8%, 8/99). Other main justifications were consistent with the thought experiment and didn't indicate overactive imagination but were nonetheless irrelevant for evaluating the relative value of reality and how our experiences feel to us on the inside. These main justifications reveal problems with using Nozick's scenario to investigate the the relative value of reality and how our experiences feel to us on the inside. These confounding factors included 'getting out every two years would be depressing' (6%, 6/99). One participant even gave the status quo-endorsing reason that "reality is always preferable, because I have done it". These results show that all of the factors that concern philosophers seem to influence at least some participants' judgments about Nozick's scenario.

Considering that Nozick stated the purpose of his experience machine thought experiment as assessing what, if anything, matters other than how our experiences feel to us on the inside, all of these justifications for not connecting are irrelevant. The fact that 60% (59/99) of the participants that chose reality stated an irrelevant reason as the main justification for their choice gives us reason to believe that Nozick's scenario is not very useful for assessing the relative prudential value of reality and how our experiences feel to us on the inside.⁷² Furthermore, as discussed above, it seems that all of the

⁷¹ To clarify, the justifications related to lack of autonomy in the machine did not include any that referred to being able to make decisions that impacted the real world. The justifications classified in this group involved thinking that the programming of the experience meant that they could not decide how to act during the time they were connected to the machine.

⁷² These stated justifications might not be the real reason for the choice the participants actually made because the main cause of their judgment could have been subconscious (as

existing experience machine scenarios elicit responses marred by some degree of status quo bias, making none of them ideal candidates for this role.

3. Reducing Imaginative Resistance and Other Irrelevant Factors

Since all of the extant experience machine scenarios appear to engender so many biased, irrelevant, and imaginatively resistant judgments, a question arises: is it possible to create an experience machine scenario that is not so heavily affected by these confounding factors? To begin investigating this question, I created a new experience machine scenario that addressed some of the imaginative resistance-related and other concerns raised in the responses to my test of Nozick's scenario. This new scenario was then tested on 93 undergraduate business students. The sample is from three streams of an undergraduate business course at Victoria University of Wellington. My Self scenario and the two questions were written on sheets that were handed out by myself and a colleague (both unknown to the students). The survey sheets were handed out to the different streams on the same day. As a part of the same experiment, three other scenarios were tested at the same time (my Friend, Cousin, and Stranger scenarios). The four scenarios were randomly distributed amongst the students. The experiment was conducted in July 2010 during the second week of term.⁷³ The Self scenario, was worded as follows:

discussed in Chapter 3). However, the justifications given match up to the reasons that some philosophers have predicted would impact participants' judgements and we might expect that some of the judgments of the few participants who confabulated their justification are offset by those of the few participants who couldn't articulate the real cause of their judgement. My seemingly successful attempt to control for these irrelevant factors (discussed below) provides at least weak evidence that some of the respondents' claims about why they didn't want to connect to the experience machine were probably true.

⁷³ The only observable difference between business and philosophy students was their likelihood to provide a justification for their choice. Only 2% (5/249) of the philosophy

It's 2062 and you are riding a hovertube to town. You have been offered a permanent spot in an Experience Machine and are now trying to decide if you should accept.

You have had a go in an Experience Machine before and know that they provide an unpredictable roller-coaster ride of remarkable experiences. When in the machine, you still made autonomous decisions and faced tough situations, such as striving for your goals and feeling grief, but your experiences were vastly more enjoyable and varied. You also recall that, while you were in the Experience Machine, you had no idea that you had gotten into a machine or that your experiences were generated by a machine.

If you accept the spot, then you would stay in an Experience Machine permanently. If you reject the spot, then you would never be offered a spot again. Your life would be the same length in an Experience Machine as it would otherwise have been.

1) Ignoring how your family, friends, any other dependents, and society in general might be affected, and assuming that

students provided no answer to the open-ended "Briefly explain your choice" question, compared to 21% (71/345) of the business students. This difference was stable across the sub-groups of business and philosophy students; none of the differences between the business student groups or between the philosophy student groups were statistically significant (to the 95% or $p = 0.050$ level using a one-tailed Fischer's exact test), whereas the difference between the business student groups and the philosophy student groups were all highly significant ($p < 0.010$). Where relevant, such as when discussing the amount of imaginative resistance demonstrated by the answers to the open-ended "Briefly explain your choice" question, this has been corrected for.

Experience Machines always work perfectly, what is the best thing for you to do *for yourself* in this situation?

Tick only one of these options:

- ☐ You should accept the spot in an Experience Machine
- ☐ You should *not* accept the spot in an Experience Machine

2) Briefly explain your choice:

Nearly 37% (34/93) of participants responding to the Self scenario chose to connect to the experience machine. Comparing the responses to the Self scenario with those to Nozick's scenario reveals a statistically significant 16% increase in participants choosing to connect to the experience machine in the Self scenario.⁷⁴ Furthermore, since participants were asked to explain their choice, the total amount of imaginative resistance and other confounding factors can also be compared. In the Self scenario, 24% (22/93) of participants provided an irrelevant main justification for their choice, compared to 47%

⁷⁴ Using a two-tailed Fischer's exact test, we observe that we can be more than 99% confident that the difference between these two scenarios is not the product of chance ($p = 0.004$). Social scientists are likely to appreciate standardised effect sizes, rather than differences between percentages, so I calculated Cohen's d effect sizes for the differences between the main scenarios (always using 1 as the value for connecting to an experience machine and 0 for not connecting). Cohen's d effect sizes for the differences between all of the main scenarios are reported in Table 2 near the end of this paper. All Cohen's d effect sizes are reported to three decimal places. As a general guide, an effect size of 0.2 is small, 0.5 is medium, and 0.8 is large. The Cohen's d effect size between the Self scenario and Nozick's scenario was $d = 0.352$.

(59/125) in Nozick's scenario.⁷⁵ The majority of this difference is explained by far fewer respondents to the Self scenario justifying their choice with the 'bad experiences are required to appreciate good experiences or to develop properly' and 'you would have no autonomy or control in the machine' reasons. Although we cannot expect the stated main justification for the participants' choices to correlate perfectly with the actual main cause of their choice (as discussed in Chapter 3), these results nevertheless give us good reason to think that the Self scenario is a better test of the relative value of reality and how our experiences feel to us on the inside.

What could we conclude about internalist accounts of Prudential Hedonism and all other internalist mental state theories of well-being based on the results of the Self scenario? It's not clear that we can draw any firm conclusions. Nearly two thirds of respondents to the Self scenario still chose to forego the experience machine, but all of the 24% (22/93) of respondents who provided an irrelevant main justification for their choice chose reality over to an experience machine life. It would be reasonable to expect that at least some of the participants might have chosen to connect to the experience machine if they had not considered the irrelevant factor they cited as justifying their choice.⁷⁶ Recall that one of the main strengths of the experience machine objection to hedonism is that the overwhelming majority agree that it's best to choose reality over an experience machine life. The results of my

⁷⁵ It should be noted that, contrary to best practice, I coded the qualitative responses myself. This is a weakness of these comparisons.

⁷⁶ This assumption will not sound reasonable to anyone who thinks that the vast majority of these justifications are likely to be confabulations—vain attempts by our deliberative mind to explain the verdict of our intuitive cognition. Readers with this belief needn't accept this part of the argument because they are likely to be more heavily swayed by the main argument about the effects of status quo and other biases on our judgments about experience machine scenarios.

experiments on Nozick's and the Self scenarios show that this strength might be lost if the participants adhered to the stipulations of the thought experiments more closely. And this is before the status quo bias has been taken into account.

4. Reducing Status Quo Bias

Assuming status quo bias is interfering with the usefulness of our intuitive responses to most kinds of experience machine thought experiments, a further question arises: is it possible to create an experience machine scenario that is relatively unlikely to elicit responses affected by status quo bias? Furthermore, any successful attempt to reduce the effects of status quo bias on responses to experience machine scenarios will provide additional evidence that they were tainted by status quo bias in the first place.

How could a thought experiment be designed in order to reduce the potential effects of status quo bias? One potential way to mitigate the effects of status quo bias is to reduce any oversensitivity to potential losses. Evolutionary considerations give us good reason to think that we should be loss averse in conditions of uncertainty; losses were probably more deleterious to our ancestors' evolutionary fitness than equivalent gains were beneficial (Chen, Lakshminaryanan, & Santos 2006). Based on this evolutionary consideration, it is reasonable to expect that the less we know and care about someone, the less likely we are to be oversensitive to the risk of them losing something. Indeed, this has been demonstrated in several experiments. The experiments show that when we make decisions for people in conditions of uncertainty our value function for gains and losses becomes increasingly flat the less we care about the people we are making decisions for (Bloomfield *et al.* 2006). That is to say that we are more likely to value equivalent losses and gains equally, just like an economist's rational agent

would, if we are impartial towards the people we are making decisions for. Therefore, I hypothesise that participants will experience less overactive imagination, and imaginative resistance to stipulations about experience machines, if they are judging the value of the potential lives of someone other than themselves. If this hypothesis is correct, then we should expect that we are more likely to think that the unfamiliar (and therefore risky) experience machine life is going to be a better idea for someone we care less about than ourselves; that participants will be more likely to think that a stranger should connect to an experience machine than a friend, and a friend more than themselves.

This hypothesis was tested by adapting the Self scenario to create versions about people that participants should care about less than themselves. These new scenarios involved making the decision about a friend, a cousin, and a stranger. All of these scenarios were tested on first year business students at the same time as the Self scenario. The Stranger scenario reads as follows:

It's 2062 and you are riding a hovertube to town. A stranger sits down next to you, introduces himself as Boris, and tells you that he has been offered a permanent spot in an Experience Machine. Although you would never actually tell Boris your opinion, you are trying to decide if you think he should accept.

You have had a go in an Experience Machine before and know that they provide an unpredictable roller-coaster ride of remarkable experiences. When in the machine, you still made autonomous decisions and faced tough situations, such as striving for your goals and feeling grief, but your experiences were vastly more enjoyable and varied. You also recall that, while you were in the Experience Machine, you had no idea that

you had gotten into a machine or that your experiences were generated by a machine.

If Boris accepts the spot, then he would stay in an Experience Machine permanently. If he rejects the spot, then he would never be offered a spot again. Boris' life would be the same length in an Experience Machine as it would otherwise have been.

1) Ignoring how Boris' family, friends, any other dependents, and society in general might be affected, and assuming that Experience Machines always work perfectly, what is the best thing for Boris to do *for himself* in this situation?

Tick only one of these options:

- ☐ Boris should accept the spot in an Experience Machine
- ☐ Boris should *not* accept the spot in an Experience Machine

2) Briefly explain your choice:

The Stranger scenario is very similar to the Self scenario, the only difference being that the choice is now about whether Boris, the stranger, should connect to an experience machine, instead of participants choosing for themselves. As predicted, participants responding to the Stranger scenario were much more likely to think that connecting to the experience machine was a good idea than those responding to the Self scenario. Over 48% (45/93) of participants responding to the Stranger scenario decided that connecting to an experience

machine made the life in question better, 12% more than participants responding to the Self scenario.⁷⁷

The results for the Friend and Cousin scenarios add further weight to the hypothesis that status quo bias can be reduced by making the scenario about someone the decision-maker cares less about. The Friend scenario makes Boris the participants' (unnamed) friend, instead of a stranger, and the cousin scenario makes Boris the participants' cousin (named Boris). Otherwise the scenarios remain the same as the Self and Stranger scenarios. It is reasonable to expect that most participants will be more oversensitive to the risk of losses to these people in order of how close they are to them, i.e. in the following order: themselves, their unnamed friend, their cousin Boris, a stranger named Boris. Figure 1 (below) shows the differences in the responses to these four scenarios and Table 1 (further below) shows the differences between these scenarios, the statistical significance of those differences (calculated using a two-tailed Fischer's exact test), and the Cohen's *d* measure of the effect sizes.

Table 1 also shows that we can only be very confident (about 95% confident) that two of the differences between these groups are not the product of chance—the differences between the Stranger and Self scenarios and the Cousin and Self scenarios. Fortunately, the difference between the Stranger and Self scenarios is the most important and the other differences still register small effects that are probably not the product of chance.

⁷⁷ Using a two-tailed Fischer's exact test, we observe that we can be about 96.5% confident that the difference between these two scenarios is not the product of chance ($p = 0.035$). The Cohen's *d* effect size between these two scenarios is $d = 0.240$.

Figure 1: Reducing Status Quo Bias in Responses to Experience Machine Thought Experiments by Decreasing the Extent to which the Decision-Maker Cares about the Subject

Table 1: Relative Differences in the Propensity to Report that Connecting to an Experience Machine is Better for Well-being than Living in Reality in Scenarios with Differences in the Extent to which the Decision-Maker Cares about the Subject

Scenarios	Difference in Connect %	Connect #/n & Connect #/n	<i>p</i> -value (3 d.p.)	Cohen's <i>d</i> effect size (3 d.p.)
Stranger & Self	11.8	45/93 & 34/93	0.035	0.240
Cousin & Self	9.7	44/95 & 34/93	0.056	0.198
Stranger & Friend	7.8	45/93 & 26/64	0.106	0.156
Cousin & Friend	5.7	44/95 & 26/64	0.146	0.115
Friend & Self	4.0	26/64 & 34/93	0.376	0.083
Stranger & Cousin	2.1	45/93 & 44/95	0.223	0.041

It might be argued that it is good to be sensitive to risk, especially in important decisions like how to spend the rest of your life. The point of this argument is that eliminating sensitivity to risk is not a virtue of my scenarios,

but a weakness. On the contrary, experience machine scenarios should be designed to eliminate oversensitivity to the risk of losses because the losses and gains should be fixed so as to better isolate the relative value of reality and how our experiences feel to us on the inside. Furthermore, many of the specific kinds of losses envisaged by participants responding to Nozick's scenario (machine failure, loss of autonomy, the experience machine's inability to provide all-important negative experiences, etc.) are either specifically ruled out in the scenario or are clearly irrelevant to comparing reality with how our experiences feel to us on the inside.

Another potential method for reducing status quo bias is to frame all of the options as equally familiar. Nozick's scenario, and all of my scenarios so far, are likely to suffer from the status quo bias because they all frame reality as the status quo, and therefore, as a much less risky option. I hypothesise that framing both options as equally familiar will produce results that are much less affected by status quo bias. This hypothesis was tested by creating a version of the Stranger scenario in which neither an experience machine life nor a real life was framed as the status quo. The Stranger No Status Quo (Stranger NSQ) scenario read as follows:

A stranger, named Boris, has just found out that he has been regularly switched between a real life and a life of machine-generated experiences (without ever being aware of the switches); 50% of his life has been spent in an Experience Machine and 50% in reality. Nearly all of Boris' most enjoyable experiences occurred while he was in an Experience Machine and nearly all of his least enjoyable experiences occurred while he was in reality. Boris now has to decide between living the

rest of his life in an Experience Machine or in reality (no more switching).

You have had a go in an Experience Machine before and know that they provide an unpredictable roller-coaster ride of remarkable experiences. When in the machine, you still made autonomous decisions and faced tough situations, such as striving for your goals and feeling grief, but your experiences were vastly more enjoyable and varied. You also recall that, while you were in the Experience Machine, you had no idea that you had gotten into a machine or that your experiences were generated by a machine.

Boris' life will be the same length in an Experience Machine as it would in reality. No matter which option Boris chooses, you can be sure of two things. First, Boris' life will be very different from *your* current life. And second, Boris will have no memory of this choice and he will think that he is in reality.

1) Ignoring how Boris' family, friends, any other dependents, and society in general might be affected, and assuming that Experience Machines always work perfectly, what is the best thing for Boris to do *for himself* in this situation?

Tick only one of these options:

- ☐ Boris should choose the Experience Machine life
- ☐ Boris should choose the real life

2) Briefly explain your choice:

As expected, when neither reality nor the experience machine were framed as the status quo, participants were more likely to think that Boris should connect to an experience machine than when reality was framed as the status quo. Over 60% (75/124) of participants responding to the Stranger NSQ scenario thought Boris should connect to an experience machine, which is 12% more than in the Stranger scenario.⁷⁸ Since the only difference between the scenarios was the framing of the status quo, there is good reason to think that the main cause of this difference was the reduction of the endowment aspect of the status quo bias—the overvaluing of what we have and what we know.

By comparing the Self scenario with the Stranger NSQ scenario, we can see a 24% increase in the proportion of participants indicating that they think an experience machine life is better than a real life.⁷⁹ Assuming that the main reason for this dramatic increase is a reduction in the effects of status quo bias, then we have good reason to believe that scenarios that frame reality as the status quo are likely to heavily bias the results in favour of reality. Indeed, this result supports De Brigard's (2010) claim that status quo bias renders Nozick's scenario relatively useless. If the results of my experiment on Nozick's scenario were adjusted to take the likely effects of status quo bias into account, nearly half of the participants would have chosen to connect to an experience machine. Again, such a result would take away the main strength of the experience machine objection to hedonism—the widespread

⁷⁸ Using a two-tailed Fischer's exact test, we observe that we can be about 97.5% confident that the difference between these two scenarios is not the product of chance ($p = 0.025$). The Cohen's d effect size between these two scenarios is $d = 0.243$.

⁷⁹ Using a two-tailed Fischer's exact test, we observe that we can be more than 99.9% confident that the difference between these two scenarios is not the product of chance ($p = 0.000$ to 3d.p.). The Cohen's d effect size between these two scenarios is $d = 0.491$.

agreement that reality is better than a life in an experience machine would be lost.

5. The Experience Machine is Dead, Long Live the Experience Machine!

Assuming that the three main hypotheses put forward so far are well supported by the evidence from my experiments, we can explain the three main reasons for the difference between the high connection rate for the Stranger NSQ scenario and the low rate for Nozick's scenario.⁸⁰ Figure 2 (below) shows the differences in the responses between my main scenarios.

Figure 2: Reducing Status Quo Bias and Other Confounding Factors in Experience Machine Thought Experiments

First, we can see that reducing the impact of irrelevant factors, such as worries about loss of autonomy and machine failure, appears to make about 16% difference (compare Self to Nozick's). Second, we can see that making the

⁸⁰ I say 'assuming' here because it is quite possible that I have inadvertently elicited some additional biases or other irrelevant factors in my new scenarios that were not existent in Nozick's scenario.

choice on behalf of a stranger instead of ourselves appears to make about 12% difference (compare Stranger to Self). As was argued, this is presumably because we are less irrationally loss averse (oversensitive to the risk of loss in conditions of uncertainty) when deciding on behalf of people we care less about. Third, we can see that neutralising what was framed as the status quo (what we are familiar with) appears to make about 12% difference (compare Stranger NSQ to Stranger). Finally, the total difference between the Stranger NSQ scenario and Nozick's scenario is about 40%, a large and very highly statistically significant difference. The differences between the scenarios, the statistical significance of those differences (calculated using a two-tailed Fischer's exact test), and the Cohen's *d* measure of effect sizes for the differences are displayed in Table 2 below.

Table 2: Relative Differences in the Propensity to Report that Connecting to an Experience Machine is Better for Well-Being than Living in Reality in the Main Scenarios

Scenarios	Difference in Connect %	Connect #/n & Connect #/n	<i>p</i> -value (3 d.p.)	Cohen's <i>d</i> effect size (3 d.p.)
Stranger NSQ & Nozick's	39.7	75/124 & 26/125	0.000	1.950
Stranger & Nozick's	27.6	45/93 & 26/125	0.000	0.603
Stranger NSQ & Self	23.9	75/124 & 34/93	0.000	0.491
Self & Nozick's	15.8	34/93 & 26/125	0.004	0.352
Stranger NSQ & Stranger	12.1	75/124 & 45/93	0.025	0.243
Stranger & Self	11.8	45/93 & 34/93	0.035	0.240

The combination of the large differences between the responses to my scenarios and Nozick's scenario with the higher level of imaginative

resistance to Nozick's scenario provides good reason to think that Nozick's scenario is not the best scenario to choose when trying to argue about the relative value of reality and how our experiences feel to us on the inside. So it seems that Kolber (1994), De Brigard (2010), and Weijers (forthcoming-a) were correct in calling for an end to the use of Nozick's scenario for evaluating internalist mental state theories of well-being. Indeed, we now have good reason to think that Nozick's scenario elicits intuitions that are about 40% off the mark, and since alternative, less biased, scenarios are available, the use of Nozick's scenario as a knockdown argument against internalist accounts of Prudential Hedonism or any other internalist mental state theories of well-being should be well and truly over. The experience machine thought experiment (based on Nozick's scenario) is dead!

Which experience machine scenario should be used for evaluating internalist mental state theories of well-being then? The Stranger NSQ scenario addresses the question of the relative prudential value between reality and how our experiences feel to us on the inside while making use of the memorable, mysterious, and intriguing concept of the experience machine. Furthermore, the Stranger NSQ scenario appears to be much less affected by the status quo bias and other irrelevant factors than Nozick's version is, and without being much more complicated. All of the existing experience machine scenarios, except for the Stranger NSQ scenario appear to be non-negligibly affected by status quo bias and other irrelevant factors. Therefore, although the Stranger NSQ scenario is possibly affected by some confounding factors, we have good reason to believe that it is relatively bias-free compared to all of the other scenarios. For these reasons, the Stranger NSQ scenario should be used, instead of any of the existing experience machine scenarios, for investigating the relative prudential value of reality

and how our experiences feel to us on the inside. Long live (the Stranger NSQ version of) the experience machine!

6. Conclusion and Implications

So, what could the results of the Stranger NSQ scenario tell us about the relative value of reality and how our experiences feel to us on the inside? And what should we think about internalist accounts of Prudential Hedonism and all other internalist mental state theories of well-being? Since 60% (75/124) of the participants thought that Boris should connect to an experience machine in the Stranger NSQ scenario, we can conclude that there is probably not widespread agreement about the relative value of reality and how our experiences feel to us on the inside. That is to say that IP2 is probably false; it is probably *not* true that the vast majority, or any kind of majority, of reasonable people report preferring reality over a life in an experience machine. Recall that the great power of the experience machine objection to hedonism (when based on Nozick's scenario) was that *nearly everyone agreed* that connecting to an experience machine was a bad choice even though a lot more enjoyment was offered by a life connected to an experience machine. Many people then inferred that directly connecting with reality must be the reason for this because reality was the only obvious difference between the lives on offer. The results of the new scenarios presented here provide good evidence that the widespread agreement about Nozick's scenario was guided at least as much by status quo bias and other irrelevant factors as it was by the value of reality.⁸¹

⁸¹ For evidence of this, compare the approximate 40% change in respondents' choices between my test of Nozick's scenario and my Stranger NSQ scenario to the proportion of respondents to Nozick's scenario whose main justification for not connecting to the experience machine was 'reality, truth, or something related' — 28% (35/125).

Therefore, contemplation of the experience machine objection to hedonism should no longer give us a *prima facie* reason for rejecting internalist accounts of Prudential Hedonism and all other internalist mental state theories of well-being. The results of the Stranger NSQ scenario certainly do *not* provide endorsement for any internalist mental state theories of well-being either, however. Nearly 22% (27/124) of participants still mentioned that having a veridical connection between their experiences and the cause of those experiences matters to them. For internalist accounts of Prudential Hedonism or any other internalist mental state theory of well-being to be true, anything that is outside of the internal aspects of our experiences (such as whether they are, unbeknownst to us, caused by a machine) has no intrinsic value and should not matter to us over and above the positive experiences it might lead to. Contemplation of the experience machine, therefore, still produces some judgements that constitute evidence against internalist accounts of Prudential Hedonism. However, since these judgments are now a minority, they should be considered defeasible instead of decisive evidence against internalist accounts of Prudential Hedonism and all other internalist mental state theories of well-being. Consider what a reworking of the experience machine objection to hedonism based on my Stranger NSQ scenario would look like. The best explanation for 40% of respondents reporting a preference for reality over a life in an experience machine might not be that reality really matters to the vast majority of reasonable people, especially if only 22% of respondents even mentioned tracking reality as the justification for their choice.

The important points here are: First, that Nozick's experience machine scenario should no longer be considered to provide conclusive or strong evidence that hedonism and all other internalist mental state theories of well-being are false. Second, that it is reasonable to believe that the experience machine objection to hedonism is very unlikely to be true. And, finally, that it

might still be true that reality matters intrinsically to the vast majority of people, regardless of what people report when surveyed about experience machine scenarios. An argument against the plausibility of Prudential Hedonism might simply run as follows:

- P1. Reality matters intrinsically to the vast majority of reasonable people. (Empirical claim)
- P2. The best explanation for reality mattering intrinsically to the vast majority of reasonable people is that reality has intrinsic prudential value
- P3. Inference to the best explanation: If a hypothesis is the best explanation of an observation, then it is rational to believe that hypothesis is true. (Standard methodological premise)
- C1. Therefore, it is rational to believe that reality has intrinsic prudential value. (*Modus ponens* P2, P3)
- P4. If internalist Prudential Hedonism is true, then the internal aspects of pleasure and pain are the only things of intrinsic prudential value (or disvalue) in a life. (Stipulated definition)
- C2. Therefore, it is rational to believe that internalist Prudential Hedonism is false. (*Modus tollens*, C1, P4)

Of course, P1 and P2 of this argument are debateable. My point is to demonstrate that our intuitions about the value of reality might still form the basis of an objection to Prudential Hedonism without the need to call on the dubious influence of exotic thought experiments like Nozick's experience machine scenario. Having said this, proponents of the argument above would probably have to provide some sort of motivation for P1 (reality mattering intrinsically to the vast majority of reasonable people) and philosophers find it hard not to use thought experiments in situations like this. The findings of

this section indicate that it might be methodologically unproblematic to use a thought experiment to motivate P1, but that, if we do rely on thought experiments to motivate P1, we should be very careful in doing so. Indeed, to be somewhat confident that the thought experiment is isolating the right prudential value comparison, it should be tested on a sample group alongside several other versions of the same scenario with relatively minor tweaks. Perhaps, using this approach, will allow for a suitable thought experiment to support P1. That thought experiment might even be an experience machine scenario, but I doubt it. Based on the discussion about intuitive cognition in Chapter 3, it seems that more realistic thought experiments might be better suited to this role.

Another important implication of the results of this section is that the way hedonism and well-being are taught in many introductory philosophy classes should change. It is common practice, in introductory philosophy classes that discuss well-being, to describe Prudential Hedonism and then promptly dismiss it based on a quick discussion of Nozick's experience machine. The main problem is not that students are inclined away from internalist accounts of Prudential Hedonism, and all other internalist mental state theories of well-being, by the way the experience machine is taught (because there are several other fairly good arguments against the plausibility of these theories, as discussed in Chapter 1). Rather, the main problem is that students are taught that unrealistic thought experiments can constitute knockdown philosophical arguments without warning them about the many biases and other irrelevant factors that might be affecting our judgements about these kinds of scenarios.

And, as for the minor problem of swaying the students' opinion about internalist accounts of Prudential Hedonism, a natural experiment occurred during my tests, which demonstrates the potential influence this kind of first exposure to experience machine scenarios can have on philosophy students.

About half of the students in my tests of Nozick's and my Stranger NSQ scenarios had *not* been exposed to the experience machine before and the rest had listened to a lecturer discuss it for about 10 minutes. The lecturer made it clear that most people decline the opportunity to connect to an experience machine and that this shows that internalist accounts of Prudential Hedonism and all other internalist mental state theories of well-being are probably false.

There was a statistically significant difference in the two groups' responses to Nozick's scenario—15% (10/67) of the group who had been lectured to about the experience machine would connect to the machine compared to 28% (16/58) of the group who had not been lectured to about it. Using a two-tailed Fischer's exact test, we can be about 97% confident that this difference was not caused by chance ($p = 0.032$). Being exposed to a short discussion about the experience machine in a lecture could influence students' judgments by encouraging them to think more deeply about the matter, but a more likely explanation for this difference is the combined pressures to agree with the majority and with the lecturer (an authority figure). Regardless of the reason, these data give us very good reason to think that we should be careful how we teach the experience machine because it does seem to make a difference to how future philosophers will think about well-being.

Interestingly, there was *not* a statistically significant difference in the two groups' responses to my Stranger NSQ scenario; 59% (40/68) of the group who had been lectured on the experience machine would connect to the machine compared to 63% (35/56) of the group who had not been lectured on it. Using a two-tailed Fischer's exact test, we observe $p = 0.204$. And with a one-tailed Fischer's exact test, we observe $p = 0.409$.⁸² Therefore, being

⁸² One-tailed Fischer's exact tests are to be used when there is *no* prior reason to expect a difference between the two sample groups and two-tailed tests are to be used when there is a prior reason to expect a difference between the two sample groups.

exposed to a short discussion about the experience machine in a lecture appeared to have little to no influence on the Stranger NSQ scenario. Perhaps the Stranger NSQ scenario was different enough from Nozick's scenario to avoid the large and statistically significant effect on the choices of the participants that responded to Nozick's scenario after listening to a lecturer discuss it for about 10 minutes.

Instead of misleading students, philosophers would be better off teaching how to assess the power of thought experiments as evidence for arguments. Indeed, it should be considered a vital component of metaphilosophy and introductory philosophy courses. Since the differences in responses elicited by Nozickian, reversed, and Stranger NSQ scenarios are so large, the experience machine thought experiment (in all of its guises) would be an excellent example for this kind of instruction. Long live (all versions of) the experience machine!

Part 2

Part 2 of this thesis focuses on practice. In particular, Part 2 is concerned with how considerations of happiness can inform certain specific areas of practice and also help us to understand *what we should do* in those circumstances.

Unlike in Part 1, there is not one specific question that acts as a focal point for all of the component chapters in Part 2. Each chapter in Part 2 identifies a separate question related to happiness, and none rely on the conclusions of the other chapters. Chapter 6 shows how considerations of happiness can be useful for guiding us in deciding what to do in a very personal endeavour; the quest for the meaning of life. Chapter 7 demonstrates how considerations of happiness can help us decide what to do in an extremely public endeavour; apportioning the responsibilities to deal with problems likely to be caused by rapid climatic change. The Postscript for Policymakers demonstrates how general philosophical considerations can help us to decide what to do regarding questions about whether, and how, happiness should guide policymaking.

Chapter 6 argues that an optimistic view about scientific and technological progress allows for two interesting new theories for the meaning of life debate. It also discusses what people with certain kinds of beliefs might want to do to achieve true meaning in life. One of these novel theories posits that causing there to be infinite happiness can be a way to lead a truly meaningful life. The discussion in Chapter 6 is not designed to encourage all readers to believe the new theories proposed, rather it is designed to provide a plausible option to people with specific existing beliefs about something else they can believe to relieve an epistemic tension they might have and to provide a guide for how they should live if they were to adopt this further belief.

Chapter 7 demonstrates how considerations of human happiness can justify why a particular set of distributive principles is the fairest way to distribute the burdens associated with adapting to and mitigating the potentially devastating effects of rapid climactic change. The main competing distributive principles discussed in the philosophical literature on climate ethics have often been thought to arrive at the same recommendations when applied to the real world. And, generally speaking, they have done. The discussion in Chapter 7, however, reveals that how the choice of distributive principle is justified can affect the specific policy recommendations that result. Based on the considerations of human happiness, Chapter 7 includes fairly specific policy recommendations about what governments should do about climate change.

This thesis also includes a Postscript for Policymakers, which is positioned after the main conclusion of the thesis. While similar to Chapters 6 and 7, in that it is a stand-alone piece of applied philosophy, the Postscript for Policymakers is different enough in its method from the rest of the thesis to warrant its partial separation from the main chapters. Compared to the other chapters (especially Chapters 2 to 7), the Postscript for Policymakers takes a much higher-level approach—it seeks to provide general answers to two very broad questions—and should be read more as a potential direction for future research than a defense of a specific thesis.

Given its broader scope and different intended audience, the Postscript for Policymakers does not include in-depth discussion of all likely objections. The two questions addressed in the Postscript for Policymakers are: should policymakers use findings from the science of happiness to guide their policy decisions, and how can they best do that? The Postscript for Policymakers relies in part on a philosophy of social science approach in addressing these questions, arguing that some of the claims of social scientists about the science

of happiness are not supported by their methods or the results they discuss. The Postscript for Policymakers concludes that findings from the science of happiness should be used to guide policymaking (with several qualifications), and it provides recommendations for how best to do this.

Chapter 6

Theories of the Meaning of Life: Optimistic Naturalism and Infinite Happiness

Chapter Summary

Naturalist theories of the meaning of life are often criticised for not setting the bar high enough for what counts as a meaningful life. Tolstoy's version of this criticism is that Naturalist theories do not describe truly meaningful lives because they do not require that we connect our finite lives with the infinite. Other versions include Naturalist theories not requiring the meaning to be objective and not being able to resolve the Absurd—the vast difference between how meaningful our actions and lives appear from subjective and objective viewpoints. These perceived weaknesses of Naturalist theories might cause would-be Naturalists to adopt Nihilism or Supernaturalism, as Tolstoy did. This chapter defends a novel view, Optimistic Naturalism, in order to refute these criticisms. Optimistic Naturalism is the idea that scientific and technological advancement will allow us to lead truly meaningful lives through the infinite consequences of our purposeful actions. Purposeful actions with infinite consequences are argued to connect our finite lives with the infinite in a way that is sufficient to confer subjective and objective meaning, and therefore true meaning, on our lives. In achieving this, Optimistic Naturalism provides one avenue for resolving the Absurd and constitutes a refutation of the criticisms mentioned above.

Furthermore, the plausibility of Optimistic Naturalism allows for an interesting new version of the Utilitarian Theory of Meaningfulness (UTM), which holds that a life is meaningful to the extent that it increases utility. My

Infinite Happiness Theory of Meaningfulness is shown to adequately defend itself against currently unanswered counterexamples to UTM in interesting ways.

1. Introduction⁸³

In the analytic debate on the meaning of life, theories are predominantly categorised according to a tripartite taxonomy: Supernaturalism; that more than a purely physical universe is required for life to be meaningful, Naturalism; that a meaningful life is possible in a purely physical universe, and Nihilism; that life cannot be meaningful.⁸⁴ A major criticism of Naturalist theories is that they offer less meaning than Supernaturalist theories. Tolstoy was amongst the first to articulate a version of this criticism. He identified the inability of Naturalistic theories to offer any kind of meaningful connection with the infinite.⁸⁵ In response to these criticisms, this chapter has three main aims:

- 1) To explain and argue for the plausibility of a novel Naturalistic position for the meaning of life debate, Optimistic Naturalism, and its three central principles.

⁸³ A paper based on this chapter has a 'major revisions required' judgment from *Sophia*.

⁸⁴ For more background on the debate and the established positions, see any of these useful reviews: (Metz 2001; 2002; 2007; 2008).

⁸⁵ "I understand that... the answer given by rational knowledge was only an indication that the answer might be got if the question were... the question of the relation of the finite to the infinite. I also understand that, no matter how irrational and monstrous the answers might be that faith gave, they had this advantage that they introduced into each answer the relation of the finite to the infinite." (Tolstoy 2000, p. 17).

- 2) To argue that Optimistic Naturalism is a counterexample to Tolstoy's criticism on the basis that it is a Naturalist theory that *does* allow for us to meaningfully connect with the infinite.
- 3) To argue that Optimistic Naturalism can resolve the tension of the Absurd—the vast difference between how meaningful our actions and lives appear from the subjective and objective viewpoints. Optimistic Naturalism can resolve the Absurd by identifying the objective meaning present in subjectively meaningful activities that have infinite consequences.

This chapter also has the supplementary aim of showing that if Optimistic Naturalism is plausible, then other philosophical debates about the meaning of life might be affected. Metz's (2003a) criticisms of the Utilitarian Theory of Meaningfulness are used as an example.

This chapter begins with some background on Tolstoy's criticism and the argument that contemplation of the Absurd causes problems for Naturalist theories of the meaning of life. Optimistic Naturalism is then explained and argued to be a refutation of these criticisms. The bulk of this chapter is devoted to explaining and arguing for two of Optimistic Naturalism's three core principles: Infinite Consequence; purposeful actions that have an infinite consequence are sufficient to make a life truly meaningful, and Scientific Optimism; continual scientific and technological advancement will allow our actions to have infinite consequences. This is followed by a brief discussion on how groups with various beliefs should act in light of Optimistic Naturalism. Finally, the interesting new version of the Utilitarian Theory of Meaningfulness (UTM) that the plausibility of Optimistic Naturalism enables is discussed. This new version of the UTM, the Infinite Happiness Theory of

Meaningfulness, is shown to adequately defend itself against currently unanswered counterexamples to UTM in interesting ways.

2. Tolstoy's Criticism and the Absurd

Tolstoy's criticism of Naturalism arose because he began questioning the importance of his actions and the very meaning of his life (Tolstoy 2000). Tolstoy was a dedicated Naturalist who became paralysed by what Nagel (1986) and others have referred to as 'the Absurd'—the vast difference between how meaningful our actions and lives appear from the subjective and objective viewpoints. Initially, Tolstoy thought that his writing and his effects on the people around him were making his actions meaningful. Subsequent recognition of the Absurd led Tolstoy to reject his subjective view—that his actions and life as a whole were meaningful. He came to believe that all Naturalistic views entailed that his life, and all the potentially meaningful products of his actions, would eventually be destroyed.⁸⁶ This led Tolstoy to conclude that Naturalist theories could not provide any meaningful answers to the question of the meaning of life.⁸⁷ Specifically, Tolstoy was looking for a theory of the meaning of life that acknowledged the importance

⁸⁶ "All my affairs, no matter what they might be, would sooner or later be forgotten, and I myself should not exist." (Tolstoy 2000, p. 13). "You are an accidentally cohering globule of something. The globule is fermenting. This fermentation the globule calls its life. The globule falls to pieces." (Tolstoy 2000, p. 15).

⁸⁷ "My situation was a terrible one. I knew that I should not find anything on the path of rational knowledge but the negation of life, and there, in faith, nothing but the negation of reason, which was still more impossible than the negation of life." (Tolstoy 2000, p. 16). "I sought in all the sciences, but far from finding what I wanted, became convinced that all who like myself had sought in knowledge for the meaning of life had found nothing." (Tolstoy 1940, p. 23).

of connecting with the infinite in a meaningful way, something none of the extant Naturalist theories offered.⁸⁸

In the end, Tolstoy's recognition of the Absurd and his disappointment with Naturalist theories led him to reject Naturalism and adopt a Supernaturalist theory of the meaning of life. He became a Christian of sorts.⁸⁹ Indeed, contemplation of the Absurd might be a major cause of dissatisfaction with Naturalist theories of the meaning of life. Belief that Naturalism cannot offer the things that make a life truly meaningful might lead those who are tolerant of supernatural beliefs to opt for Supernaturalism and those who are intolerant of them to opt for Nihilism (Boylan 2008; Metz 2008).

The act of mentally stepping back from our lives allows us to question our actions, motives and plans. Many Naturalists believe that from this objective viewpoint we can better evaluate how meaningful our actions and our lives are. From this once-removed position, we are supposed to be able to see obvious truths about what has meaning. We should see, for example, that our watching every episode of *The Simpsons* ten times in reverse to enter the Guinness Book of World Records is not as meaningful as lovingly raising a family. But, we can step back again. From this twice-removed viewpoint, we can see that our subjective standards of evaluation are still being used to decide which actions confer meaning on our lives. However, it could be

⁸⁸ While he was dealing with his dilemma, Tolstoy's conception of connecting with the infinite in a meaningful way involved living forever or creating something that persists infinitely (Flew 1963, p. 113). However, Tolstoy's conception later changed to be explicitly supernatural; "What real result will come of my life?—Eternal torment or eternal bliss. What meaning has life that death does not destroy?—Union with the eternal God: heaven." (Tolstoy 1940, p. 50). In his most recent writings, such as *What I Believe*, Tolstoy stopped discussing the infinite despite still discussing the meaning of life (Flew 1963, p. 117).

⁸⁹ Tolstoy often used Christian terminology in his later works of non-fiction and studied the Gospels extensively, but he also studied several other religions (Flew 1963, p. 116).

argued that every reasonable person would agree that lovingly raising a family is a meaningful activity.⁹⁰ Even if this claim about what other people would believe is correct, it is not obviously enough to make that action objectively meaningful. By taking another step back, we can see that the evaluative measure used is still the subjective (and therefore contingent) beliefs of humans, albeit in aggregated form. Relevantly similar aliens, or an omniscient being, might be able to point out the bias all humans are suffering from.

When we continue stepping further and further back from our subjective viewpoint in this way, two notions become salient. The first is that it is not clear our actions have the meaning we attribute them. The second is that, even if our actions are meaningful, they have a finite amount of meaning. Evaluating a finite amount of meaning from a great distance (from the far side of the universe, say) makes it seem vanishingly small—virtually meaningless from a distant objective viewpoint. Consideration of these two points (subjective meaningfulness and objective lack of meaningfulness) puts one deep in the grip of the absurdity of natural human existence. Being in the grip of the Absurd in this way might encourage would-be Naturalists to scale up their quest for objective meaning and adopt Supernaturalism or give up and opt for Nihilism.

According to Optimistic Naturalism, our actions and lives can achieve objective meaning by having an infinite consequence. Furthermore, this objective meaning can become ‘true meaning’ when coupled with the correct subjective states. For Naturalists like the mid-life Tolstoy, who are deeply affected by the Absurd, Optimistic Naturalism can provide guidance by explaining which actions should be seen as meaningful from both the subjective and objective viewpoints. For Supernaturalists and Nihilists who

⁹⁰ Darwall argues for a theory that operates along these lines (1983, chap. 11–12).

wanted to have an infinite consequence, but could not conceive of how to achieve it with Naturalism, Optimistic Naturalism explains how this can be achieved in a natural universe. For the rest of the Supernaturalists and Nihilists, and for the Naturalists who are satisfied with limiting themselves to a life that seems meaningful from the subjective viewpoint only, Optimistic Naturalism should be viewed as a challenge to their position. To these groups, the plausibility of Optimistic Naturalism would mean that they could be leading a more meaningful life by following its guidance (in some cases in addition to the guidance of their existing beliefs). The *implausibility* of Optimistic Naturalism, however, would also be of interest to these groups, since they could use it to demonstrate the futility of arguing for a ‘truly’ meaningful Naturalist theory of the meaning of life.

3. Optimistic Naturalism

Optimistic Naturalism is the idea that scientific and technological advancement will allow us to lead truly meaningful lives through the infinite consequences of our purposeful actions. Since Optimistic Naturalism holds that a meaningful life does not require more than a purely physical universe, it is a type of Naturalism. However, Optimistic Naturalism’s detailed account of how we can meaningfully connect our finite lives with the infinite stands it apart from the existing Naturalist theories.⁹¹ If plausible then, Optimistic Naturalism provides an avenue for refuting the criticisms of Tolstoy and others that Naturalist theories do not provide accounts of the meaning of life that are meaningful enough. Optimistic Naturalism also provides us with the

⁹¹ The idea that it is in principle possible to meaningfully connect with the infinite in a natural universe has been hinted at in the literature (e.g. Metz 2003b, p. 171), but has never been discussed in any detail.

chance to resolve the Absurd by explaining how some of our subjectively meaningful actions could also be objectively meaningful.

Optimistic Naturalism is based on the following three principles:

Naturalism: meaningful lives are possible in a purely physical universe.

Infinite Consequence: purposeful actions with infinite consequences are sufficient to make a life truly meaningful when accompanied by the appropriate beliefs.

Scientific Optimism: continual scientific and technological advancement will allow our actions to have infinite consequences.

Belief that all three of these principles are true makes one an Optimistic Naturalist. Optimistic Naturalism is intended to be a hybrid (subjective-objective) Naturalist theory. According to objective theories, the amount of meaning in a life is determined by factors external to the life in question. This meaning-determining role played by external factors in objective theories makes it possible for a hypothetical omniscient observer to inform someone if he or she is right or wrong about how meaningful their life is. According to subjective theories, factors within the mind of the individual are what make that person's life meaningful. Hybrid theories require the union of objective and subjective features for a life to be meaningful, as eloquently captured by the phrase "meaning arises when subjective attraction meets objective attractiveness" (Wolf 1997, p. 211). For example, a subjective theory might make the amount of meaning in a life depend on the extent to which a person thought their most important desires were satisfied. A similar objective theory

might make the amount of meaning in a life depend on the extent to which an objective list of 'ideal' desires are satisfied for that person, regardless of whether they actually have those desires. A corresponding hybrid theory might dictate that the amount of meaning in a life is proportional to the number of satisfied desires from the 'ideal' list that the person also happens to consider important.

Optimistic Naturalism is a hybrid theory because it requires the following subjective and objective elements: the person must believe that having an infinite consequence makes an action meaningful, the action must be performed with the intention that it could have an infinite consequence, and the action must have an infinite consequence. Assuming that a truly meaningful life can be defined as one that includes subjective and objective meaning derived from the same actions,⁹² Optimistic Naturalism provides us with the chance to lead a truly meaningful life. Theories of what a truly meaningful life is, defined in this way, allow us to resolve the Absurd by helping us recognise which of our actions are subjectively, objectively, and truly meaningful. This definition also helps us to better recognise the different types of meaning that a life might accrue and their relative importance. According to Optimistic Naturalism then, if a life of someone who believes that having an infinite consequence confers meaning includes any actions that were intended to and do have an infinite consequence, then that life can be considered truly meaningful.

3.1 Naturalism

Naturalism is the idea that a meaningful life is possible even if all that exists is purely physical. In line with the majority of recent analytic work on the meaning of life, this chapter assumes, rather than argues, that Naturalism is

⁹² This is argued for later.

independently plausible (Metz 2007, p. 203). A typical assumption made by Naturalists is that the meanings available in their accounts of the meaning of life are meaningful *enough*. Not all Naturalists would agree that 'true meaning' (as defined above) is necessary to achieve a meaningful enough life. Some naturalists, such as Taylor (2000), find subjective meaning sufficient to make a life meaningful. In order to resolve, rather than dissolve, the Absurd and refute, rather than ignore, the criticisms from above, however, a higher standard of meaning is required. Since this chapter is designed to refute these criticisms, only truly meaningful lives (ones in which subjective and objective meaning are correctly aligned) will be considered meaningful enough.

3.2 Infinite Consequence

Infinite Consequence is the idea that when a purposeful action with infinite consequences is accompanied by the appropriate beliefs, that action is sufficient to make the life of the actor truly meaningful. Within this broad conception of Infinite Consequence, there are three important components. First, purposefully performing an action with the beliefs that the action is meaningful and that it will have an infinite consequence is sufficient to make a life subjectively meaningful. Second, actions with infinite consequences can also be objectively meaningful, even from a distant objective viewpoint. Third, the correct alignment of subjective and objective meaning can make a life truly meaningful. The first two components of Infinite Consequence will be argued for. Then, the third component will be stipulated and motivated.

From the subjective viewpoint, how meaningful an action is depends on the mental states of the individual. As a result, the subjective viewpoint permits practically any action to be seen as meaningful to the actor. Some subjective theories of the meaning of life require that actions must comply with a

person's carefully considered or innermost desires to be able to confer meaning upon that person's life (Metz 2008). But even theories with these restrictions sanction avid-*enough* marble collectors to meaningfully use their time hunting through second hand detritus at garage sales. Subjective theories of the meaning of life are often criticised for allowing such ostensibly inane actions to confer meaning. It could be argued that avid marble collectors would recognise their hobby as absurdly meaningless if they examined it from the objective viewpoint. But many people who spend their whole lives absorbed in seemingly fruitless hobbies admit that their hobbies seem objectively meaningless and nonetheless continue to find them subjectively meaningful. It is precisely this feature of the Absurd that makes it so difficult to resolve for some people.

A more constructive approach to evaluating the meaning of such actions draws on the distinction between subjective, objective, and true meaning discussed above. If an action seems meaningful to the corresponding actor, no matter how objectively meaningless it might be, then it should be considered to confer subjective meaning on the life of the actor. Subjectively meaningful actions should only be considered truly meaningful, however, if they are also objectively meaningful. Assuming this tiered approach to evaluating meaning is reasonable, there are few, if any, grounds for arguing against any sort of actions being able to confer subjective meaning on people with suitably organised mental states.

The principle of Infinite Consequence stipulates that the coincidence of the following three conditions is sufficient to make an action subjectively meaningful: the action is hoped to have an infinite consequence, the action is accompanied by the belief that it is possible it will have an infinite consequence, and the action is accompanied by the belief that it is meaningful. Since these conditions are sufficient for subjective meaning (as opposed to

being necessary), the principle of Infinite Consequence remains silent about other potential ways to achieve subjective meaning.

Given that attaining subjective meaning in life is usually understood to be merely a matter of believing that one's life is meaningful, the principle of Infinite Consequence's addition of two further sufficiency conditions makes subjective meaning harder to obtain. However, the addition of the conditions that the action be both purposefully performed and accompanied by the belief that the action could have an infinite consequence can be easily motivated.

The kinds of actions that nearly all of us find subjectively meaningful are those that we purposefully carry out with a specific goal or consequence in mind. As we teach and care for our children, we hope that they grow up to be healthy and happy. We consider putting time and effort into lovingly raising children in this way as a particularly meaningful activity. If the intention for our actions to achieve that consequence is removed, however, the action loses its subjective meaning. If our children happen to grow up to be healthy and happy without our purposefully having tried to facilitate that outcome, then that aspect of our life seems no more meaningful. Our belief that the goal or consequence can be achieved also affects the meaningfulness of the corresponding action. Overzealous parenting might seem subjectively meaningful to the overbearing parents at the time because the goal of happy children appears achievable. When the children turn out to be miserable, however, their parents will question if their overzealous actions were as meaningful as they thought them to be.

Something that could make purposeful actions such as these more subjectively meaningful is if the intended consequence or goal is increased. Imagine an aid worker is making his life more subjectively meaningful by purposefully increasing the well-being of those in need. If the aid worker can help 10,000 needy people instead of 100, then his actions would be more

subjectively meaningful. Now, imagine a young philosopher who considers her action of writing a book to be subjectively meaningful because it will bring her fame and fortune. The more people who buy and read the philosopher's book, the more subjectively meaningful her action of writing the book will be. Those who consider some of their purposeful actions to be subjectively meaningful because they achieve a finite goal should also accept that having an infinite consequence can be subjectively meaningful. For example, if the philosopher's book becomes a central text in philosophy *for infinity*, then her action of writing it becomes even more subjectively meaningful.

As mentioned, none of this excludes other conditions that might permit actions to confer subjective meaning on a life. However, the principle of Infinite Consequence is concerned with these specific conditions because they can lead to a *truly* meaningful life. In light of this, Infinite Consequence's stipulation that purposeful actions with infinite consequences are sufficient to confer subjective meaning when accompanied by the appropriate beliefs is clearly plausible.

More importantly, it is also plausible that a life could be made objectively meaningful by containing actions that also convey subjective meaning along the lines above. The second component of Infinite Consequence stipulates that purposeful actions with infinite consequences are sufficient to make a life objectively meaningful. As mentioned earlier, many subjectively meaningful actions seem meaningless when examined from the objective viewpoint, creating an absurd asymmetry between how meaningful our actions *should* seem objectively and how they *actually do* seem to us subjectively. Indeed, the further we step back from the subjective viewpoint, the more meaningless our subjectively meaningful actions become. No matter how far we step back

from our subjective viewpoint, however, infinite consequences do not vanish into meaninglessness.

By taking a step back from the subjective viewpoint of the philosopher who desired fame and fortune, her goals seem to be idiosyncratic, tainted by subjective values, and unlikely to confer objective meaning. Many of the existing Naturalist objective theories of the meaning of life seem to have been developed from a position only one step back from the subjective viewpoint — a *near* objective viewpoint. Consider the following claims about which kinds of actions confer objective meaning on a life: actions that maximise friendship, beauty, knowledge, and some other goods (Railton 1984), actions that promote our rational nature (Hurka 1993), actions that improve the well-being of persons and sentient nonpersons (Singer 1996, ch. 4) and actions that lead us to overcome the fundamental challenges of the time (Dworkin 2000, ch. 6). Some of these claims need more detail to be usefully action-guiding. Who decides which challenges are fundamental, for example? The accounts that are detailed enough stipulate that certain actions, such as promoting friendship and beauty, are objectively meaningful. But, by taking more steps back, to a *distant* objective viewpoint, these more-detailed objective theories appear to be strongly influenced by subjective values.

The particular kinds of goals that we strive for will all seem unimportant from a distant objective standpoint, their professed significance ineluctably stained with subjective values. As we step back, the specific goals and subjective values drop away, but the significance of the size of an action's consequences remains. After a few more steps, however, even the consequences become vanishing small, all of them except for the infinite consequences, that is. No matter how far we step back, no matter how distant the objective viewpoint is that we take, infinite consequences will never vanish into insignificance. When all the values and finite consequences have

disappeared into the distance, actions with infinite consequences remain, indelibly influencing future events. Having an infinite consequence is objectively meaningful because infinite consequences will always be significant, no matter how far we step back. Note that any negative implications about other objective Naturalist theories stemming from these comments are not intended to suggest that only actions with infinite consequences can confer objective meaning on a life. I am merely arguing that an action's having infinite consequences is sufficient to make it objectively meaningful in a way that helps combat the Absurd.

The third component of Infinite Consequence is that the correct alignment of subjective and objective meaning from the first two components can make a life truly meaningful. According to the first two components of Infinite Consequence, a purposeful action with infinite consequences can be objectively meaningful and also subjectively meaningful if it is accompanied by the appropriate beliefs. The third component dictates that if *a single action* achieves *both* subjective and objective meaning in this way, then it has met the sufficient conditions to confer true meaning on the life of the actor. This component is definitional in character. As mentioned earlier, some people will be satisfied with only a subjectively meaningful life. This third component of Infinite Consequence is stipulated, rather than argued for, because this chapter is mainly addressing an issue for people who, like Tolstoy, are *not* satisfied with subjective meaning alone. Nevertheless, its inclusion will be briefly motivated.

A person, whose life is only subjectively meaningful, might rightly worry whether her actions are actually meaningful or if they merely seem that way. This worry could potentially be allayed by external or objective evaluation of her actions. When the objective evaluation is negative, then the Absurd arises.

If the evaluation is positive, however, then the person can be confident that her life is truly meaningful. In this way, true meaning is much better to have than subjective meaning alone. Similarly, true meaning is better to have than solely objective meaning. Someone whose life was objectively meaningful, but was not aware of it or hated that their life was objectively meaningful, also lacks something that would make their life more meaningful.

Even someone whose life was both subjectively and objectively meaningful might not have a truly meaningful life, however. Imagine a person whose subjective meaning came from objectively meaningless actions, and who was disgusted by his actions that were objectively meaningful. This person would not be sure that the meaning he experienced from the subjective viewpoint was actually meaningful and he would experience nothing meaningful about his objectively meaningful actions. This person's life would be more meaningful if his subjectively meaningful actions were the same as his objectively meaningful actions. If the objective and subjective meaning of this person's actions were aligned in this way, then he would get to enjoy his objectively meaningful actions and be free from the suffering caused by his doubt of the authenticity of his subjectively meaningful actions.

The concept of a truly meaningful life, as defined here, is useful because it allows us to acknowledge that the correct alignment of subjective and objective meaning is more desirable than the alternatives mentioned above. The concept of a truly meaningful life is also useful for identifying a theoretical possibility for resolving the problem of the Absurd. Indeed, the very situation of comparing the subjective and objective meaning of our actions reveals the value of actions that cause those meanings to coincide. By valuing truly meaningful actions over the others, the tension between the different perspectives of the Absurd can be resolved.

To completely resolve the Absurd in a positive way (avoiding Nihilism), a plausible theory of how subjective and objective meaning can coincide is required. The principle of Infinite Consequence provides such a theory—it describes what kinds of actions are sufficient to confer both subjective and objective meaning on a life. Recall that, Infinite Consequence is the idea that when a purposeful action with infinite consequences is accompanied by the appropriate beliefs that action is sufficient to make the life of the actor truly meaningful. But what does it really mean for our actions to have an infinite consequence?

In the context of meaningful lives for human beings, infinite consequences are the results of an action that affect humankind (or perhaps other forms of life)⁹³ and continue to do so. The consequences need not affect all humans at any moment, but need to continue to affect some humans (or perhaps other forms of life) as time goes by. The consequences need not be of infinite value in any moment of time (if that is possible), such as by making someone infinitely happy. The duration of the consequences must, however, be infinite. This could take the form of one continuous consequence or many sequential consequences. The consequences need not be affecting life constantly for infinity; periodic consequences for infinity are sufficient to count as ‘infinite consequences’ for the purposes of the principle of Infinite Consequence.

Note that living for infinity does not, by itself, entail that your life is classified as truly meaningful according to the principle of Infinite Consequence. Living for infinity would permit performing an infinite number of actions, but it does not entail that any of those particular actions would be

⁹³ Consequences for all forms of life that have the capacity to experience truly meaningful lives, such as intelligent aliens, may even be preferable to the anthropocentric view presented here.

truly meaningful. If none of those actions had infinite consequences, then none of them would be truly meaningful. Living for infinity does seem to increase the chances of performing a truly meaningful action, however.

With this description of Infinite Consequence in mind, we can see that the following purposeful actions can confer true meaning on a life because of their infinite consequences and appropriate accompanying beliefs. Imagine a musician who considers it a meaningful goal that his music influences humankind infinitely. Because the musician believes that his music is important, he gains subjective meaning from performing and recording it. If his music does in fact influence humankind for infinity, he will also gain objective meaning from these actions. Whether the musician's actions are truly meaningful, then, rests most heavily upon the condition that his music continues to influence people for infinity.⁹⁴ Ultimately, the influence his work will have on others is something that the musician cannot control. All he can do is attempt to make his work as good, and as easily available, as possible. Furthermore, even if his music is popular throughout his life, the musician will never be able to know if his music will be admired by future listeners *for infinity*. These two elements, striving to achieve some kind of effect on something outside of himself and not being able to know for sure if he has succeeded, seem apt for a truly meaningful life. At least they prevent two potential problems. First, affecting others is not always directly under our control, adding some appropriate difficulty and uncertainty to this method of achieving true meaning in life. Second, the fact that we cannot know if any of our actions will have the right kind of consequences *for infinity* means that we will be less likely to be in the potentially boring position of knowing that we have already made our lives truly meaningful and having to struggle to work

⁹⁴ But note that, since Infinite Consequence sets out sufficient conditions for truly meaningful actions, it is possible that the musician's actions are truly meaningful for some other reason.

out what to do next.⁹⁵ Indeed, lack of certainty about whether we have achieved true meaning in our life will motivate us to keep on striving to perform meaningful actions.

Imagine also a scientist who devises a new technology that allows humankind to escape a disaster that would otherwise have destroyed all sentient life on Earth, such as the supernova of the sun. All of the scientist's forebears, many of the people that she interacted with during her life, tax payers who helped to fund her research institute, other past and present scientists whose findings were relevant for her learning and her humankind-saving discovery, and many other tangentially related people have all played some degree of causal role in ensuring the on-going existence of all future humans. Performing an action with a non-trivial causal role in humankind persisting for an infinite period of time is objectively meaningful according to Infinite Consequence. However, only those like the scientist, who performed these actions with both the purpose of enabling humankind to live on for infinity, the belief that the continuation of humankind is a meaningful achievement, and the belief that they will be successful, conferred true meaning on their lives.

Both of these examples of how to achieve a truly meaningful life require that humankind *actually* lives on for infinity. If the universe became permanently inhospitable due to heat death sometime after the scientist helped us avoid the supernova of the sun, humankind would still fail to live on for infinity. If humankind does not live on for infinity, then the philosopher, the musician, and the scientist all fail to achieve their infinite consequences. Since this failure would happen after their deaths, it would not

⁹⁵ I say 'less likely' here because there may be other ways that lives become truly meaningful and we might gain true meaning from these other avenues and also come to know that we have acquired true meaning somehow.

affect the subjective meaning that they experienced during their life. It would mean, however, that their actions were probably never objectively meaningful and, thereby, probably never truly meaningful.⁹⁶ The issue of whether it is possible for human life (or perhaps other sentient life) to live on for infinity is discussed in the next section.

3.3 *Scientific Optimism*

The principle of Scientific Optimism holds that continual scientific and technological advancement will allow our actions to have infinite consequences. In light of the discussion in the previous section, the most important facet of adherence to Scientific Optimism is the belief that continual scientific and technological advancement will allow humankind (or perhaps other forms of life) to live on for infinity. Belief in Scientific Optimism does not require the belief in any specific account of how life will come to live for infinity. It merely requires the belief that continual scientific and technological advancement will make it possible *somehow*.

Scientific Optimism will be argued for here because the plausibility of Optimistic Naturalism, its usefulness for resolving the Absurd, and its ability respond to the criticisms mentioned above all depend on it. If Scientific Optimism is implausible, then belief in Infinite Consequence will provide no solace to those in the grip of the Absurd. Such people will see a theoretical solution to the tension between how meaningful their lives are from the subjective and objective viewpoints, but they will also see no physical possibility for enacting that solution.

Scientific Optimism might seem incredibly optimistic because of its stipulation that continual scientific and technological advancements will

⁹⁶ Again, I say 'probably' here because there could be some other reason why their actions were objectively meaningful.

enable humankind to live on for infinity. Readers can be forgiven for thinking that science has confirmed *as a fact* that humankind will eventually be annihilated. Indeed, most cosmologists do believe that the universe will be unable to support life indefinitely (Starobinsky 2000). The most common justification for this belief is the prediction that the universe will continue to expand and cool until there is no free energy left to support life. The leading alternate theory predicts that the Big Bang and subsequent expansion of matter will reverse at some point. If all matter is eventually drawn back together, no life is expected to survive this Big Crunch. There are other Naturalistic theories about the fate of the universe, however, including some with a more optimistic outlook about the chances of humankind's on-going survival.

One live theory in cosmology, Eternal Inflation, predicts that new baby universes will always bubble out from our existing one.⁹⁷ If this theory turns out to be true, then the right kinds of technology might enable humankind to escape into new universes whenever the existing one was becoming uninhabitable and, thereby, live on for infinity.

Several leading scientists have successfully modelled the theoretical possibility of how humankind might intentionally create a new area of inflation (usually called a baby or bubble universe).⁹⁸ Two practical problems still stand in the way of creating a useful baby universe in this way, however;

⁹⁷ Or, perhaps it would be more accurate to say that new parts of the universe are inflating all of the time while other parts come to a halt because many physicists use 'universe' to mean 'all that there is', and others use it to mean 'all that there is that obeys this set of physical laws' (Bettini 2005). See Aguirre (2006) for an accessible introduction to Eternal Inflation.

⁹⁸ Research on the topic includes: Farhi & Guth (1987), Farhi, Guth & Guven (1990), Fischler, Morgan & Polchinski, (1990a; 1990b), Guendelman & Portnoy (1999; 2001), and Sakai *et al.* (2006).

the energy required to create them and finding a way to safely travel into them.

To condense the required materials into the tiny space needed to create the baby universe would take approximately the total energy output of a galaxy (Kaku 2004). While this seems like a prohibitive amount of energy, prominent physicist Michio Kaku is confident that we will have the technology to harness this huge amount of energy well before the sun envelopes the Earth. Even if we could create a baby universe, however, the intense forces involved might make it impossible to transport humans into it before it slips out of contact with our universe (Kaku 2004). In response to this worry, Kaku suggests that we could use additional energy to stabilise the new universe long enough for humans or nanobots to enter it. If nanobots could survive the transition to a new universe, then they might also be programmed to collect and combine the raw materials necessary to create human life. If this or other similar processes prove to be physically as well as theoretically possible, then repeatedly recreating human life in new universes would allow humankind to live on for infinity.

An objection to the possibility of scientific and technological advancement enabling humankind to continue to avoid disasters in this way might arise at this point. If there is a greater than zero probability that some disaster will prevent humankind from surviving in any finite period of time, then a disaster *definitely will* prevent humankind from surviving at some point over an infinite amount of time. This objection does not provide a good reason to believe humankind could not live on for infinity. At best it provides a reason to think that it is unlikely that humankind could live on for infinity. To see why, consider rolling a die; the probability of rolling a six is quite small if the die is only rolled once, and if the die is rolled four or five times, then the

probability of rolling at least one six is much higher. Notice, however, that a robot could roll a die over and over indefinitely and it is possible that a six would *never* come up. In the same way, it is entirely possible that, despite the strong odds against it, an unavoidable humankind-destroying disaster never occurs even if we survive for infinity.

Furthermore, there is a good reason to think that the longer humankind survives (assuming that we continue to advance science and technology), the better equipped we will be to avert potentially humankind-destroying disasters. We have seen amazing scientific and technological advancements over the past century and the near future promises even more and at an increasingly faster rate. Ray Kurzweil and other technologists and philosophers predict the future merging of human mind and machine to create advanced artificial intelligence (AI+).⁹⁹ AI+ combines the strengths of human intelligence with mechanical hardware's processing power to create intelligence more powerful than anything currently existing. The creation of AI+ would most likely result in the law of accelerating returns applying to scientific and technological advancement. If this happens, then science and technology will increase exponentially and so too will our problem-solving power. If scientific and technological advancement is exponential, then Scientific Optimism becomes much more plausible. Furthermore, the objection that humankind is exceedingly likely to be wiped out by some disaster in the very distant future becomes correspondingly less convincing.

Even if AI+ is impossible to create, and science and technology do not advance exponentially, all that is required for Scientific Optimism to have some credibility is that science and technology continue to advance quickly (something that seems overwhelmingly likely). Mere linear advancement of science and technology will be enough for us to avoid local disasters, such as

⁹⁹ See Kurzweil's *The Singularity is Near* (2005).

catastrophic climate change or the Earth being consumed by the sun as it turns into a red dwarf (assuming that we actually try to survive them). Furthermore, the predicted death of our observable universe is potentially trillions of years away, leaving plenty of time to find a method for enabling humankind to live on for infinity. If humans can progress the output of aeronautical science all the way from 'jumping' to 'interplanetary travel' in a hundred years or so, then we cannot sensibly claim to be able to imagine what we could achieve in a few trillion years.

Taken individually, any specific theory about how scientific and technological progress might allow for us to meaningfully connect with the infinite seems (at least at this stage) very unlikely to be true. Indeed, firm belief that Kaku's theory is *the* way that scientific advancement will allow for us to meaningfully connect with the infinite is not plausibly justifiable. To believe in the principle of Scientific Optimism only entails belief that scientific and technological progress will make it possible for humankind (or perhaps other forms of life) to live for infinity *somehow*. Perhaps Kaku's theory will turn out to be true. Or, more plausibly, perhaps scientific advancement will allow humankind to live on for infinity in some other way. There are countless ways in which scientific and technological advancement might allow humankind to live on for infinity. Considering these myriad possibilities in combination with our continually-increasing ability to achieve things that seemed impossible only years before, belief in the plausibility of Scientific Optimism is plausibly rational (if perhaps a little optimistic).

Critics might insist that Scientific Optimism is wildly implausible. They might claim that belief in Scientific Optimism is irrational because the likelihood of science and technology enabling humankind to live on for infinity is vanishingly small. But such critics should be wary. Coupled with

the fact that no one will ever be able to know if humankind does live on for infinity, belief in Scientific Optimism is quite similar to faith in a supernatural entity. Two points in response. First, it is not sensible for anyone to claim that a logically possible future of the universe is implausible because it is unlikely. Given that any number of advances, discoveries, and unexpected events can occur in just a few years, predicting the future of the universe contains too many unknown variables to be done with any accuracy. Second, given all the possible natural and supernatural theories about the meaning of life, belief in any one particular theory also requires *a leap of faith*. Since there are infinite possible mutually exclusive theories of the meaning of life, any one particular theory only has a vanishingly small chance of being true. Furthermore, unless a very surprising advance in philosophy is made, we will never have a way to verify if any particular theory of the meaning of life is actually true. Considering these two responses, there is no reason to think that Scientific Optimism is any less plausible than other logically possible theories of the meaning of life.

If Scientific Optimism is plausible, however, it is reasonable to believe that one can resolve the Absurd and achieve a truly meaningful life by performing actions that one finds subjectively meaningful and expects to have infinite consequences. Belief in the plausibility of Scientific Optimism is required for an Optimistic Naturalist to properly resolve the vast difference between the meaning their life seems to have from the subjective and objective viewpoints. If an Optimistic Naturalist did not consider it plausible that scientific and technological advancement will enable humankind (or perhaps other forms of life) to live on for infinity, then they might not believe that their actions were objectively meaningful. Despite this lack of objective meaning, their actions might still seem meaningful to them from the subjective viewpoint, potentially leaving them in the grip of the Absurd.

4. How to Act in Light of Optimistic Naturalism

Like Tolstoy did, many Naturalists find contemplation of the Absurd deeply distressing. The feeling that he might have been wrong about how meaningful his life was, and the worry that his life might have had no objective meaning at all, depressed Tolstoy to the point where he felt paralysed (2000, pp. 11–12). Any Naturalist who finds themselves worrying about the tension between how meaningful their life seems from the subjective and objective viewpoints should consider becoming an Optimistic Naturalist. If they can find subjective meaning in any actions that could have infinite consequences, and if they believe that having infinite consequences makes that action objectively meaningful, then they have a theoretical blueprint for a bridge that crosses the gap between the subjective and objective viewpoints. To resolve the Absurd, however, these Naturalists would also have to believe that scientific and technological progress will somehow allow humankind (or possibly other forms of life) to live on for infinity and thereby provide the materials necessary to make that theoretical blueprint a reality. Any Naturalist who takes on these two beliefs, Infinite Consequence and Scientific Optimism, becomes an Optimistic Naturalist.

How then should these new Optimistic Naturalists live in order to achieve a truly meaningful life? There are many potential options for someone to achieve a truly meaningful life according to Optimistic Naturalism. All of these options, however, depend on the person purposefully performing an action that has infinite consequences.¹⁰⁰ The most obvious options include trying to significantly influence the future of humankind for infinity. This could be achieved in a variety of ways, including by creating art, contributing to existing philosophical or scientific knowledge, and even lovingly raising

¹⁰⁰ Again, there may be other options not discussed here because Optimistic Naturalism only sets out sufficient criteria for a truly meaningful life, not necessary criteria.

children with the aim that they will do the same. In order to achieve these infinite consequences, though, Optimistic Naturalists need it to be true that humankind will actually live on for infinity. Unfortunately, humankind's infinite existence is something that they could not rationally be sure of. For this reason, it would be prudent for any Optimistic Naturalist who finds subjective meaning in actively progressing science and technology to focus their efforts on just that. Since the objective and true meaningfulness of Optimistic Naturalists' lives depends on humankind living on for infinity, if they can help to increase the probability of that happening, then they probably should.

There are many ways for Optimistic Naturalists with different tastes and capabilities to contribute to the advancement of science and technology. For some Optimistic Naturalists this might mean continuing their research in a specific sub-field of physics. For others it could mean focussing on their business enterprises and using the profits to establish scientific research centres. It may even mean raising children to have a keen interest in science. No matter what the style of contribution to the advancement of science and technology, all such actions could have infinite consequences by helping to enable humankind (or possibly other forms of life) to live on for infinity. Furthermore, these contributions to the advancement of science and technology could also enable other kinds of actions, such as creating an infinitely significant work of art, to confer true meaning on people's lives.

In addition to trying to advance science and technology as much as possible, Optimistic Naturalists should also consider pursuing other compatible Naturalist theories of the meaning of life. Consideration of other theories is possible because the principle of Infinite Consequence only provided sufficient conditions for true meaning; the conditions were not described as necessary or exhaustive. Any additional theories should not be in

tension with any of the principles of Optimistic Naturalism, however, and should also bridge the gap between the subjective and objective viewpoints characteristic of the Absurd. If the other theories do not bridge the gap between the subjective and objective viewpoints, then belief in them might exacerbate the tension of the Absurd.

Not all Naturalists find contemplation of the Absurd causes them any distress, however. Naturalists who are unaffected by the Absurd tend to be satisfied with subjective meaning and often consider objective meaning to be illusory (e.g. Taylor 2000). Such strict subjective Naturalists should consider the arguments for Infinite Consequence and Scientific Optimism from above. If they find them convincing, then it would be rational to adopt Optimistic Naturalism in addition to any theories they currently use to guide their actions. If these subjective Naturalists do not find Infinite Consequence plausible, however, they at least have an idea of how far a theory might have to go to bridge the gap between the subjective and objective viewpoints and make a more meaningful life possible.

The plausibility of Optimistic Naturalism will also be important to some nihilists. Some would-be Naturalists might have turned to Nihilism after contemplation of the Absurd. Like Tolstoy did during his depression, some Nihilists believe in the following two claims. First, actions can only be rationally considered as subjectively meaningful if they are also objectively meaningful. And second, Naturalist theories of the meaning of life cannot provide objective meaning. If such Nihilists are persuaded by the arguments for Infinite Consequence and Scientific Optimism, then they should become Optimistic Naturalists. If they are not persuaded, then they might find some epistemic comfort in the idea that they now have even more reason to think that life is meaningless.

The plausibility of Infinite Consequence and Scientific Optimism is also relevant to any Supernaturalist who criticises Naturalist theories for not offering accounts of a meaningful life that are meaningful enough. Tolstoy's version of this criticism is that Naturalist theories are not truly meaningful because they cannot connect our finite lives with the infinite (2000, p. 17). The principle of Infinite Consequence reveals the theoretical blueprint for how purposeful actions with infinite consequences can in fact connect our finite lives with the infinite. Furthermore, the principle of Scientific Optimism describes why it is plausible to believe that the blueprint will somehow become a physical reality. If Infinite Consequence and Scientific Optimism are plausible, then Supernaturalists would have to cease with the criticism that Naturalist theories of the meaning of life do not offering truly meaningful lives (at least in the sense of 'truly meaningful' used in this chapter).

5. The Infinite Utility Theory of Meaningfulness

If Optimistic Naturalism is plausible, then other philosophical debates about the meaning of life are likely to be affected. For example, the plausibility of Optimistic Naturalism allows for an interesting new version of the Utilitarian Theory of Meaningfulness (UTM)—the Infinite Happiness Theory of Meaningfulness. I show that the Infinite Happiness Theory of Meaningfulness can adequately defend itself against Metz's (2003a) currently unanswered counterexamples to UTM in interesting ways.

In the only recent published discussion of the Utilitarian Theory of Meaningfulness (UTM) that I could find, Thaddeus Metz (2003a) presents several counterexamples to the utility-based theory. Metz (2003a, p. 54) describes the UTM as an objective theory, which claims: "A person's life is meaningful just to the extent that she makes those in the world better off." This description of the UTM implies that increasing well-being is the meaning

of life. Other remarks by Metz (2003a) also imply that, in line with traditional utilitarian thought, the UTM would understand happiness as constituting well-being. In this manner, the UTM provides an objective account of the meaning of life that might capture what some people mean when they claim that the meaning of life is happiness.

It should be pointed out that no defender of UTM has replied to Metz's (2003a) counterexamples. As may become clear, it seems that a direct defense of UTM along the lines familiar to normative ethicists—nuanced discussion of future consequences, possible definitions of utility, etc.—would be at least fairly successful. However, the Optimistic Naturalism-inspired version of UTM—Infinite Happiness Theory of Meaningfulness—deals with Metz's (2003a) counter examples in more unusual and interesting ways. So, I will discuss how the Infinite Happiness Theory of Meaningfulness avoids Metz's criticisms here. A full defense of the Infinite Happiness Theory of Meaningfulness will not be pursued, however, because the purpose of the discussion in this section is to highlight a potential use of Optimistic Naturalism by showing how its plausibility can affect other philosophical debates about the meaning of life.

Infinite Happiness Theory of Meaningfulness: a life is truly meaningful if it includes a purposeful action with infinite positive consequences on people's well-being and does not contain any actions with infinite negative consequences on people's well-being (where a purposeful action is defined as an intentional action that is accompanied with the beliefs that the action is meaningful and that it will have an infinite consequence).

There are several differences between the UTM and the Infinite Happiness Theory of Meaningfulness (IHTM), but both are clearly utilitarian in many

ways. Most importantly, both theories involve the maximisation of happiness or utility. However, the plausibility of Optimistic Naturalism makes maximising utility an entirely different endeavour. If Scientific Optimism is true, then continual scientific and technological advancement will allow our actions to have infinite consequences, including infinite positive and negative consequences on people's well-being. Therefore, in some situations the truth of Scientific Optimism entails that maximising utility will require causing infinite positive consequences on people's well-being!

Metz (2003a, pp. 56–57) labels his first counterexample to the UTM as 'Bear'. In Bear a hungry grizzly bear approaches you and a friend. The bear will kill one of you. You are a much slower runner than your friend. Both you and your friend will have the same net impact on other people's well-being, but if you survive you will enjoy slightly more well-being than your friend would if she survives.¹⁰¹ According to UTM, it would be more meaningful for you to trip your friend. Metz claims that this result is counter intuitive and so counts it as a reason to think that UTM is implausible. Applying the IHTM to Bear, however, we can see that it would only be more truly meaningful to trip your friend (than to die) if you would subsequently perform a purposeful action with infinite positive consequences on people's well-being (and never perform an action with infinite negative consequences on people's well-being) and your friend would not subsequently perform such an action. The result of applying the UTM seemed absurd because the gains from the immoral action were so limited, but the result of applying the IHTM seems plausible because the gains from the immoral act are quite the opposite—they are infinitely beneficial.

¹⁰¹ In order "to be complete", Metz (2003a, p. 57) also adds "the bear would find no difference in taste or nutrition between you and your friend."

Metz (2003a, pp. 58–59) then presents a duo of counterexamples to the UTM—‘Humiliation’ and ‘Prostitution’. In Humiliation several people experience great utility from denigrating you. Given that you are fairly thick-skinned and don’t have much else to do, you can maximise utility by allowing your humiliators to continue making jokes at your expense. In Prostitution, Metz (2003a, p. 59) claims that prostitutes often bring about considerable amounts of utility by selling sex and at least some of them must not suffer equivalent amounts of disutility. In Humiliation and Prostitution then, the UTM will claim that some people in some situations will be able to make their life most meaningful by allowing themselves to be humiliated or by selling sexual services. Metz (2003a) claims that Humiliation and Prostitution are counterexamples to the UTM because it seems that neither of these activities can ever make someone’s life more meaningful, let alone be the most meaningful action for them in certain circumstances. The IHTM better accommodates our intuitions about Humiliation and Prostitution because it deems the associated actions as conferring no objective, and therefore no true, meaning on a life.

The ability of the IHTM to better accommodate Metz’s (2003a) currently unanswered counterexamples to the UTM, does not necessarily mean that the IHTM is a plausible theory of the meaning of life. What the foregoing discussion does show is that the plausibility of Optimistic Naturalism can affect other philosophical debates about the meaning of life. Furthermore, the IHTM warrants further investigation because, since the demise of the UTM, there has been no objective or hybrid theory of the meaning of life that has been able to satisfactorily resolve the problem of the Absurd while also making sense of a claim commonly made by laypeople—that the meaning of life is happiness.

6. Conclusion

Optimistic Naturalism is the idea that scientific and technological advancement will allow us to meaningfully connect our finite lives with the infinite by having an infinite consequence. In this way, Optimistic Naturalism provides us with the chance to lead a truly meaningful life. Optimistic Naturalism is based on the principles of Naturalism, Infinite Consequence, and Scientific Optimism. Most of the arguments in this chapter were in support of Infinite Consequence and Scientific Optimism because both must be plausible for Optimistic Naturalism to be able to resolve the Absurd and refute the criticism that Naturalist theories do not provide accounts of the meaning of life that are meaningful enough.

The principle of Infinite Consequence dictates that a purposeful action with infinite consequences can confer both subjective and objective meaning, the combination of which confers true meaning on a life. Most importantly, actions with infinite consequences were argued to be objectively meaningful because, after all subjective values are stripped away by adopting a more and more distant objective viewpoint, infinite consequences can still be seen to have effects on humankind (or perhaps other forms of life).

The principle of Scientific Optimism holds that continual scientific and technological advancement will allow our actions to have infinite consequences, most likely enabling humankind (or perhaps other forms of life) to live on for infinity. The plausibility of Scientific Optimism was argued for using an example based on the theory of Eternal Inflation and a general argument about the unfathomable possibilities of future scientific and technological advances. In the end, however, belief in Scientific Optimism was argued to require a leap of faith, although not of the kind that would make Optimistic Naturalism implausible.

It was argued that Optimistic Naturalism refutes a general criticism of Naturalist theories of the meaning of life and also Tolstoy's more specific version of this criticism. By explaining the importance and possibility of correctly aligning naturally derived subjective and objective meaning, Optimistic Naturalism refutes the general criticism that Naturalist theories do not provide accounts of the meaning of life that are meaningful enough. Furthermore, by explaining how it is plausible that we can perform actions with infinite effects, Optimistic Naturalism refutes Tolstoy's criticism that Naturalist theories cannot connect our finite lives with the infinite.

It was also argued that those who find themselves in the grip of the Absurd should adopt Optimistic Naturalism. This was argued on the basis that Optimistic Naturalism provides both the theoretical blueprint for, and the physical possibility of, bridging the vast divide between how meaningful our actions appear from the subjective and objective viewpoints. Advice was also given to readers about the implications of Optimistic Naturalism's plausibility for their own beliefs and actions. Most importantly, Optimistic Naturalists were urged to focus their activity on the advancement of science and technology, assuming they find such activities subjectively meaningful.

Finally, the assumption that Optimistic Naturalism is plausible allowed for the creation of the Infinite Happiness Theory of Meaningfulness, which affected a philosophical debate about the meaning of life. It was argued that the IHTM could easily accommodate Metz's (2003) currently unanswered counterexamples to UTM and it was suggested that IHTM might deserve further investigation.

Chapter 7

Happiness and Climate Change: How Should the Responsibilities of Adaptation and Mitigation be Shared?

Chapter Summary

This chapter argues that considerations of happiness lead to the conclusion that a sufficientarianism-based hybrid distributional principle is the most ethically suitable principle to guide the apportionment of responsibilities to deal with the harmful effects of climate change. The proposed principle is a hybrid of the Conditional Polluter Pays Principle and the Ability to Pay Principle. The hybrid principle is explained in enough detail to be useful to policymakers and is defended against potential objections. The sufficientarian justification for the hybrid principle proposed here also informs an ability to pay objection, which is argued to be problematic for any distribution principle that doesn't prioritise the Ability to Pay Principle.

1. Introduction¹⁰²

Without global collective action on climate change, the ominous threats posed by a rapidly changing climate are predicted to have catastrophic effects for all life on Earth (IPCC 2007). Despite all major governments acknowledging the

¹⁰² This chapter is the basis for Weijers, Eng, and Das (2010).

significant role of anthropogenic emissions in producing rapid global warming,¹⁰³ relatively little has been done to reduce anthropogenic emissions.

Perhaps the best hope for reaching an enduring and widely supported international agreement on tackling climate change is to base it on the widely agreed upon principle of common but differentiated responsibilities (CBDR), Principle 7 of the Rio Declaration on Environment and Development. This principle encapsulates the international consensus that the on-going responsibility to protect the global commons is to be shared, though not necessarily evenly. In particular, the principle of CBDR highlights that industrialised states bear a greater responsibility to address the climate change problem because of the pressure they have put on the global environment and because of their financial and technological ability to take action more easily (Rio Declaration on Environment and Development 1992).¹⁰⁴ Unfortunately, serious disagreements remain about how the principle of CBDR should be interpreted. If the solution to the looming climate change problem does not include an answer to the question, ‘what is the most just way to decide what should be done about rapid climate change and who should do it?’, then it is unlikely to be an enduring solution. Since justice is

¹⁰³ In 1992, the United States and 171 other states, 108 represented by their head of state or government, participated in the Earth Summit at Rio de Janeiro (where the Rio Declaration on Environment and Development was discussed) and agreed that climate change was being affected by human activity (United Nations Department of Public Information 1997).

¹⁰⁴ Principle 7 of the Rio Declaration on Environment and Development provides the first formulation of the principle of CBDR: “In view of the different contributions to global environmental degradation, States have common but differentiated responsibilities. The developed countries acknowledge the responsibility that they bear in the international pursuit of sustainable development in view of the pressures their societies place on the global environment and of the technologies and financial resources they command” (UNFCCC 1992).

relevant to this problem, philosophers have an important role to play. This chapter is a contribution to the on-going philosophical debate about how the principle of CBDR can be interpreted in a way that is both fair and amenable to the formation of policy. This chapter also demonstrates how important real-world problems can be fruitfully analysed from the perspective of happiness.

The extant literature on how to fairly divide the responsibilities of dealing with climate change includes several widely-discussed principles of justice.¹⁰⁵ The only current agreement on these principles of justice, however, is that none of them distribute responsibilities in a way that is fair to all relevant parties, at least when considered individually (Page 2008). This has encouraged recent attempts to solve the problem of fair distribution of responsibilities by combining the main principles of justice into hybrid accounts. The purpose of these hybrid accounts is to take into consideration all of the main morally relevant factors and distribute the responsibilities of dealing with climate change in a way that is fair to all parties and amenable to translation into international policy.

This chapter follows the general approach just described. First the main principles of justice and the standard objections to them are discussed. Then I explain the hybrid account defended in this chapter. Motivated by considerations of happiness, the hybrid account is primarily based on the distributive principle of sufficientarianism. In the context of climate justice, I interpret sufficientarianism as the idea that each country should have the means to provide a minimally happy existence for each of its citizens. It is then argued that sufficientarian considerations give us good reason to think that the 'ability to pay objection' should be taken much more seriously in this

¹⁰⁵ See, for example, Singer (2008), Caney (2005), Shue (1999), Neumayer (2000), Gardiner (2004), and Page (1999; 2008).

debate. After noting that the proposed hybrid account includes enough detail to be a useful starting point for policy makers, it is defended against some potential objections.

2. Polluter Pays Principle

The polluter pays principle (PPP) apportions the responsibility for paying the costs of dealing with climate change among the parties who caused the climate-changing pollution. On first inspection, the PPP is probably the most intuitive way of thinking about the ethics of climate change. The PPP is based on the ubiquitous idea that those who cause harm to others should be morally responsible for remedying that harm. The most appealing benefit of the PPP is that it has the ability to provide the appropriate incentive to prevent polluting by directly linking moral responsibility, and the resulting accountability, to the kinds of actions that should be discouraged.

The climate-changing pollution, referred to above, should be taken to mean the emitting of greenhouse gases above some agreed upon quota. The quotas agreed upon (in the Kyoto Protocol to the United Nations Framework Convention on Climate Change and at the more-recent climate conference in Copenhagen) are all self-imposed and based on a reduction of state's absolute per capita, or per gross national product, emissions relative to some past point in time.¹⁰⁶ These arbitrary quotas are patently unfair because they fail to acknowledge that there is no good moral reason for any distribution of a common global good, like the atmosphere, other than an equal share for everyone (Singer 2008, p. 671). There is a much fairer method of creating a

¹⁰⁶ Singer (2008 p. 671), references Claussen and McNeilly (1998) and says the targets agreed on at Kyoto "were arrived at through negotiations with government leaders, and they were not based on any general principles of fairness, nor much else that can be defended on any terms other than the need to get an agreement".

quota (and one that would do more to reduce the likely catastrophic effects of climate change). This method would see the annual amount of total emissions considered safe by current United Nations Framework Convention on Climate Change (UNFCCC) estimates to be distributed to states based on their near-future population trajectory as compiled by the United Nations.¹⁰⁷ If this approach were adopted, then the per capita aspect of this method would plausibly result in current *and* future people receiving their fair share of the atmosphere. The use of near-future population trajectories instead of actual populations is meant to eliminate perverse incentives for population control.

The PPP produces entirely appropriate results when applied to current and future polluting. However, when the PPP is applied to historical emitting a problem arises: past polluters, for the most part, were not aware that their actions would have harmful consequences. In response to this fact two versions of the PPP have emerged. The full liability PPP (FPPP) assigns moral responsibility to agents to redress all of the relevant harms they have caused even when they were unaware that their actions would lead to such harm. The other version, the conditional liability PPP (CPPP) is weaker; the CPPP assigns moral responsibility only to those who *knowingly* pollute or who should have known that their greenhouse gas emitting was likely to cause harm. Such polluting will be referred to as 'culpable polluting'. Culpable polluting is to be distinguished from non-culpable polluting on the basis of whether the polluter can reasonably be held to have known that their polluting was likely to cause harm. The moral significance of this distinction

¹⁰⁷ However, as Reisinger and Larsen (2010) argue, more thought needs to be given to what total amount of atmospheric greenhouse gas emissions is desirable because all amounts have different predicted outcomes for the various ecosystems and forms of life on Earth. The setting of the 'safe' level of total atmospheric greenhouse gas emissions is a moral task that deserves greater attention than it has received.

ultimately leads to the CPPP being incorporated in the hybrid account discussed here, instead of the FPPP.

Applying the CPPP to the current climate change debate requires a method for discerning who can reasonably be held to have known that their polluting was likely to cause harm. The most conservative method for doing this would be to understand 1992 (when the Rio Declaration on Environment and Development was signed) as the date past which all states should be deemed as knowing that emitting excessive greenhouse gases is likely to cause harm. By using 1992 as the starting date for culpable polluting, the CPPP can satisfyingly deal with the problem of non-culpable polluting. However, Page (2008, p. 570) has criticised the use of this fairly recent date as the relevant starting point because he thinks it results in “harsh treatment for the newly industrialised populations and lax treatment of those residing in countries of transition”. To move the starting date further back would decrease Page’s fairness concerns but would exacerbate the unfairness to polluters who were truly not aware of the consequences of their actions. The combination of these two concerns makes it difficult to specify a fair date after which states should be deemed as knowing that greenhouse gas emissions over a certain level are likely to cause harm.

The foremost problem for the CPPP is that it fails to designate sufficient moral responsibility to adequately address the problem of climate change (Caney 2005). A large portion of the polluting was caused before 1992 and the CPPP cannot assign the responsibility to deal with that pollution to anyone. The resulting insufficient allocation of responsibility might influence some readers to favour the FPPP, but the FPPP suffers from the same problem (albeit to a lesser extent). Many unknowing and knowing polluters are dead and therefore cannot pay to mitigate damage that their polluting will help to bring about. The FPPP cannot assign the responsibility to deal with that

pollution to anyone else and so the FPPP also suffers from the insufficiency problem.

A common response to this insufficiency problem has been to argue that individuals currently residing in states that are primarily responsible for climate change should be held morally responsible for pollution that was caused by the previous generations of those states. But this intergenerational transition of responsibility seems unfair; why should the mere fact that someone lives in a country whose previous citizens polluted make them responsible for the polluting?

It might be argued that this intergenerational objection presupposes that the relevant moral agents are individuals as opposed to states. A collectivist approach to the PPP would view states as the relevant moral agent for the current climate change debate, considerably decreasing the significance of this intergenerational problem. A collectivist approach to the PPP has some intuitive plausibility given that any future agreement the UNFCCC reaches will distribute the responsibilities for dealing with climate change among states in the first instance.

Applying this collectivist version of the PPP reveals that, because of their relatively long history of greenhouse gas emitting, the developed nations have the primary responsibility for mitigating and adapting to rapid climate change. These states should pay, on this collectivist version of the PPP, because they have caused, and will continue to cause, harms due to the high concentrations of greenhouse gases they have released into the atmosphere by their historic polluting. Caney (2005) has argued against a collectivist approach to the PPP on the grounds that it would be unfair to the current citizens of a historically polluting state to have to pay for damages done by their forebears. He asks, "individuals cannot inherit debts from parents or grandparents, so why should this be any different?" (Caney 2005, p. 760).

Caney is essentially restating the intergenerational problem from above, which fails to adequately engage with the rationale of collectivist views.

As a part of a collective, an individual is usually entitled to some benefits, but those benefits come at the cost of certain responsibilities. The entitlements of new citizens of New Zealand, by birth or grant, include the benefits of social welfare, a public health system, and the freedom to live in a naturally beautiful country. However, these new citizens also accrue several responsibilities, including abiding by the law and paying taxes. As a rule, the responsibilities of being a part of a collective come ineluctably hand-in-hand with the benefits. Therefore, individuals who did not vote for the creation of the benefits that they are now enjoying (as a part of a collective) should understand that with those benefits come responsibilities, and that acceptance of the benefits entails acceptance of the whole package. Therefore, while citizens of industrialised countries are innocent of historic polluting, the collective of which they are a part is not. One may decide to opt out of the collective (of both the benefits and the responsibilities), but no one is entitled to opt out of the responsibilities only. In short, one can respond to Caney's worries about the unfairness of collective versions of PPP as follows. If individuals born into rich countries can make the case that it is unfair to require them to pay for harms they did not cause, then individuals born into poor, non-polluting countries can make an even stronger case that it is unfair that they lack so many benefits enjoyed by individuals of rich countries *solely* because of accidents of birth. A collectivist PPP will not be pursued further, however, because both individual and collective versions of PPP are susceptible to a different objection (discussed later in the chapter): the 'ability to pay objection'.

3. Beneficiary Pays Principle

According to the beneficiary pays principle (BPP), agents who benefit from historic polluting should bear the moral responsibility for dealing with the problems caused by that polluting. The BPP easily avoids the intergenerational problem because it assigns moral responsibility to those who benefit from polluting, regardless of whether they have caused any. According to the BPP, the response to Caney's innocent complainer should be, 'I agree that you did not cause historic pollution, but you have benefited from it, and *that* is why you have the moral responsibility to deal with it'.

The strongest ethical rationale for the BPP is based on the idea of minimising the unearned inequalities that have resulted from historical polluting. Unearned inequalities are differences in welfare between agents that have come about because of circumstances beyond the agents' control. As it stands, historical polluting has had two main effects: it has damaged the global ecosystem, causing the climate to warm rapidly, and it has enabled the polluting states to industrialise quickly, causing their current citizens to enjoy a higher quality of life. The BPP answers the problem of who should pay the costs of mitigating and adapting to the effects of the rapid climate change caused by historic polluting, by arguing that it is fairer if agents pay in proportion to the extent that they have benefitted. Given that the most obvious alternative to this approach involves both rich (benefitting) and poor (non-benefitting) states paying equally to cover the costs of dealing with climate change, the BPP seems to capture an important moral distinction.

Many philosophers have criticised the application of the BPP to historical polluting on different intergenerational grounds; namely, the non-identity problem (e.g. Caney 2005). The non-identity problem as applied to the BPP argues that it is impossible for anyone to benefit from historic polluting because if historic polluting had not occurred, then different people would

have been born (different to the current so-called benefiteres). Some responses to the non-identity problem have been given (e.g. Page 2011, pp. 423–424), but their success is irrelevant to this chapter because (as argued below) the BPP, like the PPP, is better rejected on the grounds of the ‘ability to pay objection’.

4. Ability to Pay Principle

The ability to pay principle (APP) regards states’ per capita production capacity (or some other measure of welfare) as the only moral consideration in sharing the responsibilities of remedying the adverse effects of climate change. The APP requires that all and only those who can afford to pay for mitigating and adapting to climate change should pay, and that they should pay in proportion to their ability to pay. Adoption of the APP would result in the Annex I (developed) states paying for historic greenhouse gas emissions,¹⁰⁸ which accords with the element of the principle of CBDR that calls on developed states to bear more responsibility for dealing with climate change *because* they have the ability to do so.¹⁰⁹

There are several ways to discern a state’s ability to pay and even more ways to justify the various methods. I believe that the most promising method for discerning a state’s ability to pay is morally justified by the notion of sufficientarianism. Sufficientarianism is a principle of distribution which holds that benefits and burdens should be shared in such a way that as many people as possible (including future people) have sufficient resources to

¹⁰⁸ For a list of Annex I states, see UNFCCC (no date).

¹⁰⁹ “The developed countries acknowledge the responsibility that they bear in the international pursuit of sustainable development in view of the... technologies and financial resources they command” (UNFCCC 1992).

achieve a certain level of well-being (Page 2007).¹¹⁰ A sufficientarian would argue that a government's primary moral responsibility is to ensure its citizens have a quality of life sufficient for a reasonable level of well-being. Defining a reasonable level of well-being is not a simple task, but considerations of happiness can help.

The potentially catastrophic effects of climate change are not considered a problem because they prevent a few happy people from becoming very happy. Rather, agreement that action on climate change is required is usually motivated by concerns about the millions of people that appear likely to experience the opposite of happiness—suffering. These concerns reflect the widespread notion that suffering seems more harmful than an equivalent amount of happiness seems beneficial. Indeed, the larger disutility of losses compared to the utility of equivalent gains in many domains of life is well established in economics (e.g. Ding, Charoenwong, & Seetoh 2004). Reflected also is the idea that deprivation of the basic physical means for life makes happiness next to impossible; when hurricanes destroy homes in areas where citizens are uninsured and already on the brink of poverty, suffering becomes very likely and happiness becomes a dream for the distant future. People who

¹¹⁰ For more discussion of sufficientarianism, see Frankfurt (1987), Crisp (2003), Page (2008), and Shue (1992; 1999). Sufficientarianism might seem similar to egalitarianism and prioritarianism, but the three views are not the same. See Page's (2007) elegant discussion on this. An egalitarian views states of affairs as being increasingly just as they make individuals' well-being (or some other good) increasingly equal (and vice versa). A prioritarian sees changes in states of affairs as increasingly just as they increasingly prioritise the improvement of the well-being of those who are worst off at every iteration of redistribution. A sufficientarian views states of affairs as being increasingly just as fewer people reside below a sufficient level of well-being. Therefore, in a situation of many unequally rich (but all very rich) agents, egalitarians and prioritarians would recommend redistributing the wealth, but a sufficientarian would not.

have lost their homes, livelihoods, or loved ones score very badly on tests of happiness regardless of how happiness is measured.¹¹¹

With these considerations in mind, it seems that securing a sufficient level of happiness means securing the basic means for life, including housing, sustenance, and protecting people's ability to work and connect with loved ones. For practical purposes, a state's ability to provide these basics for its citizens should be measured by its per capita production. Although often considered anathema to the newer measures of happiness, measures of production have (at least currently) practical advantages over the newer measures of happiness; they are relatively easy to calculate and adequately reflect a state's ability to provide most of the goods that increase its citizens' well-being.¹¹² Furthermore, when the happiness of the worst off is being considered, the basic physical means for life provide a very reliable proxy for happiness, and the traditional measures of production provide a good proxy for a state's ability to provide these basics.

For practical purposes, the level of per capita production considered sufficient to be able to provide these basic goods should be based on international agreement. However, the level of sufficient production should be far enough above the 'poverty line' that malnutrition and preventable diseases are not commonplace, and it should also not exceed the threshold where per capita real income begins to make little difference to subjective well-being.¹¹³ This is because further gains in happiness (and further

¹¹¹ See Diener (2009), Frey (2008), and Graham (2009) for reviews.

¹¹² In the future, a state's ability to provide a sufficient quality of life might include consideration of its natural and cultural resources, such as pristine wilderness and celebrated heritage. The current difficulty quantifying the effects of such non-commercial goods on well-being prevents their inclusion in measures of sufficiency for now.

¹¹³ Of course, estimating the exact threshold for sufficient well-being is so fraught with difficulties that it is probably impossible to get right (Casal 2007, pp. 312–18; c.f. Page 2008,

decreasing of suffering) past this point are no longer best achieved by continuing to increase productivity.¹¹⁴ Indeed, once this point is reached, there is good reason to believe that people are leading more than a sufficiently decent life (based on subjective well-being studies) and also that per capita production becomes a much worse proxy for happiness (since the correlations between happiness and income continue to weaken past this point). Henceforth, any use of 'APP' will mean 'sufficientarian-supported ability to pay principle', unless specified otherwise.

When a state has a sufficient level of production to provide this level of well-being for its own citizens, a sufficientarian would then argue that the state has a moral responsibility to ensure citizens of other states, and future citizens of all states, can also reach this level of well-being. As noted by Shue (1999, p. 542), this responsibility could be either weak or strong, where the strong version calls for positive action to assist others below the level of sufficiency and the weak version requires only that states are not interfered with when attempting to reach the level of sufficiency. For the APP, the strong version applies; the ability to pay for preventing the damage that rapid climate change is likely to cause creates a moral responsibility to do so. I propose that a state's ability to pay for helping other states deal with problems, such as climate change, be understood as the degree to which a state's per capita production exceeds the agreed level of sufficient per capita production. Of course, it could be the case that the government of a very wealthy state distributes its plentiful goods in such a way that some or even many of its citizens are left without the resources required for a sufficiently

p. 565). However, enough work has been done on this to create estimates that are well-enough grounded to proceed with. For example, Baer *et al.* (2007) argue that the minimum should be US\$9,000 (in 1995 dollars).

¹¹⁴ Again, see Diener (2009), Frey (2008), and Graham (2009) for reviews.

good life. Although I do not wish to trivialise this issue, I set it aside here as a matter to be resolved between citizens and their governments. Thus, for present purposes, within-state distribution of income does not affect the objective assessment of whether a state has the ability to pay for protecting the global commons.

Using the idea of sufficientarianism as the moral justification for the APP, it can be seen that if the citizens of a poor state do not have a decent standard of living, then that state has no obligation to pay for helping citizens of other states. In contrast, if a rich country has the ability to provide *more* than a minimally happy life for its citizens, it is obliged to help pay for the prevention of harm to citizens of less fortunate states. A consequence of this sufficientarian justification for the APP is that a very poor country, such as Bhutan, could start polluting now without incurring any moral responsibility to mitigate its greenhouse gas emissions or pay to help others adapt to the rapidly changing climate. In fact, this view advocates that Bhutan's hypothetical polluting be paid for by all states that achieve above a certain level of production (and to the degree that their relative productions exceed this level). Although this consequence might seem unfair, and thereby pose a problem for the APP, it actually highlights a benefit of it. The current debate about climate change would never have come about if it were not for the potentially catastrophic consequences for humans, and the APP is designed to minimise the number of people living in appalling and thereby potentially catastrophic circumstances. The APP, as defined here, ensures that people who are most likely to suffer as a result of rapid climate change are the central concern of any agreement on dealing with climate change.¹¹⁵ Focusing on

¹¹⁵ This result nicely reflects several of the principles of the Rio Declaration on Environment and Development (UNFCCC 1992). Principle 1 states: "Human beings are at the centre of concern for sustainable development". Principle 5 states: "All States and all people shall cooperate in the essential task of eradicating poverty as an indispensable requirement for

those who are suffering in this way reflects the moral desire to have as few people as possible suffer from the existence of 'radical inequalities', which Nagel (1977) describes as situations in which there is enough of some good for everyone but some parties have much more than enough and others have less than enough.

Although the APP captures one vital moral consideration, it fails to address another. In particular, the APP fails to assign fair distributions of moral responsibility in the case of rich states with an equal ability to pay but differing levels of culpable carbon-emitting. According to the APP, richer states' ability to pay for the costs of mitigating and adapting to climate change justifies their doing so *regardless* of whether they knowingly contribute to the problem. The omission of this consideration departs from the original text of the principle of CBDR, which ties developed nations' responsibilities for dealing with the costs of climate change to their greater role in producing it (as well as their greater ability to pay). More importantly, the failure to consider who knowingly created the climate-changing pollution creates unfair burdens on equally rich but non-polluting states.

5. Ability to Pay Objection to the Polluter Pays Principle and Beneficiary Pays Principle

Despite its weaknesses, considerations of a sufficientarianism-based ability to pay principle highlight an important problem for both the PPP and the BPP.

sustainable development, in order to decrease the disparities in standards of living and better meet the needs of the majority of the people of the world". Principle 6 states: "The special situation and needs of developing countries, particularly the least developed and those most environmentally vulnerable, shall be given special priority. International actions in the field of environment and development should also address the interests and needs of all countries".

The ability to pay objection has not been afforded much attention in the literature, especially as an objection to the BPP.¹¹⁶ I argue that the ability to pay objection reveals how the PPP can unfairly assign moral responsibility to some agents without the ability to pay for it and that the BPP can fail to assign moral responsibility to some agents that do have the ability to pay for it.

If the notion of sufficientarianism is taken seriously, then very poor states should focus on their primary moral responsibility; the welfare of their citizens, and not the welfare of citizens belonging to other states. However, if the BPP, or either the individualist or collective version of the PPP, are taken seriously, then some currently polluting, but still very poor, states will be morally obliged to help much richer states deal with the deleterious effects of climate change. According to the PPP, even if a state lacks the resources to provide a minimally happy life for its citizens, it must pay for any polluting that it does. Indeed, this remains the case even if the polluting is the result of efforts to improve the abysmal living conditions of the state's citizens. As a practical matter, no beneficent (or even slightly self-interested democratic) government would voluntarily deprive its own citizens of the basic goods of life to relieve some richer people from (what is to them) a tiny burden. And, more importantly, it would be unfair to obligate such a government to do so. Imagine a state that emits greenhouse gases above the accepted per capita quota in an attempt to rebuild its capacity to offer basic services to its citizens after being ravaged by famine and war. *Even if it is currently polluting*, it would be grossly unjust to require the government of such a state to give resources badly needed by its own citizens to some international fund so other citizens (most of whom already enjoy a happier life) do not have to contribute as much themselves. The injustice in this case arises because the

¹¹⁶ For example, Caney (2005) uses a weaker version of this objection to argue that the PPP needs to be supplemented with the APP, but he does not use this to object to the BPP.

plight of the very worst off is ignored so that those who are already better off can receive some benefit—a clear instance of exacerbating a Nagellian radical inequality.

Another version of the ability to pay objection reveals how the BPP can also fail to assign moral responsibility to some agents that do have the ability to pay for it and seem equally culpable as other agents who are expected to pay. Imagine two states that have the same ability to pay, although one state has acquired its ability through historical non-polluting activities and the other from historical pollution-causing industrialisation. Let us further imagine that the latter state switched over to entirely renewable energy in the relatively recent past, and that none of its current citizens have ever produced any greenhouse gases. According to the BPP, the historically polluting state should bear moral responsibility for mitigating and adapting to climate change because it has benefited from greenhouse gas emitting, while the non-polluting state should bear none. It might be argued that it is fair for citizens of such non-polluting states to bear no responsibility to deal with climate change because they have neither caused nor benefited from it. However, a comparison of these citizens with the citizens of the benefiting state reveals that both are equally innocent of polluting and both enjoy unearned benefits. In both cases, the polluting and non-polluting actions that led to the benefits were beyond the current citizens' influence. So, in both cases, the benefits the current citizens enjoy are unearned. In the same way, neither generation of citizens caused the historic polluting (or non-polluting) that may have led to the current pollution, so they are equally (totally) innocent of causing the pollution. To assign more moral responsibility to the citizens of one state because the benefits that they have *just happen* to have come from historical greenhouse gas emitting, instead of some other non-polluting actions, is clearly unfair. Why should the citizens of a historically polluting state have to

pay more for their unearned benefits when they had an equal (total) lack of ability to affect how those benefits came about?

The PPP and BPP both represent what appear to be important moral considerations for the climate change debate; both polluting and benefiting from polluting appear to create some moral responsibility to deal with the harmful consequences of that polluting. However, it has been argued here that the often-neglected ability to pay objection shows us two things.

First, although polluting states should generally pay, in some circumstances they should *not* have to pay; specifically when they lack the ability to provide the material requirements for a minimally happy life to their citizens. As a result of this, if the PPP were used to explain the principle of CBDR by itself, then the consequent apportionment of moral responsibility to deal with climate change might be unfair.

And second, having benefited *from polluting* (as opposed to any other historical actions of our forebears) is actually not an important moral consideration because the sources of our intergenerational benefits are always essentially out of our control. This finding reveals that, on the face of it, the plausible moral justification for the BPP—that those who benefit *from polluting* have more responsibility to deal with the resulting pollution—does not reflect our considered judgements about what is really important in deciding who should have to bear the responsibility to deal with climate change.

6. The Hybrid Account

So far, it has been argued that none of the three main principles of justice discussed in climate ethics are without major problems of unfairness. The most natural explanation for this is that more than one moral consideration is relevant to the issue of distributing the responsibilities of dealing with climate change. Thus, it is unsurprising that some authors have offered hybrid

accounts that reflect what they believe to be the relevant combination of moral considerations for this issue (e.g. Caney 2005; Page 2008).

The hybrid account proposed here (henceforth the hybrid account) combines elements of the APP and conditional liability PPP (CPPP). By combining these principles, the hybrid account fits nicely with the dual rationale behind the principle of CBDR. Specifically, it matches the only two reasons given in the principle of CBDR for why developed states should bear the lion's share of the responsibility to deal with climate change: developed states put greater pressure on the environment and they possess greater technological and financial abilities to facilitate the mitigation of and adaptation to rapid climate change. The hybrid account bears similarities to that of Caney (2005), but differs in specific details. Most importantly, the hybrid account is justified differently to Caney's hybrid account. In contrast to Caney's rights-based approach, the hybrid account is informed by considerations of happiness and is based on the distributive principle of sufficientarianism. Indeed, sufficientarianism is used as the underlying moral justification for the inclusion of the APP and the CPPP and for giving priority to the APP over the CPPP.

The main potential problem of rapid climate change is the increased numbers of people who will likely lead miserable lives as a result. Bearing this in mind, the hybrid account appeals to the APP to ensure those who already lack a minimally happy life are not put under more pressure. For similar reasons, the hybrid account appeals to the CPPP to encourage much lower levels of emissions and, thereby, decrease the chances of catastrophic climatic changes that would plunge even more people (especially future people) into desperate poverty.

The hybrid account is also potentially useful to policy makers. Specifically, it should be useful as an aid to the fair assignment of responsibilities in

current and possible future situations for both developed and developing states. To do this the moral agents to which the hybrid account applies need to be states. Although applying the hybrid account directly to individuals is theoretically unproblematic, it suffers from obvious practical problems. Suffice it to say that arriving at a global agreement on climate change is difficult enough when the negotiators are tens of states, let alone billions of individuals.

Given that the relevant moral agents are states, I propose the following process for deciding how the responsibilities of dealing with climate change should be distributed. First, assess which states have the ability to pay. I propose that there should be three broad categories of ability to pay (no ability to pay, clear ability to pay, and unclear ability to pay) each of which corresponds to a state's ability to provide a minimally decent quality of life for its current citizens.

States with no ability to pay should not be assigned any moral responsibility to deal with climate change because of their overriding responsibility to raise the standard of living of their own citizens to a sufficient level.

States that have a clear ability to pay should have to pay for their own greenhouse gas emissions above their quota and their share of any remaining costs (based on how many other states are in this category and how much each of these states exceeds the lower limit of the band).

States that have an unclear ability to pay should have to pay for their own greenhouse gas emissions above their quota, but they should not have to help the rich states pay to deal with any outstanding pollution. I propose that 3,000 international dollars of gross domestic product (purchasing power parity) per capita (GDP-PPP-PC) should be the upper limit of the no ability to pay band and 7,000 international dollars of GDP-PPP-PC should be the lower limit of

the clear ability to pay category.¹¹⁷ States with an unclear ability to pay would be those with 3,000–7,000 international dollars of GDP-PPP-PC. For perspective, in 2009, the United States' GDP-PPP-PC was 46,433 international dollars, China's was 6,546 international dollars, and India's was 2,932 international dollars (India's projected GDP-PPP-PC for 2014 is 4,285 international dollars).¹¹⁸ Although I have set these thresholds according to what I believe roughly corresponds to what it takes to provide a minimally decent quality of life, a complete justification of these suggested figures would require much more research than is available and is certainly beyond the scope of this chapter.

After assessing which states have the ability to pay, the amount of existing pollution that has been 'culpably caused' (that is, caused after 1992) by each state with a 'clear' or an 'unclear' ability to pay should be gauged. Any agent with the ability to pay for dealing with the pollution it has caused must do so in full. Any remaining pollution that needs to be dealt with must have been created by past generations, the currently very poor, or rogue non-complying states. Since neither the dead people nor the poor states have the ability to pay for dealing with this pollution, and the rogue states refuse to pay, the question arises of who should have to pay for it.

It might be suggested that no one should have the moral responsibility to pay for the outstanding pollution. However, if the precautionary principle and the right to development principle from the Rio Declaration on

¹¹⁷ International dollars is a hypothetical currency that has purchasing power equivalent to the US dollar at a particular time. A state's purchasing power parity-adjusted annual gross domestic product is a measure of how many standardised baskets of goods that state could afford to buy at domestic prices if its total production output for the year were in money. Indexing this measure to international dollars allows for rough but meaningful across-state and across-time comparisons of how well a state can provide the basics of life.

¹¹⁸ These figures are from the International Monetary Fund (IMF) (2009).

Environment and Development are adopted,¹¹⁹ it must be concluded that someone must pay for the outstanding pollution to avoid the risk of an environmental catastrophe that could have devastating effects for billions of future people. Assuming, then, that someone should bear the responsibility to pay for dealing with the outstanding pollution, a fair method for deciding *who* those bearers should be is required.

Appealing once again to the idea of sufficientarianism, and more specifically to the idea that states that have fulfilled their responsibility to raise their own citizens' welfare to an acceptable standard consequently have a responsibility to ensure that all people (including future people) can reach this level of welfare, the distribution should be on an ability to pay basis. Therefore, after those who can pay for the costs of their own polluting have done so, the remaining costs should be distributed between those who can afford to pay them (states with a 'clear' ability to pay) and apportioned based on each state's degree of ability to pay. These costs should include the cost of creating institutions to encourage and, if necessary, enforce compliance of rogue states, as Caney (2005) recommends. This method of distribution is the fairest because any other method would result in the possibility of the governments of the worst-off states in the world having to forego their primary moral concern (their citizens' welfare) for the sake of generally much wealthier people elsewhere and in the future (the ability to pay objection from above).

¹¹⁹ Principle 15 of the Rio Declaration on Environment and Development states: "In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation". Principle 3 states: "The right to development must be fulfilled so as to equitably meet developmental and environmental needs of present and future generations" (UNFCCC 1992).

In practical terms, the hybrid account results in the rich states paying for their own polluting and then sharing the costs associated with both the minor amounts of polluting caused by rogue states and very poor states and the historic pollution that was unknowingly caused by previous generations. The sufficiency-based hybrid account gives primacy to the APP over the CPPP. This results in the hybrid account deeming it morally permissible for undeveloped and developing states to knowingly pollute, *but only* if that pollution is likely to result in higher well-being for their citizens. Because undeveloped countries can justifiably prioritise meeting the basic needs of their own citizens over the less urgent needs of future people, they can knowingly pollute on the hybrid account without incurring the moral responsibility to deal with that pollution.

Contrasting the hybrid account against those that prioritise the CPPP over the APP, the main difference between them is that the hybrid account goes further to eliminate radical inequalities in the essentials of life between the very rich and the very poor. On the CPPP prioritised-hybrid accounts, very poor states do not get to fast-track themselves to a minimally good quality life for their citizens through rapid industrialisation because that would entail polluting above their per capita quota for a time. This goes against Principle 5 of the Rio Declaration on Environment and Development, which encapsulates the global agreement to try to eradicate poverty.¹²⁰ More importantly, though,

¹²⁰ Principle 5 of the Rio Declaration on Environment and Development states: "All States and all people shall cooperate in the essential task of eradicating poverty as an indispensable requirement for sustainable development, in order to decrease the disparities in standards of living and better meet the needs of the majority of the people of the world". However, Principle 2 would endorse a CPPP-APP hybrid account that prioritises CPPP over our account. Principle 2 states: "States have, in accordance with the Charter of the United Nations and the principles of international law, the sovereign right to exploit their own resources pursuant to their own environmental and developmental policies, and the responsibility to

denying poor states the benefits of rapid industrialisation is unfair because it robs their citizens of the opportunity to catch up with citizens of developed states (which were not prevented from industrialising in the past). Without the chance to rapidly industrialise, undeveloped states will remain economically insignificant and continue to be forced to draw the short straw on international trade agreements.

Furthermore, if the hybrid account is enacted, then developed states should help developing and undeveloped states adapt and develop so that they can meet the basic needs of their citizens *without* polluting. On the hybrid account, the wealthy countries have the responsibility for dealing with all outstanding emissions, such as those created by states without the ability to provide a minimally good life for their citizens. Therefore, it is in their best interest to ensure that they fulfil that responsibility by giving enough technology and training to developing countries to provide a strong incentive for them to industrialise in a way that creates minimal greenhouse gas emissions. This plausible way to fulfil the moral responsibility to deal with historical emissions will help to ensure that very poor countries still get the welfare benefits of industrialisation and that few if any extra problems are created for future people.

7. Dealing with Some Potential Objections

As discussed above, an implication of the hybrid account is that a very poor country could emit greenhouse gases above its per capita allowance and incur no moral responsibility to deal with the effects of that emitting. Shue warns that:

ensure that activities within their jurisdiction or control do not cause damage to the environment of other States or of areas beyond the limits of national jurisdiction" (UNFCCC 1992).

If whoever makes a mess receives the benefits and does not pay the costs, not only does he have no incentive to avoid making as many messes as he likes, but he is also unfair to whoever does pay the costs. (Shue 1999, p. 533)

On the first point, the hybrid account allows poor states to make a mess only if it increases the well-being of their citizens and it encourages rich states to incentivise low pollution-causing industrialisation for the poor states through technology transfer. Furthermore, the following stipulation can be added in order to enhance both of the above points: 'poor states are permitted to emit greenhouse gases over the per capita limit *only* if no other means is available to increase their well-being for a similar cost'. Most importantly, poor states are only permitted to pollute without incurring the responsibility to pay for the resulting damage for a limited time (until they can provide a minimally happy life for their citizens).

On the second point, in the situations in which the rich pay for the pollution of the very poor they do so (and do so *fairly*) because of their moral responsibility to the people of the world who do not have access to the basics for a minimally happy life. This responsibility arises because the ability to secure the goods of a minimally happy life for ourselves and for others, which the citizens of rich states have, is mainly a product of chance, not deservedness. Moreover, these inequalities have been preserved and exacerbated to the further detriment of people who happen to be born into very poor countries.

Also mentioned above, Page (2008) has criticised the use of 1992 as the year after which states should have known that greenhouse gas emissions over a certain level create standing harms. According to Page (2008, p. 570), using 1992 results in "harsh treatment for the newly industrialised

populations and lax treatment of those residing in countries of transition". Although this creates a problem for the PPP by itself, it is not a problem for the hybrid account. On the hybrid account, newly industrialised states are likely to have a low degree of ability to pay compared to states that industrialised over 100 years ago. So, newly industrialised states will only have the responsibility to pay for a fraction of what the more established developed states have to pay. Furthermore, if a newly industrialised state happens to be as wealthy as the states that industrialised long ago, then they *should* have to pay as much as the more established states because they are all lucky enough to enjoy the benefits of wealth that came to them through the actions of previous generations (which were completely outside of their control).

As for the states in transition (those that are very close to the upper limit of not having the ability to pay), the hybrid account makes it morally permissible for them to pollute in order to complete the transition. However, they will incur the moral responsibility to pay for dealing with their own pollution as soon as they reach the threshold of sufficiency (when they will have an 'unclear' ability to pay—at 3,000 international dollars of GDP-PPP-PC). With this in mind, states in transition would be better off accepting technological assistance from developed states so they can complete the transition in a low pollution-causing manner and not have to undergo a comprehensive and costly energy-production transition when they achieve the level of production that allows them to provide a minimally happy life for their citizens. Therefore, on the hybrid account, states in transition and very poor states are assigned less responsibility than developed states, because this distribution of responsibilities will help to reduce the most important inequalities between states, namely, the unequal distribution of the basic goods for a minimally happy life.

8. Conclusion

In this chapter, it has been argued that a hybrid account made up of the APP and the CPPP is the best way to interpret the principle of CBDR. More specifically, the hybrid account provides the fairest guidance for sharing the responsibilities of dealing with climate change. The hybrid account is unusual in its prioritisation the APP over the CPPP. This prioritisation stems from a distinctive rationale for why the hybrid account on offer here should be preferred over each of the individual principles that have been discussed in the literature; that rationale was that the most important moral consideration in the debate over climate justice is that each government's primary responsibility is to raise its own citizens' welfare to a sufficient level for them to have a minimally happy life. This conclusion was reached after considering the importance of eliminating extreme and on-going unhappiness in people's lives and what is required to bring this about.

The most important practical implication of the hybrid account is that undeveloped and developing states can continue to pollute without incurring any moral responsibility to deal with the effects of that pollution, as that polluting is the best way to achieve the agreed upon level of welfare for their citizens. This result is grounded in the conviction that the fundamental purpose of the current climate justice debate is (and should continue to be) to ensure that a minimally acceptable level of happiness is and will continue to be attainable for all people.

Conclusion

This thesis has discussed several distinct theoretical and applied happiness-related questions and, in doing so, has demonstrated the potential breadth and depth of happiness-related areas of inquiry.

Regarding theoretical happiness-related questions, the aims of this thesis have been to elucidate the multifarious definitions of hedonism and to show that (contrary to popular opinion amongst philosophers) the experience machine objection to hedonism should not be viewed as a convincing objection to internalist accounts of Prudential Hedonism or any other internalist mental state theories of well-being.

Regarding the applied happiness-related questions, this thesis aimed to demonstrate that considerations of happiness can help us to better understand what we might do to pursue a truly meaningful life and to fairly distribute the burdens of climate change. Furthermore, in the Postscript for Policymakers that follows, philosophical considerations are used to guide policymakers in their use of the findings from the science of happiness.

Taken together, the assorted chapters of this thesis also demonstrate that interdisciplinary analyses can help resolve some of the problems that can occur when research on an area of inquiry becomes insulated from other relevant disciplines. Although many philosophers have investigated research from related disciplines at some point, very few are currently making a point of taking a fully interdisciplinary approach whenever it is appropriate (Tiberius & Plakias 2009, p. 402).¹²¹ Particularly for research on happiness and well-being—where there is a wealth of relevant research being produced by

¹²¹ Dan Haybron, Valerie Tiberius, and Erik Angner are perhaps the only obvious exceptions (e.g. Angner 2011; Haybron 2007; 2008b; 2008c; Haybron & Tiberius 2012; Tiberius 2008; Tiberius & Plakias 2009).

researchers in other disciplines—this mono-disciplinary orthodoxy seems to be unfortunate at best and unproductive at worst.

Part 1 of this thesis, for example, is a sustained analysis of how psychological insights and methods can challenge a received belief in philosophy that has held sway for nearly 40 years. Indeed, by the end of Chapter 5 the psychological insights and experimental methods that I employ hopefully make it clear that Nozick's original experience machine thought experiment tends to elicit biased judgements. And this has serious implications for the widely-believed experience machine objection to hedonism.

Chapter 6 also demonstrates the usefulness of interdisciplinary analysis. Scientific and technological theories from physics and related sciences are used to open up new possibilities in the old philosophical debate about the meaning of life. The fact that the scientific and technological theories are not entirely new gives some reason to suspect that philosophers working on the meaning of life might benefit from looking to other disciplines for inspiration to enliven their old debates.

The opportunities for future research on happiness-related questions are clearly numerous. And the importance of investigating happiness-related questions has not diminished because happiness remains possibly the most plausible candidate for an "objective, non-instrumental, non-conditional good" (Moore & Crisp 1996, p. 606). In particular, the application of the science of happiness to public policy is an area that will benefit from the input of philosophers as well as academics from other disciplines. The Postscript for Policymakers includes a discussion of the role that philosophers should play in this very important happiness-related area of inquiry.

One of the main difficulties for future research on these questions will be the ability to keep up to date not only with the philosophical work on

happiness but also with all of the other research that is relevant to the particular happiness-related question under investigation. Still, the importance of research on happiness-related questions is unquestionably worth the considerable effort required to stay abreast of the current research; knowledge about happiness remains both a valuable commodity and something that is intrinsically interesting to pursue. Indeed, unlike Darrin McMahon (2006, p. xi)—who claims that, at times while writing *Happiness: A History*, he was “forced to confront the irony that writing a book on happiness might make [him] miserable”—I have thoroughly enjoyed tackling happiness-related questions and hope to continue doing so as long as I am able.

Postscript for Policymakers

The Science of Happiness for Policymakers

Postscript Summary

With several national governments and multinational organisations investigating new measures of progress and well-being to inform policymaking, some researchers have called for measures of happiness to be among those investigated. This Postscript for Policymakers provides a high-level tour of this debate. Some objections to the most promising measures of happiness are discussed. Given that the intended audience of this Postscript for Policymakers is policymakers, and given that this interdisciplinary debate is vast in scope, analysis of some of the objections is not nearly as fine-grained or thorough as the analyses in Chapters 2 to 7 of this thesis.

After providing a brief high-level analysis of this debate, it is argued that measures of happiness should play an important role in policymaking. However, it is also noted that several important obstacles need to be overcome before any measure of happiness could play such a role. The problems are manifold and will require considerable interdisciplinary work to overcome. Nevertheless, it is clear that the importance of happiness should make this work a priority.

1. Introduction¹²²

Traditional economic indicators of progress are widely seen to be insufficient as indicators of well-being (Michalos 2011; Stiglitz, Sen, & Fitoussi 2009). Even

¹²² This chapter provides the basis for Turton (2009), Weijers (2010), and Weijers (2011d). Turton was my previous name.

if the traditional economic measures of per capita production, income, and wealth took all relevant production, income, and wealth into account (which they don't), they would still fail to capture the value of our relationships, health, and happiness—all of which are typically viewed as important for well-being. Many alternate measures of well-being have been developed over the years to address this problem, and governments have slowly incorporated some of them into the policymaking process at various stages. Over the last few years, the debate about whether happiness should be measured and used as an indicator of progress, and to inform policymaking, has intensified. This chapter provides a high-level tour of this debate, includes an investigation into the severity of the problems of using measures of happiness for policymaking, and pays particular attention to measures of subjective well-being (predominantly survey questions about how happy or satisfied respondents are with their lives).

First, I discuss how we have arrived at the point where using measures of happiness is being seriously considered by policymakers. Then I explain the most promising methods for measuring happiness. Following this, I discuss several important criticisms of these measures. Some are found to be misguided, but others are found to be problems that need addressing before measures of happiness can usefully be employed by policymakers. The main criticisms addressed include: happiness data cannot tell us anything, we cannot know what measures of happiness are really measuring, and that the wrong kind of happiness is being measured. Finally, I provide recommendations for the role that suitably-improved measures of happiness could and should play in policymaking and what steps should be undertaken to suitably improve these measures.

This chapter concludes that happiness should indeed be measured and used by governments and civil servants to inform policymaking. However,

much complex interdisciplinary and international research will be required before measures of happiness can fruitfully play such a role.

1.1 Why Measure Happiness?

The limits of Gross Domestic Product (GDP) and Gross National Product (GNP) have been recognised by economists, politicians, and others for a long time (Michalos 2011). In particular, GDP and GNP have been criticised as measures of progress because of their myopic focus on production. These measures were never intended to be complete measures of progress, rather they were intended to reveal the rate at which the economy was growing or shrinking in a country (England 1998). However, the pursuit of economic growth has dominated the agendas of nearly all national governments and as a result GDP became the main indicator of progress almost by default. Some of the problems with GDP and GNP were emotively elucidated by Robert F. Kennedy in his famous speech at the University of Kansas in 1968:

“But even if we act to erase material poverty, there is another greater task; it is to confront the poverty of satisfaction—purpose and dignity—that afflicts us all. Too much and for too long, we seemed to have surrendered personal excellence and community values in the mere accumulation of material things. Our Gross National Product, now, is over \$800 billion dollars a year, but that Gross National Product—if we judge the United States of America by that—that Gross National Product counts air pollution and cigarette advertising and ambulances to clear our highways of carnage. It counts special locks for our doors and the jails for the people who break them. It counts the destruction of the redwood and the loss of our natural wonder in chaotic sprawl. It counts napalm and counts nuclear

warheads and armored cars for the police to fight the riots in our cities. It counts... the television programs, which glorify violence in order to sell toys to our children. Yet the gross national product does not allow for the health of our children, the quality of their education or the joy of their play. It does not include the beauty of our poetry or the strength of our marriages, the intelligence of our public debate or the integrity of our public officials. It measures neither our wit nor our courage, neither our wisdom nor our learning, neither our compassion nor our devotion to our country; it measures everything, in short, except that which makes life worthwhile. And it can tell us everything about America except why we are proud that we are Americans.”¹²³

In response to these and other shortcomings of GDP and GNP, economists, national statisticians, and interested non-governmental organisations began to investigate and measure many other aspects related to progress. In addition to broadening and refining the existing range of economic measures, these initiatives also led to the collection of data related to individual well-being, quality of life, and even happiness. For many years now, academics from several disciplines and various countries, and even some civil servants, have been increasingly pushing for these new measures of well-being to play more important roles in policymaking. And, over the last few years, politicians have finally begun to listen. In 2008, French President Nicolas Sarkozy chartered the Commission on the Measurement of Economic Performance and

¹²³ Transcript of Robert F. Kennedy's speech is available from:

http://www.glaserprogress.org/program_areas/pdf/Remarks_of_Robert_F_Kennedy.pdf

Social Progress. The Commission, headed by Nobel-winning economist Joseph Stiglitz, advised that their report is...

... addressed, first of all, to political leaders. In this time of crises, when new political narratives are necessary to identify where our societies should go, the report advocates a shift of emphasis from a “production-oriented” measurement system to one focused on the well-being of current and future generations, i.e. toward broader measures of social progress. (Stiglitz, Sen, & Fitoussi 2009, p. 10)¹²⁴

The Commission mentions measures of subjective well-being briefly, but positively, encouraging national statistical offices to “incorporate questions to capture people’s life evaluations, hedonic experiences and priorities in their own surveys” because “[m]easures of subjective well-being provide key information about people’s quality of life.” (Stiglitz, Sen, & Fitoussi 2009, p. 58). The commission’s report seems to have had the required effect on British Prime Minister David Cameron, who gave a speech announcing the investigation of and subsequent measurement of well-being (including subjective well-being) by the British government on 25 November 2010:

[F]rom April next year, we’ll start measuring our progress as a country, not just by how our economy is growing, but by how our lives are improving; not just by our standard of living, but by our quality of life.¹²⁵

¹²⁴ The report is available from: http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf

¹²⁵ Nowhere in his speech does Cameron mention subjective well-being, but the questions, his responses, and the related media coverage of the speech reveal that the most important part of Cameron’s initiative (often referred to as his ‘happiness agenda’) is his request of the

Demonstrating that civil servants can also take the lead on incorporating new measures of well-being into policymaking, the Treasuries of Australia and New Zealand have independently developed a Wellbeing Framework and a Living Standards Framework, respectively (Australian Treasury 2006; Treasury 2011). Endowed with the vision of encouraging “higher living standards for New Zealanders” (Treasury 2010, p. 1) since its inception, the New Zealand Treasury finally produced the Living Standards Framework in 2011 to provide guidance on what improving living standards actually amounts to (Treasury 2011). The Living Standards Framework outlines the importance of human, social, and natural capital in addition to traditional economic and physical capital for increasing living standards (Treasury 2011). The Living Standards Framework also sets out the role of subjective measures of well-being as providing a “cross-check of what is important to individuals” (Treasury 2011, p. 1).

Surprisingly, supranational organisations may even end up leading the way on promoting the use of broader measures of well-being in policymaking. In April this year, the United Nations (UN) General Assembly will hold a high-level meeting on ‘Happiness and Wellbeing: Defining a New Economic Paradigm’. The agenda for this meeting includes the use of measures subjective well-being. Indeed, the World Happiness Report commissioned for the meeting contains a chapter on The State of World Happiness that relies exclusively on research using measures of subjective well-being because “they capture best how people rate the quality of their lives” (Helliwell & Wang forthcoming).

United Kingdom’s Office of National Statistics to start measuring subjective well-being. The official transcript of David Cameron’s speech is available from:
<http://www.number10.gov.uk/news/pm-speech-on-well-being/>

Far ahead of the UN, the Organisation for Economic Cooperation and Development (OECD), has identified measures of subjective well-being as essential for a complete understanding of well-being:

For over fifty years, the [OECD] has helped governments design better policies for better lives for their citizens... Ever since the OECD started out in 1961, GDP has been the main factor by which it has measured and understood economic and social progress. But it has failed to capture many of the factors that influence people's lives, such as security, leisure, income distribution and a clean environment... The OECD Better Life Initiative allows a better understanding of what drives the well-being of people and nations and what needs to be done to achieve greater progress for all... [T]he OECD has identified 11 dimensions as being essential to well-being, [including] overall satisfaction with life... (OECD 2012).¹²⁶

Lord Richard Layard is the most prominent proponent of the use of measures of subjective well-being, recommending that “quality of life, as people experience it, has got to be a key measure of progress and a central objective for any government” (Layard 2011, no page).¹²⁷ Indeed, Layard has even argued that measures of subjective well-being should be the *main* yardstick for public policy because happiness is the most important goal in life for most of us (Layard 2005, pp. 224–225). Happiness, Layard claims, is “what people

¹²⁶ The OECD Better Life Index Executive Summary is available here:

http://oecdbetterlifeindex.org/wpsystem/wp-content/uploads/2012/02/YourBetterLifeIndex_ExecutiveSummary3.pdf

¹²⁷ Similar positions are held by Derek Bok (2010), Ed Diener (2011), and Bruno Frey (2008), but none of these authors take the more extreme stance presented by Layard in his (2005) book *Happiness: Lessons from a New Science*.

want for their children and for their fellow citizens” and thereby concludes that “the greatest happiness of all” deserves to be the ultimate goal of governments and policymakers (Layard 2005, pp. 124–125). Layard understands happiness as meaning “feeling good—enjoying life and wanting the feeling to be maintained” and believes that the emerging field of happiness science has come far enough for us to be able to accurately measure this kind of happiness using subjective survey questions about happiness and satisfaction with life (2005, p. 12). According to Layard, and many others, we should measure happiness precisely because we should be using data from those measures to inform policymaking.¹²⁸

Given the reasonable assumptions that sustainably, justly, and equitably increasing the well-being of people should be at least a very important goal for public policy, and given that our subjective judgements about happiness are at least a fairly important part of our well-being, it is clear that policymakers should at least investigate the practicality of measuring subjective happiness. Furthermore, if the citizenry of a democratic state demand that its government includes happiness as one of its overarching goals, then there is overwhelming reason for policymakers to find a way to make measuring happiness feasible.

David’s Cameron’s pronouncement that subjective well-being will be measured in the United Kingdom has been met with many negative and positive comments on blog sites, but the wider public seems to support the notion that happiness should be measured and the results of that

¹²⁸ The use of subjective measures of happiness as the only or ultimate criterion to assess specific policies or progress in general has been criticised by philosophers and economists on many grounds, including most of the traditional philosophical objections to Prudential Hedonism (e.g. Hausman 2010; Diener & Scollon 2003; Frey & Stutzer 2007). But see also Veenhoven (2010) for a discussion of how the practice of maximising happiness may avoid many of these theoretical problems.

measurement should inform policymaking. In a 2005 BBC opinion poll, 1001 participants were asked whether the government's main objective should be the "greatest happiness" or the "greatest wealth" and 81% thought that happiness should be the main goal (Easton 2006, no page).¹²⁹ This result closely resembles that of a poll taken on The Economist's website during a debate between Richard Layard and Paul Ormerod, in which the motion "new measures of economic and social progress are needed for the 21st-century economy" received 83% of the support from the online audience (of unspecified size).¹³⁰ If these results are representative of popular opinion, then, in democracies at least, we should investigate the practicality of measuring happiness for policymaking. Furthermore, since happiness is ubiquitously understood as being a subjective state, then this means we should investigate the practicality of measuring subjective well-being for policymaking.

1.2 The Basics of Measuring Subjective Well-Being

Whether the use of measures of subjective well-being for policymaking should be pursued depends not only on how important happiness is to people, but also on whether happiness can actually be efficiently and effectively measured. Of all the potential methods of measuring happiness, only questions asking for survey respondents' judgment about how happy or satisfied they are with their life, are practical on anything but a very small scale. I intend online and smartphone surveys to be included in this. And

¹²⁹ The survey results are available here:

http://news.bbc.co.uk/1/hi/shared/bsp/hi/pdfs/29_03_06_happiness_gfcpoll.pdf

¹³⁰ The audience was well aware that measures of subjective well-being were the new measures in question, since that is all that the debaters and commentators discussed. See the results of the debate here: <http://www.economist.com/debate/overview/204>

indeed as these methods of surveying are becoming cheaper and more accessible, they are being used more and more. Collecting happiness data with behavioural measures, such as expert observations, or any of the neuroimaging techniques, would be prohibitively expensive. Furthermore, it is far from clear that any of these more objective measures are any better at capturing how happy someone is than simply asking them. Indeed, the success of neuroimaging measures of happiness is sometimes assessed by the size of their respective correlations with the participants' responses to subjective well-being survey questions (e.g. Urry *et al.* 2004).

Subjective measures of well-being can be global or domain-specific. Global measures aim to assess respondents' judgments of their lives as a whole, while domain-specific measures target limited aspects of respondents' lives, such as their work lives or their family lives. Although domain-specific measures undoubtedly have their uses, the focus here will be on global measures because they provide a better approximation of the term 'happiness' as it is normally understood.

There is a wide range of global subjective well-being questions, but most are subtle variants of general questions about happiness or satisfaction with life. For example, the United States' General Social Survey asks: "Taken all together, how would you say things are these days? Would you say that you are very happy, pretty happy, or not too happy?" (Kahneman & Krueger 2006, p. 6). The World Values Survey asks: "All things considered, how satisfied are you with your life as a whole these days?", and uses a response scale ranging from "1 (not at all satisfied)" to "10 (very satisfied)" (e.g. Inglehart *et al.* 2008). The subtle variations on these questions usually amount to changing the number of points on the response scale or slightly adjusting the wording of the question. For example, the World Values Survey also asks the following question about happiness: "Taking all things together, would

you say you are..." with a 4-point response scale: "Very happy... Rather happy... Not very happy... Not at all happy."¹³¹

It is widely acknowledged that global subjective well-being questions elicit responses that are biased by contextual factors, the wording of questions, the order and type of preceding questions, and respondents' current mood (Kahneman & Krueger 2006; Schwarz & Strack 1999). Experiments have shown, for example, that contextual factors, such as the weather (Schwarz & Clore 1983) and unexpectedly finding a dime (Schwarz 1987), significantly affect how satisfied participants reported being *with their whole lives*. Experiments on the variability of self-reported satisfaction with life within individuals have demonstrated that people's reported satisfaction with life as a whole changes dramatically over a period of a few weeks. Kahneman and Krueger (2006, p. 7), for example, found that 218 women who were interviewed twice over two weeks reported life satisfaction scores that correlated only moderately with each other (0.59). Such large differences in how a lot of these women reported judging their life as whole imply that current mood and recent events probably affected their judgments considerably.

Fortunately, large representative samples and careful survey creation can avoid most of these problems. Many of these potential biases can be avoided because they are random biases that tend to affect different people at different times. By conducting surveys on large representative samples, the impact of random bias on the usability of the results is considerably reduced. In this way, thorough sampling can eliminate the potential bias associated with personal variations in mood, and localised variation in important events

¹³¹ See the questions for the World Values Survey here:

http://www.worldvaluessurvey.org/wvs/articles/folder_published/survey_2005/files/WVSQuest_RootVers.pdf

(sports teams winning etc.) and the weather. Variability caused by the weather can also be reduced by getting participants to acknowledge the weather before completing the survey (Schwarz & Clore 1983). The effects of recent events and participants' current mood can also be reduced by using a battery of questions about satisfaction with life, instead of just one question (Lucas, Diener, & Suh 1996; Schimmack & Oishi 2005). In order to prevent the significant (but usually small) effect specific questions have on subsequent responses to the more important global questions, researchers usually put global questions first on their surveys (Schimmack & Oishi 2005). Finally, the different results that different wordings of subjective well-being questions produce is a complex issue that is discussed later in this chapter.

2. Johns and Ormerod's Criticisms of Using Measures of Subjective Well-Being to Inform Policymaking

Many criticisms have been levelled at the use of happiness science to inform policy and the most pertinent will be discussed here. In a recent article, Helen Johns and Paul Ormerod make the strong claim that "time series data on happiness tells us nothing" and that it should not be used to inform policymaking (Johns & Ormerod 2008, p. 140).¹³² Their argument is based on three main points: that statistically significant correlations between time series happiness data and other important socioeconomic indicators cannot be found, that the nature of happiness scales makes them insensitive and difficult to compare to most other economic data, and that using time series happiness data for policymaking creates several undesirable problems. These three criticisms are discussed in this section.

¹³² By 'time series data', Johns and Ormerod (2008) mean a collection of answers to the same set of happiness survey questions asked periodically over time.

2.1 Is There a Relationship Between Happiness and Other Variables that We Would Expect to Observe a Relationship Between?

Johns and Ormerod (2008) begin by criticising the argument: 'If there is no correlation between economic growth and happiness, then economic growth must not bring happiness'. Johns and Ormerod (2008) then point to six major social trends that should have affected average happiness through recent history, but (according to them) have not done so. Johns and Ormerod (2008) argue that happiness researchers should admit that either no government's actions since World War Two have ever affected their citizens' happiness or that time series happiness data is completely useless. Their implication is that the first option is untenable, leaving us no choice but to agree that the happiness data must be useless.

Considering that some of the six social trends that Johns and Ormerod (2008) point out are expected to increase happiness, and the others decrease it, it is hardly surprising that relationships between any of these individual trends and average happiness over time are not obvious in correlative analyses that do not control for the other social trends! Most economists who study happiness do not make crude arguments like the one criticised by Johns and Ormerod (2008) above. Rather, unexpected findings are usually posed as questions inviting further investigation (e.g. 'why does it appear that a considerable increase in real income has not made United States citizens any happier over the last 50 years?'). These investigations often use multivariate regression analyses and control for the other factors known to be, or suspected of being, related to happiness in order to isolate the effects of the variables that are being studied. When studies comparing societal trends to happiness are carried out in this manner, like those discussed below, statistically significant relationships are discovered.

Johns and Ormerod (2008) show evidence for a statistically insignificant relationship between income inequality and well-being in the United States over the last 30 years, but fail to mention the results that *do* show a significant relationship between income inequality (using the same measure) and happiness in Europe during the same time period. In Europe, rising income inequality significantly explains a small amount of the variation in reported happiness; income inequality generally made Europeans less happy (Alesina, Di Tella, & MacCulloch 2004). Why the difference between the results from Europe and the United States? Economist Bruno Frey (2008, pp. 57–58) puts it down to United States citizens' greater belief in social mobility; they don't mind the inequality because they (mostly mistakenly) believe that they will be one of the rich folk in the near future.

Johns and Ormerod (2008) believe that the best explanation for the supposed lack of correlation between happiness and these socioeconomic indicators is that, in its current state, the happiness data is simply not worth the paper that it's printed on. The fact of the matter is, however, that statistically significant relationships between reported happiness and many other socio-political factors *have* been discovered in careful studies that properly isolate the variables in question (Dolan, Peasgood, & White 2008; e.g. e.g. Frey 2008). Nevertheless, Johns and Ormerod's two main criticisms of the construction of happiness scales (why they think the data is not worth anything) will now be addressed.

2.2 Are Happiness Scales Insensitive and Too Hard to Compare with Other Economic Measures?

There is some legitimacy to Johns and Ormerod's (2008) claim that time series happiness data is insensitive, but much of their argument is misleading on

this point.¹³³ Johns and Ormerod (2008) give the example of a 3-point scale and then proceed to discuss how insensitive scales with only three options are. First of all, many happiness scales have four or more options, like the example above from the World Values Survey, and many subjective well-being scales have ten or eleven options.¹³⁴

Secondly, Johns and Ormerod (2008, p. 141) assert that to observe a 10% increase from 2.2 in average happiness on a 3-point scale, 22% (net) of a population would have to place themselves in a higher category, an increase that they consider “very difficult” to imagine occurring over “a few years”.¹³⁵ Well, they are not the only ones. An enduring 10% increase in happiness is a lot to ask for in a few years, *regardless of the scale used*. On a 4-point scale 30% of respondents would have to judge themselves as a category happier and on a 10-point scale 90% would have to go up a category or 22.5% would have to go up four categories. Naturally, the gap between categories gets smaller as

¹³³ A statistical measure of a dependant phenomenon can be said to be insensitive if its results do not demonstrate statistically significant changes in response to changes of input variables that we have good reason to believe should effect a statistically significant change in the dependant phenomenon.

¹³⁴ See, for example, the life satisfaction question from the World Values Survey (question V22). See <<http://www.worldvaluessurvey.org/>> for more information and access to the full data set. See also the well-being questions used in the Gallup World Poll (Gallup Inc. 2008, p. 5). Gallup World Poll questions are available from: <http://media.gallup.com/dataviz/www/WP_Questions_WHITE.pdf>.

¹³⁵ It should be noted that another way to interpret the data suggests a smaller increase. It would take 22% of a population to change if a 10% increase in happiness meant multiplying the current average of 2.2 by 1.1 (which equals 2.42). But consider that the average comes from a scale limited to values between 1.0 and 3.0. The accessible part of the scale can be split into 10 equal portions of 0.2. This clearly indicates that a 10% movement on the scale would increase the average from 2.2 to 2.4, which would only require 20% of respondents to opt for a higher happiness category.

the number of them to choose between gets larger, but the proportion of respondents required to report higher happiness increases too. With 3-point scales, researchers simply have to pay more attention to smaller changes in the average value—thank goodness for decimal places!

Despite being hard to imagine, a 10% increase in average reported happiness (on 4-point scales) has actually occurred in some countries over just a few years, including Lithuania (1997–1999), Mexico (1996–2000), and Slovenia (1992–1995). Furthermore, since 1980 at least 21 countries have reported a 10% or more increase in happiness over longer periods of time, including Johns and Ormerod’s home country of Great Britain (1998–2006) (Ingleheart *et al.* 2008). In light of these results, the claim that time-series happiness data is too insensitive to capture trends is totally unfounded.

Having said this, the general consensus in the psychological community is that 3-point scales are not ideal for measuring well-being (Cummins & Gullone 2002). Fortunately, measures of happiness, and especially of subjective well-being more generally, are increasingly using much more precise and robust measures, such as the Subjective Well-Being construct used by Inglehart and colleagues, which combines a 4-point happiness scale and a 10-point life satisfaction scale (Ingleheart *et al.* 2008), and the 4-question Satisfaction With Life Scale of Ed Diener and colleagues (Diener *et al.* 1985).

Johns and Ormerod’s (2008) other criticism of happiness scales is that the type of data they produce is not easily comparable to most of the data that economists use because happiness scales are discrete and bounded. They are correct that data from discrete scales can be difficult to fruitfully compare with non-discrete data. This is because it is harder to find statistically significant results in such comparisons. However, this problem is one faced by data from many non-happiness-related domains, most of which are widely considered to contain useful information when studied carefully.

On the point of time series happiness data being bounded, Johns and Ormerod (2008) admit that short term trends in reported happiness data might exist, but correctly note that no trend in happiness data can continually persist using the present measuring technique. This is because if happiness increased until everyone rated themselves as happy as possible on a bounded scale, then they could not communicate any increase in happiness from that point. Of course this is true, but the chances of everyone reporting maximum happiness on any realistic scale do not seem high enough to warrant discontinuing the collection and analysis of subjective happiness data.

Based on these arguments for the insensitivity of time series happiness data, Johns and Ormerod (2008) then make the misleading assertion that time series happiness data cannot show trends and are thereby useless for comparing to trend-showing traditional economic indicators.¹³⁶ The problem here for Johns and Ormerod is that happiness data do in fact exhibit trends in many countries. Inglehart and colleagues' recent study of reported happiness in 52 countries from 1981–2007 revealed that nearly all of them exhibit upwards trends in happiness (Inglehart *et al.* 2008). They also suggest several reasons for why this trend might have been missed by some researchers. The oldest data on happiness come from the most developed countries, such as the United States, all of which had already passed the point of economic development where gains in happiness could be easily attained through economic growth. Furthermore, increases in tolerance and democratisation (which help increase a sense of freedom and happiness) have been relatively recent and do not always have significant effects on other measures of well-being (Inglehart *et al.* 2008).

¹³⁶ Their reasoning for why this is the case is discussed in the next subsection.

2.3 *Is Time Series Happiness Data Useful for Policymaking?*

In addition to the claim that happiness measures are useless, Johns and Ormerod (2008, p. 142) also assert that happiness data should not be used in policymaking. Johns and Ormerod justify this claim by arguing that governments will inevitably “influence” the data, which is only possible because the data don’t contain any “real information”. Presumably they mean that governments will ‘cook’ the happiness ‘books’, as opposed to create policies that make their citizens happier (thereby influencing them to report higher levels of happiness). Naturally, governments will attempt to present happiness data in the best light (for them at that time), just like corporate directors *and governments* now do with financial data. Without the actual falsification of the data, such unscrupulous behaviour cannot continue for long without being spotted. And, happiness data are just as open to falsification as financial data. However, if every set of happiness data contained as much information as a random set of numbers as Johns and Ormerod (2008) claim, then it would be somewhat more difficult to identify happiness ‘book cooking’ (mainly because no one would care about it).

But is time series happiness data really indistinguishable from a purely random series of numbers like Johns and Ormerod (2008) allege? They claim that time series happiness data provide a flat autocorrelation¹³⁷ and no statistically significant individual values, which make it impossible to create accurate forecasts from it. However, they just tested one set of happiness data from one country. Reported happiness in many countries exhibits clear trends over time, as discussed above, so implying that *all* time series happiness data provide a flat autocorrelation is very misleading. Furthermore, our ability to forecast time series happiness data is constantly increasing due to careful

¹³⁷ Autocorrelations can reveal trends in a data series that are hidden by noise. A flat autocorrelation indicates that there is no trend over time in the data series.

comparisons of changes in reported happiness over time between different countries (or distinct groups within countries). By identifying changes in happiness that occur in some countries or groups, but not others experiencing very similar conditions, variables that might explain the variance in happiness can be isolated. The prevailing conditions in some of these comparisons even allow for the direction of causality to be assessed and thereby begin to provide useful information for predicting the effects on future reported happiness of some upcoming changes in circumstances.

Economist David Dorn and his colleagues recently showed not only that the more democratic countries in their study had higher average reported happiness values, but also that as the countries in the study became more democratic, their average reported happiness values increased too (Dorn *et al.* 2007). By observing how happiness and other variables interact within populations, and comparing that with the interactions observed under similar circumstances in other populations, we can gain valuable insight as to the direction of causality between reported happiness and the other variable.

So, Johns and Ormerod (2008) have not yet provided any good reason for us to believe that happiness data contains no information or that happiness data could not be useful for informing policymaking. However, they go on to cite an unpublished mathematical paper, showing that the variation observed in time series happiness data can be completely explained by sampling error. What the paper actually shows is that some of the variation of happiness over time in *one* set of 3-point time series happiness data from *one* country could be explained by sampling error.¹³⁸ Johns and Ormerod (2008) implicitly generalise this lone result to *all* time series happiness data. Since many other sets of time series happiness data are based on more sensitive scales, show

¹³⁸ The unpublished paper is available on request from Helen Johns: general.hj@googlemail.com (Johns & Ormerod 2008, p. 142).

clear trends, and show more variation from year to year, generalising such a finding to *all* time series happiness data is very misleading.¹³⁹

Knowing that not all time series measures of happiness are as sensitive as they could be and that (as with many social science measures) some of the variation in the resulting data is probably due to sampling error, should time series happiness data ever be used for public policy? At most, combining these findings should result in the conclusion that *some* happiness studies should not be used to guide policy because their results should not inspire enough confidence. However, other time series happiness studies can be useful for policymaking in many ways (Frey 2008, Chap. 13). For example, by carefully comparing results from several populations in circumstances as similar as possible and implementing the policy change in only some of those populations, changes in reported happiness can be recorded and compared. If other variables are sufficiently controlled for and the changes in happiness are statistically significant, then useful information can be gained about how the policy might affect other populations in similar circumstances.

3. Important Problems for Using Measures of Subjective Well-Being to Inform Policymaking

Even though Johns and Omerod's (2008) criticisms of measuring happiness are relatively easily dealt with, there are other more serious critiques that need to be addressed. First I turn to the problem of whether subjective well-being surveys really measure happiness and then to the problem of which version of happiness we should be measuring.

¹³⁹ I provide further criticism of Johns and Ormerod's (2008) interpretation of Johns' unpublished study in Turton (2009).

3.1 Do Measures of Happiness Really Measure Happiness?

As discussed, there are currently a number of different methods that are claimed to be measures of well-being. These measures include brain scans, daily reports about how participants have been feeling, the opinions of participants' friends or colleagues, the opinion of an expert, the amount participants smile and, most commonly, survey questions. The survey questions are often markedly different. They might ask about happiness, satisfaction with life, the degree to which one would choose to change one's life and many other distinct notions. It is the findings produced by analyses of these surveys that are usually recommended as aids to inform policymaking. But should we trust that these surveys are actually measuring happiness or well-being and not something else entirely? And should we trust that one person's happiness is the same as another's?

Kroll (2010), Layard (2003; 2005) and Bok (2010) all argue that we should have faith in answers to survey questions about well-being because they are significantly correlated with many other measures of well-being, both within and between large groups of people. Importantly, these researchers view the correlations found in (the unarguably objective) neuroimaging studies as adding considerable support to the idea that subjective measures of well-being are assessing something that is real and that we all experience. Layard (2005, p. 17) states:

Sceptics may still question whether happiness is really an objective feeling that can be properly compared between people. To reassure doubters, we can turn to modern brain physiology with its sensational new insights into what is happening when a person feels happy or unhappy.

Indeed, it is true that most of the diverse measures of well-being mentioned above are significantly correlated for most test participants. However, that truth does not indicate that any of the different measures are actually evaluating well-being or happiness. The important point that Kroll, Layard, and Bok ignore is that although most of the correlations in the neuroimaging studies are statistically significant, they are not particularly large. Furthermore, the neuroimaging studies that Layard (2005, pp. 17–19) and others cite usually compare neuroimaging results with various cues that are expected to create various feelings, including ‘approach’ or ‘withdrawal’, in their participants, not with measures of subjective well-being.

In the only comprehensive study of correlations between neuroimaging and measurement of subjective well-being, several prominent subjective well-being measures were compared with electroencephalogram data from 84 right-handed adults aged 57–60. Correlating highest with the neuroimaging results were the results for a measure of psychological flourishing—Carol Ryff’s (1989) Scales of Psychological Well-Being. The correlation was highly significant ($p < 0.01$) and moderate in size (0.33) (Urry *et al.* 2004, p. 370). Following close behind was Ed Diener and colleagues’ (1985) Satisfaction With Life Scale, which correlated with the neuroimaging results by 0.30 and was also highly significant ($p < 0.01$) (Urry *et al.* 2004, p. 370). A smaller (0.21) and slightly less significant ($p < 0.05$) correlation was also found with the positive affect component of Watson, Clark, and Tellegen’s (1988) Positive and Negative Affect Schedule (Urry *et al.* 2004, p. 370). It should be noted that, unlike the questions that directly ask about happiness, none of the measures of subjective well-being used in Urry and colleagues’ (2004) study are equivalent to the folk notion of happiness (although the measure of positive affect would come the closest).

The significant, but relatively small, correlations between the neuroimaging results and these measures of subjective well-being tell us two things. First, the neuroimaging results and these measures of subjective well-being are very likely to be tracking phenomena that are related in some positive way. This is shown by the fact that the correlation is positive and highly statistically significant. Second, the phenomena being tracked are clearly distinct. The high statistical significance of the results should make us confident that the various measures are *not* measuring exactly the same thing. If the size of the correlations were much higher, at least above 0.60,¹⁴⁰ and the statistical significance remained very high, then we would expect to observe the phenomena measured by the different tests to covary more closely and, thereby, give the impression of being the same thing. Positive correlations of 0.33 (roughly) mean that we should expect an increase in the results of the neuroimaging measure to be usually accompanied by a relatively smaller increase of the subjective well-being measure. This is the kind of relationship we expect from distinct but positively related variables, not from two different measures of the same phenomenon.

Perhaps most telling of all on the question of what support neuroimaging provides for the objectivity of happiness is that the cognitive scientists who carry out neuroimaging studies rarely claim to be testing happiness or well-being. Much more commonly they claim to be investigating the neural correlates of pleasure and pain or approach and withdrawal behaviour, as is the case with the neuroimaging study that Layard discusses the most (Davidson *et al.* 1990; c.f. Layard 2005). But even if experimental neuroimaging studies were carried out until a measure of neurological activity correlated very highly and statistically significantly with a subjective

¹⁴⁰ 0.60 is the level above which behavioural scientists usually deem results to be 'highly related' (Cohen 1988).

measure of happiness, it would be presumptuous to declare it the discovery of an objective measure of happiness. Rather, such a neurological measure should be understood as an objective measure of the propensity to report subjective happiness. Understood this way, it's much less obvious how objective neuroimaging results are supposed to give us confidence that a measure of subjective well-being is accurately performing the task that we want it to.¹⁴¹

Nevertheless, the confluence of the various correlations between measures of subjective well-being, aspects of bio-physical health, neuroimaging data, observers' reports, and behavioural analyses points toward there being something tangible to measure (Layard 2010). Or, at least, a few closely related tangible things to measure.¹⁴² Frey and Stutzer (2002) also provide a (now dated but nonetheless excellent) summary of how measures of subjective well-being are reliable *enough* to provide useful economic and policy insights despite all of the potential problems discussed so far. Perhaps most encouraging in this regard is the study of Oswald and Wu (2010, p. 579), which reported a highly significant (p approximately = 0.0001) and relatively large (0.6) correlation between objective quality of life factors and subjective satisfaction with life in the United States of America. This careful study of over 1.3 million data points "suggests that subjective well-being data contain genuine information about the quality of people's lives", which is exactly what policymakers should be interested in (Oswald & Wu 2010, p. 579). A further problem remains, however. None of the many existing measures of

¹⁴¹ A similar presentation of this issue appears in Weijers & Jarden (2011, pp. 56–57) and Feldman 2010, chap. 13).

¹⁴² Indeed, this is surely how these results are best understood. The fact that different measures of well-being actually measure different *aspects* of well-being is not necessarily a drawback, as discussed below.

subjective well-being measure the same thing, so it is far from obvious which one or ones, if any, actually measure well-being.

3.2 How Do We Know if We Are Measuring the Right Kind of Happiness?

The fact that different questions in well-being surveys are not measuring the same phenomena is no revelation to many researchers. It's an understatement to say that philosophers have been discussing the merits of various conceptions of well-being for a long time. To a philosopher, the different kinds of questions used in subjective well-being surveys often endorse one particular philosophical conception of well-being. It might be suggested that these philosophical differences are trifling distractions, since all of the measures are assessing something that is obviously good, but there is a serious problem with this suggestion.

What should policymakers do when findings based on different measures of well-being imply different policies? This problem might be easily pushed aside if it were not for the fact that many subjective well-being surveys produce results that appear to contradict the existing findings of the science of well-being. For example, according to Kroll (2010) it is well established that increases in income have no effect on the overall self-reported well-being of countries with per capita GDP over 10,000 Euros. However, well-being researchers using different questions have found a positive significant relationship between income and self-reported well-being amongst both poor and rich countries (e.g. Deaton 2010). Since both of these findings are statistically significant, the 'apparently' very similar questions are the most likely cause of the divergent results. In this particular case, the findings that increasing income improves well-being well above 10,000 Euros are based on data from Gallup World Polls. The question about well-being in Gallup World Polls is worded in a way that is more likely to elicit comparisons with

all other people in the world, not just the respondents' immediate reference group. Many studies have shown that our reported satisfaction with life is significantly affected by whatever reference group is most salient to us at the time of the survey (Graham & Pettinato 2002; Kahneman & Krueger 2006; e.g. Merton 1957). Therefore, it should come as no surprise that Gallup Word Polls' wording of their subjective well-being question makes a difference in this way.

Indeed, other researchers have good evidence that the more a measure of subjective well-being asks about the respondents' emotional lives, and the less it encourages them to engage in cognitive deliberation about how satisfied they are with their life compared to what it might have been, then the smaller and less significant the relationship between increases in income and increases in subjective well-being becomes (Diener *et al.* 2010). When the science of well-being produces contradictory findings in this way, it creates a problem for policymakers. To prevent this from occurring, researchers analysing well-being survey data should never generalise findings from different questions unless those questions really are asking about the same phenomena. But what should policymakers intent on using the science of well-being do when the findings are contradictory?

Policymakers should always investigate the original surveys to find out if the well-being findings are about one conception of well-being. Such an investigation might lead to the identification of different questions about well-being as the source of conflict. Discovery of such conflicts illuminates the fundamental problems policymakers intent on using the science of well-being face. Which question about subjective well-being is the most appropriate basis for policymaking? What exactly is well-being anyway?

Despite thousands of years spent pondering these questions, philosophers have not yet come up with definitive answers. They have, however, identified

several main conceptions of well-being, along with their advantages and their disadvantages. This knowledge should be shared and discussed widely, with the public and with social scientists working on happiness and well-being.¹⁴³ With a greater awareness of these problems and a deeper understanding of what well-being might consist of, citizens can exercise their democratic rights and lobby their governments to adopt their preferred conception. When this occurs, governments can encourage the use of appropriate measures of well-being. Only then can well-being researchers be confident that they are producing findings that are really relevant for policymaking. And only then can policymakers get the most out of the science of well-being.

4. What Role Should Happiness Play in Policymaking?

In any democratic society the (hopefully informed) citizens should decide what conceptions of happiness or well-being are important and the extent of the role any such conceptions should play in policymaking. But in order to educate citizens and encourage effective evidence-based policymaking, academics and top-level civil servants need to better clarify the various conceptions of happiness and well-being and whether we can accurately and efficiently measure them. After these issues have been clarified, how circumstances and policies affect happiness or well-being (as defined in each case) should also be investigated to help better understand where each concept of well-being fits in the economic landscape, and also to promote public debate on the relevant merits of certain kinds of happiness and other goods. Philosophers, psychologists, economists, statisticians and

¹⁴³ Dan Haybron and Valerie Tiberius provide some excellent examples of how philosophers can engage with social scientists and policymakers to inform them about the philosophical theories and debates about well-being (Haybron 2008c; Haybron & Tiberius 2012; Tiberius 2004; Tiberius & Hall 2010).

policymakers should work together on this in order to pool their collective expertise and progress most effectively. But what can governments and policymakers do to speed up this process?

In a recent interview for the *International Journal of Wellbeing*, Nobel Laureate Daniel Kahneman has expressed fears that governments will insert only one of the widely-used subjective well-being questions in their censuses (e.g. How satisfied are you with your life as a whole? On a 1-5 Likert scale) (Jarden 2011). He worries that this kind of measure won't be sensitive enough to reveal anything interesting, even in the face of dramatically changing traditional economic indicators. Kahneman's main concern is that both politicians and the public will view the insensitivity of such measures as a reason to reject all future use of subjective measures of well-being. So, how can governments ensure that they ask the right question from the start?

For starters, one question will simply not be enough. As Martin Seligman discusses in *Flourish* (2011) there is considerable evidence in favour of a 'dashboard' approach because it seems that there are several distinct and roughly equally important dimensions of well-being. He outlines the following five aspects of well-being as being worthy of inclusion in a dashboard of subjective well-being indicators: positive emotion, engagement, meaning, positive relationships, and accomplishment (Forgeard *et al.* 2001). The United Kingdom's Office for National Statistics (ONS) appears to have followed the dashboard approach to some extent, asking four subjective well-being questions in its ONS Opinion Survey, each of which represents a different group of philosophical theories about well-being (Beaumont 2011). The ONS has not yet confirmed which questions will capture data on subjective well-being in the long-run, with consultation still ongoing. It seems very likely, though, that several subjective well-being questions will be chosen. This cautious approach is surely a good one. The truth of the matter is

that more research needs to be done to finalise both the list of distinct aspects of well-being and the most accurate and effective way to measure each of them with as few questions as possible. So, the question now becomes: how can governments ensure that they ask the right *questions* from the start?

I propose a ten-year international collaborative effort to answer this question. The importance of this issue means that ten years should not be considered excessive. Governments in several countries in Europe and around the world already conduct longitudinal panel surveys and some of them even use these surveys to collect data on both economic and subjective well-being indicators. The best of these kinds of surveys for our purposes are those in which each respondent, and everyone else in their household, has to complete the survey every three months or so for at least three years. The number and types of questions would have to be expanded, however. The generic objective economic and demographic questions would need to be accompanied by subjective versions of the same questions, subjective questions about the events that have recently impacted their lives, and batteries of subjective well-being measures for each potentially important aspect of well-being.¹⁴⁴

Comparing the various measures of subjective well-being to the existing philosophical theories of well-being would be a useful way to assess whether the existing measures cover the scope of all possibly important conceptions of well-being. Such an assessment would have to be carried out carefully because of the profound differences between philosophical accounts of well-being that are nonetheless grouped together. As discussed in Chapter 1, there is a variety of types of Prudential Hedonism. While a simple “How happy

¹⁴⁴ Ed Diener’s (2006) ‘Guidelines for National indicators of Subjective Well-Being and Ill-being’ is a good starting point for more specific guidance on the creation of such a collection of measures of subjective well-being.

have you been these days?" question might be thought to cover all hedonistic theories, other measures better capture the ideas behind specific hedonistic theories of well-being. For example, Kahneman and Krueger's (2006) U-index (which asks about the proportion of time respondents spend in a negative emotional state) would be a fairly good measure of Epicurus' pain-minimising Prudential Hedonism, while Watson and colleagues' (1988) Positive and Negative Affect Scale would better represent Bentham's net-pleasant-feelings-based Prudential Hedonism.

If several governments around the world conducted this kind of survey (with the same questions) very useful information would result. Not only would we be able to assess how perceptions mediate the effect of objective economic changes on individuals and groups, we would also be able to see what kinds of objective economic and demographic factors affect the various aspects of well-being and whether these results are cross-culturally robust. The combination of surveying the same participants, and those they live with, over time provides more definitive information on how what happens to people and those around them affects their well-being.

Furthermore, the measures of the components of well-being could be refined for accuracy and sensitivity during this process. The questions could be updated and improved with each new round of the surveys, until the ten years is up. At the end, individual governments could decide which of the measures of particular aspects of well-being are most relevant to their respective countries. They would know if the measures were already robust enough to be used as a question in the census, or if the questions needed further refinement, and could then act accordingly.

Indeed, the OECD's current work on creating guidelines for measuring subjective well-being combined with the political impetus created at the UN

meeting on Happiness and Wellbeing: Defining a New Economic Paradigm in April this year might result in such a widespread collaborative effort.

If this kind of approach were adopted, and well-being dashboards were incorporated into censuses, then political parties of the not-too-distant future might even be able to differentiate themselves by giving precedence to the promotion of certain aspects of well-being over others. To better facilitate public debate on the merits of these various measures of well-being, philosophers, psychologists and economists should hold interdisciplinary public lectures and forums on different conceptions of well-being. Assuming that the public gain sufficient knowledge about the different aspects and conceptions of well-being being measured, the constant collecting of data on these well-being indicators could be a great way to measure a government's effectiveness (comparing the net benefit to well-being indicators with the net cost to capital stocks, such as natural resources, infrastructure, etc.) (Weijers 2012). In this way, measures of subjective well-being could be used to allow policymakers to consider the potential impacts of a policy on the happiness as well as the wealth of citizens. It will take a considerable amount of interdisciplinary work to reach this point but, once there, policymakers will have an incredibly useful set of tools at their disposal and citizens might just become happier because of it.

References

- Aguirre, A. (2006). Where Did It All Come From?, *Sky & Telescope*, November: 36–41.
- Alesina, A., Di Tella, R., & MacCulloch, R. (2004). Inequality and Happiness: Are Europeans and Americans Different?, *Journal of Public Economics*, 88: 2009–2042.
- Angner, E. (2011). Are Subjective Measures of Well-Being ‘Direct’?, *Australasian Journal of Philosophy*, 89(1): 115–130.
- Attfield, R. (1987). *A Theory of Value and Obligation*. London: Croom Helm.
- Australian Treasury (2006). *Wellbeing and Public Policy*. Canberra: Australian Treasury. Available from:
[http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/d62854438d178820ca25727a0021d0ea/\\$FILE/Wellbeing%20and%20Public%20Policy%20-%20June%202006.pdf](http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/d62854438d178820ca25727a0021d0ea/$FILE/Wellbeing%20and%20Public%20Policy%20-%20June%202006.pdf)
- Baber, H. (2008). The Experience Machine Deconstructed, *Philosophy in the Contemporary World*, 15(1): 132–137.
- Baer, P., Athanasiou, T., & Kartha, S. (2007). *The Right to Development in a Climate Constrained World: The Greenhouse Development Rights Framework*. Berlin: Heinrich Böll Foundation.
- Baggini, J. & Fosl, P. S. (2007). *The Ethics Toolkit: A Compendium of Ethical Concepts and Methods*. Wiley-Blackwell.
- Barber, A. (2011). Hedonism and the Experience Machine, *Philosophical Papers*, 40(2): 257–278.
- Beaumont, J. (2011). *Measuring National Well-Being: Discussion Paper on Domains and Measures*, Office for National Statistics, 31 October 2011. Available from: http://www.ons.gov.uk/ons/dcp171766_240726.pdf

- Becker, L. C. (1992). Good Lives: Prolegomena, in E. F. Paul, F. D. Miller, Jr., & J. Paul (eds.), *The Good Life and the Human Good*. Cambridge: Cambridge University Press: 15–37.
- Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*. Reprinted 1948, Oxford, UK: Blackwell. Reprinted 2005, Adamant Media Corporation.
- Bentham, J. (1825). *The Rationale of Reward*. London: John and H. L. Hunt.
- Berridge, K. & Kringelbach, M. (2011). Building a Neuroscience of Pleasure and Well-Being, *Psychology of Well-Being: Theory, Research and Practice*, 1(1): 1–26.
- Bettini, S. (2005). A Cosmic Archipelago: Multiverse Scenarios in the History of Modern Cosmology, *arXiv:physics/0510111*.
- Blackburn, S. (2005). Hedonism, in S. Blackburn (ed.), *Oxford Dictionary of Philosophy*. Oxford University Press: 161.
- Blake, R. M. (1926). Why Not Hedonism?: A Protest, *International Journal of Ethics*, 37(1): 1–18.
- Bloomfield A. N., Sager, J. A., Bartels, D. M., & Medin, D. L. (2006). Caring about Framing Effects, *Mind & Society*, 5(2): 123–138.
- Bok, D. (2010). *The Politics of Happiness: What Governments can Learn from the New Research on Well-Being*. Princeton, NJ: Princeton University Press.
- Bok, S. (2010). *Exploring Happiness: From Aristotle to Brain Science*. New Haven: Yale University Press.
- Bostrom, N. & Ord, T. (2006). The Reversal Test: Eliminating Status Quo Bias in Applied Ethics, *Ethics*, 116: 656–679.
- Boylan, M. (2008). *The Good, the True and the Beautiful: A Quest for Meaning*. New York: Continuum International.
- Bradbury, R. (1957). *Dandelion Wine*. Doubleday.

- Bramble, B. (forthcoming). The Distinctive Feeling Theory of Pleasure, *Philosophical Studies*, Online First 8 June 2011. Available from: <http://www.springerlink.com/content/65qq253t7262h514/fulltext.pdf>
- Brandt, R. (1966). The Concept of Welfare, in S. R. Krupp (ed.), *The Structure of Economic Science*. Prentice-Hall: 257–276.
- Brink, D. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.
- Brink, D. (2006). Some Forms and Limits of Consequentialism, in *The Oxford Handbook of Ethical Theory*, D. Copp (ed.). Oxford: Clarendon Press: 380–423.
- Broad, C. D. (1959). *Five Types of Ethical Theory*. Paterson, NJ: Littfield Adams.
- Bronsteen, J., Buccafusco, C., & Masur, J. (2010). Welfare as Happiness, *Georgetown Law Journal*, 98: 1583–1641.
- Brülde, B. (2007). Happiness Theories of the Good Life, *Journal of Happiness Studies*, 8: 15–49.
- Bruner, J. (1960). *The Process of Education*. Cambridge, MA: Harvard University Press.
- Cameron, D. (2010). *A Transcript of a Speech Given by the Prime Minister on Wellbeing on 25 November 2010*. Available from: <http://www.number10.gov.uk/news/pm-speech-on-well-being/>
- Caney, S. (2005). Cosmopolitan Justice, Responsibility, and Global Climate Change, *Leiden Journal of International Law*, 18: 747–775.
- Casal, P. (2007). Why Sufficiency is Not Enough, *Ethics*, 117(2): 296–326.
- Chen, M. K., Lakshminaryanan, V., & Santos, L. R. (2006). The Evolution of Our Preferences: Evidence from Capuchin Monkey Trading Behavior, *Journal of Political Economy*, 114: 517–537.

- Churchland, P. (1996). Towards a Cognitive Neurobiology of the Moral Virtues, *Topoi*, 17: 83–96.
- Claussen, E., & McNeilly, L. (1998). *Equity and Global Climate Change: The Complex Elements of Global Fairness*. Arlington, VA: PEW Center for Global Climate Change.
- Cohen, J. (1988). *Statistical Power Analysis for Behavioural Sciences* (2nd ed.). Hillsdale, NJ: Earlbaum.
- Craig, A. D. (2004). Human Feelings: Why are Some More Aware than Others?, *Trends in Cognitive Sciences*, 8: 239–241.
- Crandall, C. S., Eidelman, S., Skitka, L. J., & Morgan, S. G. (2009). Status Quo Framing Increases Support for Torture, *Social Influence*, 4: 1–10.
- Crisp, R. (2003). Equality, Priority, and Compassion, *Ethics*, 113(4): 745–763.
- Crisp, R. (2006a). Hedonism Reconsidered, *Philosophy and Phenomenological Research*, LXXIII(3): 619–645.
- Crisp, R. (2006b). *Reasons and the Good*. Oxford: Oxford University Press.
- Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., & Dolan, R. J. (2004). Neural Systems Supporting Interoceptive Awareness, *Nature Reviews Neuroscience*, 7: 189–195.
- Cummins, R. & Gullone, E. (eds.) (2002). *The Universality of Subjective Wellbeing Indicators: A Multi-Disciplinary and Multi-National Perspective*. Kluwer Academic Publishers.
- Darwall, S. (1983). *Impartial Reason*. Ithaca: Cornell University Press.
- Darwall, S. (1997). Self Interest and Self Concern, *Social Philosophy and Policy*, 14: 158–178.
- Davidson, R., Ekman, P., Saron, C., Senulius, S., & Friesen, W. (1990). Emotional Expression and Brain Physiology I: Approach/Withdrawal and

- Cerebral Asymmetry, *Journal of Personality and Social Psychology*, 58: 330–341.
- Davis, W. (1981). Pleasure and Happiness, *Philosophical Studies*, 39(3): 305–317.
- De Brigard, F. (2010). If You Like it, Does it Matter if it's Real?, *Philosophical Psychology*, 23(1): 43–57.
- Deaton, A. (2010). Income, Aging, Health and Well-Being Around the World: Evidence from the Gallup World Poll, NBER Chapters, *Research Findings in the Economics of Aging*. National Bureau of Economic Research: 235–263.
- Dennett, D. (1980): The Milk of Human Intentionality, *Behavioral and Brain Sciences*, 3(3): 428–430.
- Diener, E. (2006). Guidelines for National Indicators of Subjective Well-Being and Ill-Being, *Applied Research in Quality of Life*, 1: 151–157.
- Diener, E. (2009). *Culture and Wellbeing: The Collective Works of Ed Diener*. Springer.
- Diener, E. (2011). Happiness, Economist Debates, *The Economist*, 18 May 2011. Available from: <http://www.economist.com/debate/days/view/702>
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale, *Journal of Personality Assessment*, 49: 71–75.
- Diener, E. Kahneman, D., Tov, W., & Arora, R. (2010). Income's Differential Impact on Judgments of Life Versus Affective Well-Being, in E. Diener, D. Kahneman, & J. F. Helliwell (eds.), *International Differences in Well-Being*. New York: Oxford University Press: 3–15.
- Diener, E. & Scollon, C. (2003). *Subjective Well-Being is Desirable, But Not the Summum Bonum*. Unpublished Manuscript. University of Minnesota Workshop on Well-Being. Available from: http://scholar.googleusercontent.com/scholar?q=cache:2gjgl65fdpQJ:scholar.google.com/&hl=en&as_sdt=0,5

- Ding, D. K., Charoenwong, C., & Seetoh, R. (2004). Prospect Theory, Analyst Forecasts, and Stock Returns, *Journal of Multinational Financial Management*, 14: 425–442.
- Dolan, P., Peasgood, T., & White, M. (2008). Do We Really Know What Makes Us Happy?: A Review of the Economic Literature on the Factors Associated with Subjective Well-Being, *Journal of Economic Psychology*, 29: 94–122.
- Donner, W. (2006). Mill's Theory of Value, in H. R. West (ed.), *The Blackwell Guide to Mill's Utilitarianism*. Oxford: Blackwell.
- Dorn, D., Fischer, J., Kirchgässner, G., & Sousa-Poza, A. (2007). Is it Culture or Democracy?: The Impact of Democracy, Income and Culture on Happiness, *Social Indicators Research*, 82(3): 505–526.
- Dorsey, D. (2010). Preferences, Welfare, and the Status-Quo Bias, *Australasian Journal of Philosophy*, 88(3): 535–554.
- Druckman, J. N. (2001). Evaluating Framing Effects, *Journal of Economic Psychology*, 22: 91–101.
- Duncker, K. (1941). On Pleasure, Emotion, and Striving, *Philosophy and Phenomenological Research*, 1: 391–430.
- Dworkin, R. (2000). *Sovereign Virtue*. Cambridge: Harvard University Press.
- Easton, M. (2006). *Britain's Happiness in Decline*, BBC News, 2 May 2006.
Available from:
http://news.bbc.co.uk/2/hi/programmes/happiness_formula/4771908.stm
- England, R. W. (1998). Alternatives to Gross Domestic Product: A Critical Survey, in F. Ackerman, D. Kiron, N. Goodwin, J. Harris, & K. P. Gallagher (eds.), *Human Wellbeing and Economic Goals* (vol. 3). Island Press: 373–402.

- Farhi, E., Guth, A. (1987). An Obstacle to Creating a Universe in the Laboratory, *Physics Letters B*, 183(2): 149–155.
- Farhi, E., Guth, A. & Guven, J. (1990). Is it Possible to Create a Universe in the Laboratory by Quantum Tunneling?, *Nuclear Physics B*, 339: 417–490.
- Feldman, F. (1997). On the Intrinsic Value of Pleasures, *Ethics*, 107(3): 448–466.
- Feldman, F. (1997). *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy*. Cambridge: Cambridge University Press.
- Feldman, F. (2002). The Good Life: A Defense of Attitudinal Hedonism, *Philosophy and Phenomenological Research*, LXV(3): 604–628.
- Feldman, F. (2004). *Pleasure and the Good Life*. Oxford: Clarendon Press.
- Feldman, F. (2011). What We Learn from the Experience Machine, in R. M. Bader & J. Meadowcroft (eds.), *The Cambridge Companion to Nozick's Anarchy, State, and Utopia*. Cambridge University Press.
- Finnis, J. (1980). *Natural Law and Natural Rights*. Oxford: Clarendon Press.
- Finnis, J. (1983). *Fundamentals of Ethics*. Washington, DC: Georgetown University Press.
- Fischler, W., Morgan, D., & Polchinski, J. (1990a). Quantum Nucleation of False Vacuum Bubbles, *Physics Review D*, 41: 2638.
- Fischler, W., Morgan, D., & Polchinski, J. (1990b). Quantization of False-Vacuum Bubbles: A Hamiltonian Treatment of Gravitational Tunneling, *Physics Review D*, 42: 4042.
- Fletcher, G. (2007). Wrongness, Welfarism and Evolution: Crisp on Reasons and the Good, *Ratio* (new series), 20(3): 341–347.
- Flew, A. (1963). Tolstoi and the Meaning of Life, *Ethics*, 73(2): 110–118.
- Forgeard, M. J. C., Jayawickreme, E., Kern, M., & Seligman, M. E. P. (2011). Doing the Right Thing: Measuring Wellbeing for Public Policy, *International Journal of Wellbeing*, 1(1): 79–106. doi:10.5502/ijw.v1i1.15

- Frankena, W. (1973). *Ethics* (2nd ed.). New Jersey: Prentice-Hall.
- Frankfurt, H. (1987). Equality as a Moral Ideal, *Ethics*, 98(1): 21–43.
- Frey, B. S. (2008). *Happiness: A Revolution in Economics*. MIT Press.
- Frey, B. S. & Stutzer, A. (2002). What Can Economists Learn from Happiness Research?, *Journal of Economic Literature*, 40(2): 402–435.
- Frey, B. S. & Stutzer, A. (2007). Should National Happiness be Maximized?, Institute for Empirical Research in Economics, *University of Zürich, Working Paper Series*, Working Paper No. 306, March 2007. Available from: <http://www.iew.uzh.ch/wp/iewwp306.pdf>
- Gallup Inc. (2008). *World Poll Questions*, Washington, DC: Gallup Inc.
Available from:
http://media.gallup.com/dataviz/www/WP_Questions_WHITE.pdf
- Gardiner, S. (2004). Ethics and Global Climate Change, *Ethics*, 114: 555–600.
- Gilbert, D. (2006). *Stumbling on Happiness*. New York: Alfred A. Knopf.
- Gilovich, T., Griffin, D. W., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.
- Graham, C. (2009). *Happiness Around the World: The Paradox of Happy Peasants and Miserable Millionaires*. Oxford University Press.
- Graham, C. & Pettinato, S. (2002). Frustrated Achievers: Winners, Losers and Subjective Well-Being in New Market Economies, *Journal of Development Studies*, XXXVIII: 100–140.
- Griffin, J. (1986). *Well-Being: Its Meaning, Measurement and Moral Importance*. Oxford: Clarendon Press.
- Guendelman, E. & Portnoy, J. (1999). The Universe out of an Elementary Particle?, *Classical Quantum Gravity*, 16: 3315.
- Guendelman, E. & Portnoy, J. (2001). Almost Classical Creation of a Universe, *Modern Physics Letters A*, 16: 1079.

- Chattopadhyaya, D. (ed.) (1990). *Cārvāka/Lokāyata: An Anthology of Source Materials and Some Recent Studies*. New Delhi: Indian Council of Philosophical Research in association with Rddhi-India Calcutta, 1990.
- Hartman, R. S., Doane, M. J., & Woo, C.-H. (1991). Consumer Rationality and the Status Quo, *The Quarterly Journal of Economics*, 106(1): 141–162.
- Haslett, D. W. (1990). What is Utility?, *Economics and Philosophy*, 6(1): 65–94.
- Hausman, D. M. (2010). Hedonism and Welfare Economics, *Economics and Philosophy*, 26: 321–344.
- Haybron, D. M. (2001). Happiness and Pleasure, *Philosophy and Phenomenological Research*, 62(3): 501–528.
- Haybron, D. M. (2007). Life Satisfaction, Ethical Reflection, and the Science of Happiness, *Journal of Happiness Studies*, 8: 99–138.
- Haybron, D. M. (2008a). Happiness, the Self and Human Flourishing, *Utilitas*, 20(1): 21–49.
- Haybron, D. M. (2008b). *The Pursuit of Unhappiness*. Oxford: Oxford University Press.
- Haybron, D. M. (2008c). Philosophy and the Science of Subjective Well-Being, in M. Eid & R. Larsen (eds.), *The Science of Subjective Well-Being*. New York: Guilford: 17–43.
- Haybron, D. M. & Tiberius, V. (2012). Normative Foundations for Well-Being Policy, *Papers on Economics and Evolution*, #1202. Available from: <ftp://papers.econ.mpg.de/evo/discussionpapers/2012-02.pdf>
- Hewitt, S. (2009). What Do Our Intuitions about the Experience Machine Really Tell Us about Hedonism?, *Philosophical Studies*, 151(3): 331–349.
- Helliwell, J. F. & Wang, S. (forthcoming). The State of World Happiness, in J. F. Helliwell (ed.), *World Happiness Report*. Commissioned for the United Nations.

- Hirsch, M. S. & Liebert, R. M. (1998). The Physical and Psychological Experience of Pain: The Effects of Labeling and Cold Pressor Temperature on Three Pain Measures in College Women, *Pain*, 77(1): 41–48.
- Hodgson, G. M. (1988). *Economics and institutions*. Polity Press, Cambridge.
- Hofstadter, D. & Dennett, D. (1981). *The Mind's I: Fantasies and Reflections on Self and Soul*. New York: Basic Books.
- Hooker, B. (2000). *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Clarendon.
- Hume, D. (1777). *An Enquiry Concerning the Principles of Morals*. Reprinted in T. L. Beauchamp (ed.), Clarendon Press, 1998.
- Hurka, T. (1993). *Perfectionism*. New York: Oxford University Press.
- Hurka, T. (2011). *The Best Things in Life: A Guide to What Really Matters*. New York: Oxford University Press.
- IMF (International Monetary Fund) (2009). *World Economic Outlook, October 2009: Sustaining the Recovery*. Washington, DC: International Monetary Fund. Available from:
www.imf.org/external/pubs/cat/longres.cfm?sk=22576.0
- Ingleheart, R., Foa, R., Peterson, C., & Welzel, C. (2008). Development, Freedom, and Rising Happiness: A Global Perspective (1981–2007), *Perspectives on Psychological Science*, 3(4): 264–285.
- Inwood, B. & Gerson, L. P. (eds.) (1994). *The Epicurus Reader: Selected Writings and Testimonia*. Indianapolis, Hackett Publishing.
- IPCC (Intergovernmental Panel on Climate Change) (2007). *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press.
- James, W. (1950). *The Principles of Psychology*. New York: Dover.

- Jarden, A. (2011). An Interview with Daniel Kahneman, *International Journal of Wellbeing*, 1(1): 186–188. doi:10.5502/ijw.v1i1.22
- Jarden, A., Jose, P., Kashdan, T., Simpson, O., McLachlan, K., & Mackenzie, A. (no date). International Wellbeing Study: Further Assessment Study. Unpublished raw data.
- Johns, H. & Ormerod, P. (2008). The Unhappy Thing about Happiness Studies, *Real-World Economics Review*, 46, 139–146.
- Jollimore, T. (2004). Meaningless Happiness and Meaningful Suffering, *The Southern Journal of Philosophy*, 42(3): 333–347.
- Kagan, S. (1994). Me and My Life, *Proceedings of the Aristotelian Society* (new series), 94: 309–324.
- Kagan, S. (1998). *Normative Ethics*. Boulder: Westview Press.
- Kagan, S. (2009). Well-Being as Enjoying the Good, *Philosophical Perspectives: Ethics*, 23: 253–272.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias, *The Journal of Economic Perspectives*, 5(1): 193–206.
- Kahneman, D. & Krueger, A. B. (2006). Developments in the Measurement of Subjective Well-Being, *The Journal of Economic Perspectives*, 20(1): 3–24.
- Kahneman, D. & Tversky, A. (2000). *Choices, Values, and Frames*. Cambridge: Cambridge University Press.
- Kaku, M. (2004). *Parallel Worlds: A Journey Through Creation, Higher Dimensions, and the Future of the Cosmos*. Doubleday: New York.
- Katz, L. D. (2009). Pleasure, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition). Available from: <http://plato.stanford.edu/archives/fall2009/entries/pleasure/>

- Kawall, J. (1999). The Experience Machine and Mental State Theories of Well-Being, *The Journal of Value Inquiry*, 33: 381–387.
- Keller, S. (2009). Welfare as Success, *Noûs*, 43(4): 656–683.
- Kempf, A. & Ruenzi, S. (2006). Status Quo Bias and the Number of Alternatives: An Empirical Illustration from the Mutual Fund Industry, *Journal of Behavioral Finance*, 7: 204–213.
- Kennedy, R. F. (1968). *Speech at University of Kansas*, March 18. Transcript available from:
http://www.glaserprogress.org/program_areas/pdf/Remarks_of_Robert_F_Kennedy.pdf
- Knetsch, J. L. (1989). The Endowment Effect and Evidence of Nonreversible Indifference Curves, *American Economic Review*, 79: 1277–1284.
- Knetsch, J. L. & Sinden, J. A. (1984). Willingness to Pay and Compensation Demand: Experimental Evidence of an Unexpected Disparity in Measures of Value, *Quarterly Journal of Economics*, 99: 507–521.
- Knobe, J. (2007). Experimental Philosophy, *Philosophy Compass*, 2(1): 81–92.
- Kolber, A. (1994). Mental Statism and the Experience Machine, *Bard Journal of Social Sciences*, 3(3–4): 10–17.
- Kraut, R. (2007). *What is Good and Why: The Ethics of Well-Being*. Cambridge, MA: Harvard University Press.
- Kringelbach, M. L. & Berridge, K. B. (eds.) (2010). *Pleasures of the Brain*. Oxford University Press.
- Kroll, C. (2010). Social Democracy and Happiness, *Social Europe Journal*. Available from: <http://www.social-europe.eu/2010/01/social-democracy-and-happiness/>
- Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. New York: Viking Penguin.

- Kymlicka, W. (1990). *Contemporary Political Philosophy: An Introduction*. Oxford: Oxford University Press.
- Laertius, D. (1925a). Aristippus, in Volume I, Book II of his *Lives of the Eminent Philosophers*, trans. by R. D. Hicks. The Loeb Classical Library, Harvard University Press.
- Laertius, D. (1925b). Epicurus, in Volume II, Book X of his *Lives of the Eminent Philosophers*, trans. by R. D. Hicks. The Loeb Classical Library, Harvard University Press.
- Layard, R. (2003). *Happiness: Has Social Science a Clue?*, Lionel Robbins Memorial Lectures. Available from:
<http://cep.lse.ac.uk/events/lectures/layard/RL030303.pdf>
- Layard, R. (2005). *Happiness: Lessons from a New Science*. Penguin.
- Layard, R. (2010). Measuring Subjective Well-Being, *Science*, 327(5965): 534–535.
- Layard, R. (2011). Happiness, Economist Debates, *The Economist*, 17 May 2011. Available from: <http://www.economist.com/debate/days/view/698>
- Lieberman, M. (2000). Intuition: A Social Cognitive Neuroscience Approach, *Psychological Bulletin*, 126: 109–137.
- Lucas, R. E., Diener, E., & Suh, E. M. (1996). Discriminant Validity of Well-Being Measures, *Journal of Personality and Social Psychology*, 71(3): 616–28.
- Macpherson, F. (ed.) (2011). *The Senses: Classic and Contemporary Philosophical Perspectives*. Oxford University Press.
- McMahon, D. (2006). *Happiness: A History*. Atlantic Press.
- Mendola, J. (2006). Intuitive Hedonism, *Philosophical Studies*, 128(2): 441–477.
- Merton, R. K. (1957). *Social Theory and Social Structure*. Glencoe: The Free Press of Glencoe.

- Mettrick, E. F. (1928). G. E. Moore and Intrinsic Goodness, *International Journal of Ethics*, 38(4): 389–400.
- Metz, T. (2001). The Concept of a Meaningful Life, *American Philosophical Quarterly*, 38(2): 137–153.
- Metz, T. (2002). Recent Work on the Meaning of Life, *Ethics*, 112: 781–814.
- Metz, T. (2003a). Utilitarianism and the Meaning of Life, *Utilitas*, 15(1): 50–70.
- Metz, T. (2003b). The Immortality Requirement for Life's Meaning, *Ratio* (new series), XVI(0034–0006): 161–177.
- Metz, T. (2007). New Developments in the Meaning of Life, *Philosophy Compass*, 2: 196–217.
- Metz, T. (2008). The Meaning of Life, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition). Available from: <http://plato.stanford.edu/archives/fall2008/entries/life-meaning/>
- Michalos, A. C. (2011). What Did Stiglitz, Sen and Fitoussi Get Right and What Did They Get Wrong?, *Social Indicators Research*, 102: 117–129.
- Mill, J. S. (1838). Bentham, *London and Westminster Review*, August 1838.
- Mill, J. S. (1861). *Utilitarianism*. Indianapolis: Bobbs-Merrill, 1957.
- Mill, J. S. (1969). *Autobiography*. J. Stillinger (ed.). Boston: Houghton Mifflin Company.
- Mitsis, P. (1988). *Epicurus' Ethical Theory: The Pleasures of Invulnerability*. Ithaca: Cornell University Press.
- Moore, A. (2011). Hedonism, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Available from: <http://plato.stanford.edu/archives/win2011/entries/hedonism/>
- Moore, A. & Crisp, R. (1996). Welfarism in Moral Theory, *Australasian Journal of Philosophy*, 74(4): 598–613.
- Moore, G. E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

- Murray, E. J. (1964). *Motivation and Emotion*. New Jersey: Prentice-Hall.
- Myers, D. G. (2002). *Intuition: Its Powers and Perils*. New Haven: Yale University Press, 2004.
- Nadelhoffer, T. & Nahmias, E. (2007). The Past and Future of Experimental Philosophy, *Philosophical Explorations*, 10(2): 123–149.
- Nagel, T. (1970). Death, *Noûs*, 4(1): 73–80.
- Nagel, T. (1977). Poverty and Food: Why Charity is Not Enough, in P. Brown & H. Shue (eds.), *Food Policy: The Responsibility of the United States in the Life and Death Choices*. New York: Free Press: 54–62.
- Nagel, T. (1979). Death, in his *Mortal Questions*. Cambridge University Press: 1–10.
- Nagel, T. (1986). *The View from Nowhere*. New York: Oxford University Press.
- Neumayer, E. (2000). In Defence of Historical Accountability for Greenhouse Gas Emissions, *Ecological Economics*, 33: 185–192.
- Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises, *Review of General Psychology*, 2: 175–220.
- Nozick, R. (1974). *Anarchy, State, and Utopia*. Oxford: Blackwell, 1991.
- Nozick, R. (1989). *The Examined Life: Philosophical Meditations*. New York: Simon and Schuster.
- OECD (Organisation for Economic Cooperation and Development) (2012). *Your Better Life Index: Executive Summary*. Organisation for Economic Cooperation and Development. Available from: http://oecdbetterlifeindex.org/wpsystem/wp-content/uploads/2012/02/YourBetterLifeIndex_ExecutiveSummary3.pdf
- O’Keefe, T. (2002). The Cyrenaics on Pleasure, Happiness, and Future-Concern, *Phronesis*, 47(4): 395–416.

- O'Keefe, T. (2005a). Aristippus, *Internet Encyclopedia of Philosophy*. Available from: <http://www.iep.utm.edu/aristip/>
- O'Keefe, T. (2005b). Cyrenaics, *Internet Encyclopedia of Philosophy*. Available from: <http://www.iep.utm.edu/cyren/>
- Oswald, A. J. & Wu, S. (2010). Objective Confirmation of Subjective Measures of Human Well-Being: Evidence from the U.S.A., *Science*, 327(5965): 576–579.
- Page, E. (1999). Intergenerational Justice and Climate Change, *Political Studies*, 47: 53–66.
- Page, E. (2007). Justice Between Generations: Investigating a Sufficientarianism Approach, *Journal of Global Ethics*, 3(1): 3–20.
- Page, E. (2008). Distributing the Burdens of Climate Change, *Environmental Politics*, 17(4): 556–575.
- Page, E. (2011). Climatic Justice and the Fair Distribution of Atmospheric Burdens: A Conjunctive Account, *The Monist*, 94(3): 412–432.
- Pence, G. (2005). Hedonism, in G. Pence (ed.), *A Dictionary of Common Philosophical Terms* (International Edition). McGraw-Hill: 25.
- Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An Empirical Study of Moral Intuitions: Toward an Evolutionary Ethics, *Journal of Personality and Social Psychology*, 64(3): 467–478.
- Plato (1937). Philebus, in *The Dialogues of Plato*, trans. by B. Jowett. New York: Random House.
- Plato (1974). Book II: Preliminaries, in *The Republic*, trans. by D. Lee (2nd ed., revised). Penguin Books, 1983.
- Popper, K. (1952). *The Open Society and its Enemies* (2nd ed., revised). London: Routledge & Kegan Paul.

- Porter, L. W., Bigley, G. A., & Steers, R. M. (2003). *Motivation and Work Behaviour* (7th ed.). Boston: McGraw-Hill Irwin.
- Railton, P. (1984). Alienation, Consequentialism, and the Demands of Morality, *Philosophy & Public Affairs*, 13(2): 134–171.
- Reisiniger, A. & Larsen, H. (2010). Recognising Ethics to Help a Constructive Climate Change Debate, in J. Boston, A. Bradstock, & D. Eng (eds.), *Public Policy: Why Ethics Matters*. ANU E Press: 117–139.
- Rivera-lópez, E. (2007). Are Mental State Welfarism and Our Concern for Nonexperiential Goals Incompatible?, *Pacific Philosophical Quarterly*, 88: 74–91.
- Ryff, C. D. (1989). Happiness is Everything, or is it? Explorations on the Meaning of Psychological Well-Being, *Journal of Personality and Social Psychology*, 69: 719–727.
- Sakai, N., Nakao, K.-I., Ishihara, H., & Kobayashi, M. (2006). Is it Possible to Create a Universe out of a Monopole in the Laboratory?, *Physics Review D*, 74: 024026.
- Samuelson, W., & Zeckhauser, R. (1988). Status Quo Bias in Decision Making, *Journal of Risk and Uncertainty*, 1: 7–59.
- Sarch, A. (2011). Internalism about a Person's Good: Don't Believe it, *Philosophical Studies*, 154: 161–184.
- Sarma, D. (2011). *Classical Indian Philosophy: A Reader*. Columbia University Press.
- Savery, W. (1934). A Defense of Hedonism, *International Journal of Ethics*, 45(1): 1–26.
- Schimmack, U. & Oishi, S. (2005). The Influence of Chronically and Temporarily Accessible Information on Life Satisfaction Judgments, *Journal of Personality and Social Psychology*, 89(3): 395–406.

- Schwarz, N. & Clore, G. L. (1983). Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States, *Journal of Personality and Social Psychology*, 45(3): 513–23.
- Schwarz, N. & Strack, F. (1999). Reports of Subjective Well-Being: Judgemental Processes and Their Methodological Implications, in D. Kahneman, E. Diener, & N. Schwarz (eds.), *Wellbeing: The Foundations of Hedonic Psychology*. New York: Russell Sage: 61–84.
- Schwarz, N. (1987). *Stimmung als Information* [Mood as Information]. Heidelberg: Springer Verlag.
- Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *Philosophical Review*, 117: 245–273.
- Seligman, M. E. P. (2011). *Flourish*. New York: Simon & Schuster.
- Shaw, W. H. (1948). *Moore on Right and Wrong: The Normative Ethics of G.E. Moore*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1995.
- Shue, H. (1992). The Unavoidability of Justice, in A. Hurrell & B. Kingsbury (eds.), *The International Politics of the Environment*. Oxford: Oxford University Press: 373–397.
- Shue, H. (1999). Global Environment and International Inequality, *International Affairs*, 75(3): 531–545.
- Sidgwick, H. (1907). *The Methods of Ethics* (7th Edition). Macmillan.
- Silverstein, M. (2000). In Defense of Happiness: A Response to the Experience Machine, *Social Theory and Practice*, 26(2): 279–300.
- Singer, I. (1996). *Meaning in Life, Volume 1: The Creation of Value*. Baltimore: Johns Hopkins University Press.
- Singer, P. (2008). One atmosphere, in T. Brooks (ed.), *The Global Justice Reader*. Oxford: Blackwell: 667–688. (First published in P. Singer. 2002. *One World*:

- The Ethics of Globalization*. New Haven: Yale University Press: 14–50, 205–208.)
- Singer, P. (1990). *Animal Liberation* (2nd Edition). New York: Avon Books.
- Sinnott-Armstrong, W. (2011). Consequentialism, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition). Available from: <http://plato.stanford.edu/archives/win2011/entries/consequentialism/>
- Smart, J. J. C. (1973). An Outline of a System of Utilitarian Ethics, in J. J. C. Smart & B. Williams (eds.), *Utilitarianism: For and Against*. Cambridge: Cambridge University Press: 3–74.
- Smart, R. N. (1958). Negative Utilitarianism, *Mind*, 67(268): 542–543.
- Smith, B. (2011). Can We Test the Experience Machine?, *Ethical Perspectives*, 18(1): 29–51.
- Smith, J. & Sosa, E. (eds.) (1969). *Mill's Utilitarianism: Text and Criticism*. Belmont CA: Wadsworth.
- Smuts, A. (2011). The Feels Good Theory of Pleasure, *Philosophical Studies*, 155(2): 241–265.
- Sobel, D. (1999). Pleasure as a Mental State, *Utilitas*, 11(2): 230–234.
- Sobel, D. (2002). Varieties of Hedonism, *Journal of Social Philosophy*, 33(2): 240–256.
- Sober, E. & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge: Harvard University Press.
- Starobinsky, A. (2000). Future and Origin of Our Universe: Modern View, *Gravitation and Cosmology*, 6: 157–163.
- Stiglitz, J. E., Sen, A., & Fitoussi, J.-P. (eds.) (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Paris: Commission on the Measurement of Economic Performance and Social Progress.

- Sumner, L. (1996). *Welfare, Happiness, & Ethics*. Oxford: Clarendon press.
- Sunstein, C. R. (2005). Moral Heuristics, *Behavioral and Brain Sciences*, 28: 531–573.
- Sweet, W. (2008). Jeremy Bentham, *Internet Encyclopedia of Philosophy*.
Available from: <http://www.iep.utm.edu/bentham/#H4>
- Tännsjö, T. (1998). *Hedonistic Utilitarianism*. Edinburgh: Edinburgh University Press.
- Tännsjö, T. (2007). Narrow Hedonism, *Journal of Happiness Studies*, 8: 79–98.
- Thomson, G. (1987). *Needs*. London: Routledge and Kegan Paul.
- Tiberius, V. (2004). Cultural Differences and Philosophical Accounts of Well-Being, *The Journal of Happiness Studies*, 5: 293–314.
- Tiberius, V. (2006). Well-Being: Psychological Research for Philosophers, *Philosophy Compass*, 1: 493–505
- Tiberius, V. (2008). *The Reflective Life: Living Wisely Within our Limits*. New York: Oxford University Press.
- Tiberius, V. & Hall, A. (2010): Normative Theory and Psychological Research: Hedonism, Eudaimonism, and Why it Matters, *Journal of Positive Psychology*, 5(3): 212–225.
- Tolstoy, L. (1940). *My Confession*, trans. by A. Maude. London and New York: Oxford University Press.
- Tolstoy, L. (2000). *My Confession*, in E. D. Klemke (ed.), *The Meaning of Life* (2nd Edition). New York: Oxford University Press: 11–20. Reprint from the 1905 Leo Wiener translation published by J M Dent in London.
- Treasury (2010). *Statement of Intent 2010–2015*. Wellington, New Zealand: Treasury.
- Treasury (2011). *Working Towards Higher Living Standards for New Zealanders*. Wellington, New Zealand: Treasury. Available from:

<http://www.treasury.govt.nz/publications/research-policy/tp/higherlivingstandards/tp-hls-may11.pdf>

Turton, D. (2009). The Real Dirt on Happiness Studies: A Reply to 'The Unhappy thing about Happiness Economics', *Real-World Economics Review*, 49: 83–89.

Tversky, A. & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model, *Quarterly Journal of Economics*, 106(4): 1039–1061.

UN (United Nations) Department of Public Information (1997). *UN Conference on Environment and Development (1992)* (revised 23 May 1997). Available from: www.un.org/geninfo/bp/enviro.html

UNFCCC (United Nations Framework Convention on Climate Change) (1992). *Report of the United Nations Conference on Environment and Development*. United Nations Framework Convention on Climate Change. A/CONF.151/26 (Vol. I). Available from: www.poptel.org.uk/nssd/otherdocuments/Agenda21.doc

UNFCCC (United Nations Framework Convention on Climate Change) (No date). *List of Annex I Parties to the Convention*. United Nations Framework Convention on Climate Change. Available from: http://unfccc.int/parties_and_observers/parties/annex_i/items/2774.php

Unger, P. (1996). *Living High and Letting Die*. New York: Oxford University Press.

Urry, H. L., Nitschke, J. B., Dolski, I., Jackson, D. C., Dalton, K. M., Mueller, C. J., Rosenkranz, M. A., Ryff, C. D., Singer, B. H., & Davidson, R. J. (2004). Making a Life Worth Living: Neural Correlates of Well-Being, *Psychological Science*, 15(6): 367–372.

- Van Lange, P. A. M., Tarris, T. W., & Vonk, R. (1997). Dilemmas of Academic Practice: Perceptions of Superiority among Social Psychologists, *European Journal of Social Psychology*, 27: 675–685.
- Van Wyk, R. N. (1990). *Introduction to Ethics*. New York: St. Martin's Press.
- Veenhoven, R. (2010). Greater Happiness for a Greater Number: Is That Possible and Desirable?, *Journal of Happiness Studies*, 11: 605–629.
- Vesey, G. & Foulkes, P. (1990). Hedonism, in G. Vesey & P. Foulkes (eds.), *Collins Dictionary of Philosophy*. London and Glasgow: Collins: 131–132.
- Vroom, V. (1964). *Work and Motivation*. New York: John Wiley & Sons.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scale, *Journal of Personality and Social Psychology*, 54: 1063–1070.
- Weijers, D. (2010). A Warning to Policy-Makers: What Exactly is Well-Being Anyway?, *Social Europe Journal*, 10 September 2010. Available from: <http://www.social-europe.eu/2010/09/a-warning-to-policy-makers-what-exactly-is-well-being-anyway/>
- Weijers, D. (2011a). Hedonism, *Internet Encyclopedia of Philosophy*. Available from: <http://www.iep.utm.edu/hedonism/>
- Weijers, D. (2011b). The Experience Machine Objection to Hedonism, in M. Bruce & S. Barbone (eds.), *Just the Arguments*. Wiley-Blackwell: 229–231.
- Weijers, D. (2011c). Reality Doesn't Really Matter, in D. K. Johnson (ed.), *Inception and Philosophy*. Wiley-Blackwell: 92–107.
- Weijers, D. (2011d). National Accounts of Well-Being: What Should We Put in the Census?, *Social Europe Journal*, 15 February 2011. Available from: <http://www.social-europe.eu/2011/02/national-accounts-of-well-being-what-should-we-put-in-the-census/>

- Weijers, D. (2012). National Accounts that Make Governments More Efficient by Making Them More Accountable. Available from the official Bhutanese government's website for the UN meeting on 'Happiness and Wellbeing: Defining a New Economic Paradigm':
<http://www.2apr.gov.bt/images/stories/2page/danweijers.docx>
- Weijers, D. (forthcoming-a). Intuitive Biases in Judgements about Thought Experiments: The Experience Machine Revisited, *Philosophical Writings*.
- Weijers, D. (forthcoming-b). We Can Test the Experience Machine: Reply to Smith, *Ethical Perspectives*, 19(1).
- Weijers, D., Eng, D., & Das, R. (2010). Sharing the Responsibility of Dealing with Climate Change: Interpreting the Principle of Common but Differentiated Responsibilities, in J. Boston, A. Bradstock, & D. Eng (eds.), *Public Policy: Why Ethics Matters*, ANU E Press: 141–158.
- Weijers, D. & Jarden, A. (2011). Wipe that Smile Off Your Face, *The Philosopher's Magazine*, 52: 53–58.
- Wild, J. (1927). The Resurrection of Hedonism, *International Journal of Ethics*, 38(1): 11–26.
- Wolf, S. (1997). Happiness and Meaning: Two Aspects of the Good Life, *Social Philosophy and Policy*, 14: 207–225.
- Woodward, J. & Allman, J. (2007). Moral Intuition: Its Neural Substrates and Normative Significance, *Journal of Physiology-Paris*, 101(4–6): 179–202.