9-2009

On the Objectivity of Welfare

Alexander F. Sarch
University of Massachusetts Amherst, asarch@philos.umass.edu

ON THE OBJECTIVITY OF WELFARE

A Dissertation Presented

by

ALEXANDER F. SARCH

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2009

Philosophy

ON THE OBJECTIVITY OF WELFARE

A Dissertation Presented

by

ALEXANDER F. SARCH

Approved as to content and style by:

Fred Feldman, Chair

Hilary Kornblith, Member

Gareth Matthews, Member

David Lahti, Member

Phillip Bricker, Department Head
Philosophy

DEDICATION

To my Mom, Dad and Stepfather, with love.

ACKNOWLEDGEMENTS

There are many people to whom I owe a huge debt for their help in writing this dissertation. First and foremost, I thank Fred Feldman. He read and commented on multiple versions of every chapter of this dissertation. The emails he wrote to me with suggestions and criticisms could fill several books. I am tremendously grateful not only for all the time he dedicated to helping me with this project, but also for his refusal to let me get away with mistakes, imprecision or carelessness. He inspired me to be consumed by this work, and he matched me with his own investment in my projects. I cannot adequately express how grateful I am for his help.

I am also heavily indebted to my other committee members for all their help and support. I want to thank Hilary Kornblith for many helpful conversations about various parts of this dissertation, as well as all the words of encouragement he gave me throughout grad school. I am very grateful to Gary Matthews for his support and enthusiasm, and to David Lahti for many helpful conversations and suggestions.

I owe a thousand thanks to John Arthur Skard for the many discussions about the Discount/Inflation Theory presented in chapter 8, and for his tireless help in constructing an accurate mathematical representation of this theory (and several others besides). Chapter 8 would not have turned out as well as it did had it not been for his generous assistance.

Many members of the philosophical community at UMass also deserve my heartfelt thanks. Thanks especially to Kelly Trogdon for countless discussions (mostly late at night at Packard's) about various crazy ideas of mine, and thanks to both him and Sam Cowling for all their encouragement and for helping me stay motivated. Thanks to Phil Bricker, Pete Graham and Chris Meachem for helpful discussions about various chapters. Thanks to Ernesto Garcia, Michael Rubin, Scott Hill, Kristian Olsen and James Patten for taking part in a colloquium on the paper that eventually became appendix to chapter 3, and thanks to Dan Doviak, Michael Rubin, Scott Hill, Jeff Dunn and Kristoffer Ahlstrom for many helpful conversations. Many thanks also to Zoe Geva for all her wonderful encouragement and insights.

I owe a special thanks to Anna White-Nockleby for her love and support, as well as for the skeptical conversations that compelled me to write chapter 1.

Finally, I am tremendously grateful to my mother, my step-father, my father and my sister for their unconditional support, as well as for putting up with me throughout the stressful process of writing this dissertation. They have always been enormously caring and supportive, and have all given me (in their own ways) exceptional amounts of help in writing this dissertation. I would not have been able to complete the project without their love and encouragement.

ABSTRACT

‘ON THE OBJECTIVITY OF WELFARE’

SEPTEMBER 2009

ALEXANDER SARCH, B.A., CORNELL UNIVERSITY

M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Fred Feldman

This dissertation is structured in such a way as to gradually home in on the true theory of welfare. I start with the whole field of possible theories of welfare and then proceed by narrowing down the options in a series of steps.

The first step, undertaken in chapter 2, is to argue that the true theory of welfare must be what I call a *partly response independent* theory. First I reject the *entirely response independent* theories because there are widely-shared intuitions suggesting that some psychological responses are indeed relevant to welfare. Then I reject the *entirely response dependent* theories because there are other central intuitions suggesting that our welfare is not determined *solely* by our psychological responses. Thus I reach the preliminary conclusion that welfare must involve *some* response independent (or objective) component.

The next step is to consider the most promising theories in the partly response independent category. In particular, I formulate, refine and ultimately reject what seem to

be the main *monistic* theories that have been proposed in this category. In chapter 4, I reject the Adjusted-Enjoyment Theories of Welfare because they cannot account for the claim that a life containing no pleasure or pain can still contain a positive amount of welfare (e.g. if it's a particularly successful life). Then in chapters 5-7, I discuss Desire Satisfaction theories of welfare. I argue that even the most promising of these theories – e.g. Worthiness Adjusted Desire Satisfactionism – are problematic because they cannot accommodate the claim that a life containing no success with respect to worthwhile projects can still contain a positive amount of welfare (e.g. if it's a particularly pleasant life).

Finally, I suggest that in order to accommodate the intuitions that led to the rejection of all these other theories of welfare, what is needed is a multi-component theory. In the final chapter, I formulate a multi-component theory that is particularly promising. Not only does it avoid the problems of the monistic theories discussed earlier, but, by incorporating a number of novel mathematical devices, it avoids problems that undermine several other initially promising multi-component theories of welfare.

TABLE OF CONTENTS

APPENDICES

LIST OF FIGURES

FIGURE

INTRODUCTION

This dissertation has the title it does because my main aim here is to investigate and ultimately defend the view that the true theory of welfare must involve at least some objective component. The dissertation is structured in such a way as to gradually home in on the true theory of welfare. I start with the whole field of possible theories of welfare and then proceed by narrowing down the options in a series of steps.

The first step, which is undertaken mainly in chapter 2, is to argue that the true theory of welfare (whatever it is) is to be found among the *partly response independent* theories. On the one hand, I reject the *entirely response independent* theories on the grounds that there are widely-shared intuitions that suggest that some psychological responses are indeed relevant to welfare. And on the other hand, I reject the *entirely response dependent* theories on the grounds that there are other central intuitions that suggest that our welfare is not determined *solely* by our psychological responses. On the basis of these considerations, I reach the preliminary conclusion that welfare must involve *some* response independent (or objective) component. (In chapter 3, I defend this preliminary conclusion from a common sort of objection.)

The next step is to consider the most promising theories of welfare that philosophers have put forward in the partly response independent category. In particular, I formulate, refine and ultimately reject what seem to be the main *monistic* theories in this category. In chapter 4, I discuss Adjusted-Enjoyment Theories of Welfare (like Feldman's theory DAIAH) and reject them because they cannot account for the claim that a life containing no pleasure or pain can still contain a positive amount of welfare (e.g. if it's a particularly successful life). Then in chapters 5-7, I discuss Desire Satisfaction theories of welfare. I argue that even the most promising of these theories – such as Worthiness Adjusted Desire Satisfactionism (a partly response independent theory) – is problematic because they cannot accommodate the claim that a life containing no success with respect to worthwhile projects can still contain a positive amount of welfare (e.g. if it's a particularly pleasant life).

Finally, I suggest that in order to accommodate the intuitions that led to the rejection of all these other theories of welfare, what is needed is a multi-component theory of welfare. In the final chapter, I formulate a multi-component theory of welfare that I think is particularly promising. Not only does it avoid the problems of the monistic theories discussed earlier, but, by incorporating a number of novel mathematical devices, it avoids problems that undermine several other initially promising multi-component theories of welfare. I've discovered that I am not very good at thinking of names for theories, so I called my theory 'the Discount/Inflation Version of the Happiness and Success Theory of Welfare', or DIVHSTW. (My apologies to anyone who tries to pronounce this.)

Having sketched the overarching structure of the dissertation, it might be helpful to now provide a brief explanation of the main aims of each chapter:

Chapter 1: The first chapter of this dissertation deals primarily with methodology. The standard methodology that philosophers use to defend or attack various theories of welfare is that of reflective equilibrium. This method has recently begun to come under attack, however. While I am sympathetic to some of these criticisms, I think it is unlikely that we will arrive at a more successful methodology for dealing with ethical questions any time soon.

Nonetheless, a significant obstacle remains when the method of reflective equilibrium is applied specifically to the topic of welfare. The problem is that we don't have a handle on what the concept of welfare *is*. More specifically, we are not in a position to say what distinguishes the concept of welfare from the concept of other sorts of value, such as moral value, aesthetic value, perfectionist value (excellence), and so on. Why is this a problem? According to the method of reflective equilibrium, theories about what welfare consists in (like hedonism, desire-satisfactionism, or objective list theories) are to be evaluated by appeal to how well they account for our intuitions about whether and how much various things would enhance one's welfare. However, before we can test a theory of welfare against intuitions of this sort, we need to know precisely what they are supposed to be intuitions *about*. That is, we need to get clear on what the concept of welfare *is*, and what distinguishes it from other sorts of value, before we will be in a position to evaluate theories of welfare by appeal to our intuitions about that concept. In this chapter, I show why five well-known attempts to say what is distinctive of the concept of welfare fail. I then conclude by proposing my own account of how the concept of welfare value may be distinguished from other types of value. If my account is successful, the upshot is that we may continue to use the method of reflective equilibrium in theorizing about welfare.

Chapter 2**:** In this chapter, I show why the standard taxonomy for theories of welfare is unsystematic and incomplete, and then I develop a new taxonomy that is both systematic and exhaustive. In particular, my taxonomy groups theories of welfare into the *entirely response dependent theories* (one important subclass of which is the mental state theories), the *partly response independent theories* and the *entirely response independent theories*. This taxonomy is not merely for show; it does some work. In particular, my taxonomy is useful because it groups together theories that share certain fundamental flaws. Thus my taxonomy provides a convenient way to offer blanket arguments against whole categories of theories. In this chapter, I use my taxonomy to argue that the true theory of welfare, whatever it is, must fall under the category of the *partly response independent theories*. By clearing the playing field early on in this way, I will be able to focus my investigation on the theories that have the most going for them.

Chapter 3: The conclusion I argue for in chapter 2 amounts to the claim that welfare must have at least some ‘objective’ component. However, the entirely response dependent theories of welfare, which have a more ‘subjective’ character, seem to be in vogue at present, and so many philosophers may find it difficult to accept the conclusion of chapter 2. The main reason for this lack of sympathy for response independent theories, I suggest, is a tacit acceptance of some kind of *internalism* (whether about reasons or about welfare). So in order to bolster my argument for the idea that the true theory of welfare must be partly response independent (i.e. that welfare must include some kind of ‘objective’ component), I devote this chapter to arguing against internalism.

More specifically, what I aim to accomplish in chapter 3 is this. Some might have thought that a well-known view called *internalism about reasons* gives reason to prefer the response dependent theories (like Hedonism or Desire Satisfactionism) over the response independent theories. This is an argument that many philosophers seem to tacitly accept but no one has defended in depth. Nonetheless, I think an uncritical acceptance of it is what underlies much of the resistance to the response independent theories of well-being. I begin by formulating the argument as precisely and plausibly as I can, but then I go on to reject it. For, as it turns out, not only are the response independent theories of welfare incompatible with internalism about reasons, but the response *de*pendent theories are too! Since no theory of welfare is compatible with internalism about reasons, I suggest that we should reject the view altogether.

This chapter is followed by an appendix in which I discuss Connie Rosati's recent attempt to directly defend internalism about welfare. She presents five *prima facie* formidable arguments in favor of her favorite version of internalism. But I argue that all five arguments, on closer inspection, are unsound. Rosati thus fails to establish internalism about welfare.

Chapter 4**:** Up to this point, I have been concerned to argue that the true theory of welfare must have an 'objective' component, i.e. that the true theory of welfare is a partly response independent theory. In the remaining chapters of this dissertation, I consider three of the main kinds of partly response independent theory that philosophers have defended: *adjusted enjoyment theories, objectively restricted desire satisfactionism,* and *hybrid theories*.

Chapter 4 deals with the Adjusted Enjoyment Theories. I argue against Mill's entirely response *de*pendent Adjusted Enjoyment Theory, and then consider four recent attempts to defend a partly response *in*dependent Adjusted Enjoyment Theory. Parfit, Darwall, Adams and Feldman all defend partly response independent Adjusted Enjoyment Theories. Parfit's and Darwall's are underdeveloped. Adams' theory of welfare is problematic because Adams' account of one of the key notions that his theory rests on, viz. the notion of excellence, suffers from major difficulties. Feldman's theory, called Desert-Adjusted Intrinsic Attitudinal Hedonism, is thus left as the most plausible Adjusted Enjoyment Theory on offer. Some work needs to be done to provide a systematic account of the notion of desert, which Feldman's theory crucially depends upon. But that work can be done, I think, and I attempt to do it in this chapter.

Nonetheless, I conclude the chapter by raising an objection that threatens every Adjusted Enjoyment Theory, even Feldman's. The objection is that no theory of this type is compatible with the intuition that a life that is entirely devoid of enjoyment may still be worth living to some extent (perhaps merely quite a small extent). The upshot is that while Feldman's theory is compelling, it cannot represent the whole truth about welfare.

***Chapters 5-6*:** These two chapters deal with technical problems for Desire Satisfactionism. Chapter 5 deals with problems for the view that concern desire and time, while chapter 6 attempts to solve three other difficult technical problems for the view: the problem of double counting, the problem of partially fulfilled desires, and the problem of irrelevant desires. I formulate a view called Cloud Desire Satisfactionism, which I think solves all these problems. Insofar as you want to be a Desire Satisfactionist, I think you should be a Cloud Desire Satisfactionist.

***Chapter** 7*: My aim in this chapter is to argue that a certain partly response independent version of the desire satisfaction theory*,* namely *Worthiness Adjusted Cloud Desire Satisfactionism* (WACDS), is the most promising theory in the desire satisfaction family. This should not come as a surprise, considering the arguments I gave in chapter 2 for the claim that the true theory of welfare is to be found in the partly response independent category. But the argument of chapter 2 was sweeping, while the present chapter proceeds more carefully. I argue that the desire satisfactionist can avoid certain kinds of problem cases (involving intuitively defective desires) only by formulating a version of the theory that is partly response *independent*. Nonetheless, I argue that even the best version of Desire Satisfactionism – viz. WACDS – faces another problem that is serious enough to warrant rejecting the view altogether.

Chapter 8: In this chapter, I discuss the question of how to develop a theory of well-being that fits with the intuitions that led us, in previous chapters, to reject other influential theories of well-being. I endorse a type of theory for which a fitting label is 'the Happiness and Success Theory'. What makes a theory belong to this type is that it makes welfare be a function of two things: how happy you are (i.e. how good you feel) and how successful you are in accomplishing worthwhile goals. A number of philosophers have proposed multi-component theories of well-being of this sort, but no one I know of has stated any such multi-component theory in full detail. In particular, none of them discuss the question of how the *math* in such a theory should be worked out. But this is important because the many different ways in which the math can be done for multi-component theories provides a rich set of resources for dealing with problem cases. I formulate several versions of the Happiness and Success Theory that make use of mathematical resources of this sort to avoid a number of potential problems. In fact, I end up endorsing one of them – namely, the Discount/Inflation version of the Happiness and Success Theory – because this theory seems to be able to avoid the main problems of virtually every other theory of welfare that I consider in this dissertation.

CHAPTER 1

METHODOLOGICAL FOUNDATIONS

In the chapters that follow, I will be considering a number of theories about individual well-being. These theories are evaluative ones: they purport to tell us what determines the degree to which a life goes well (or poorly) for the one who lives it. Moreover, I will be considering a number of arguments in favor of or against various theories of well-being. I will claim that some of these theories are more plausible than others. At times, I will even suggest that some of them are beyond repair and should be rejected altogether. Before I get into the business of actually *doing* moral philosophy, however, I think I need to say something about the nature of my project.

1.1 Normative Ethics and Metaethics

In metaethics, one seeks answers to higher-order questions about various evaluative concepts like moral rightness, well-being, virtue, and so on. Thus, a metaethicist might ask linguistic questions like 'What do terms like "morally right action" or "a life high in well-being" mean?' and 'Are sentences like "action *a* is morally wrong" or "person A is better off than person B" capable of being true or false in the standard way, or are they true or false only relative to the beliefs of some group of people?'. A metaethicist might also ask metaphysical questions like 'Are properties like moral rightness or the property of being welfare-enhancing *real* properties, or are they merely fictions?' and 'Are moral properties natural properties or non-natural properties?' Moreover, a metaethicist might ask epistemological questions like 'What is the process by which we come to know moral facts, assuming there are any?'

The business of metaethics is to be contrasted with the project of normative ethics, or substantive moral philosophy. Whereas the metaethicist is concerned with what the metaphysical status of moral properties is, the normative ethicist is concerned with the substantive question of what *makes* something instantiate a given moral property. What is it in virtue of which something possesses the property of moral rightness, say, or the property of intrinsic goodness for a person? Thus when it comes to moral rightness, a normative ethicist is concerned to discover what natural (descriptive) property it is that moral rightness supervenes on. That is, she is looking for a non-trivial true statement of the form 'An action, *a*, is morally right if and only if *a* instantiates N', where 'N' stands for some complex natural property (like utility maximization, the property of not using anybody merely as a means, or what have you). Such a statement is a *criterion* of moral rightness. When it comes to the property of being welfare-enhancing, too, the normative ethicist is interested in discovering what natural properties it supervenes on. She is looking for a non-trivial true statement of the form 'X in itself enhances person P's well-being if and only if X instantiates N*', where 'N*' stands for some complex natural property (like being an episode of pleasure, or being a state of desire satisfaction, or what have you). Such a statement would be a *criterion* of intrinsic goodness for a person.

This dissertation is primarily concerned with normative ethics. I will be discussing various theories about what natural property it is that well-being supervenes on. And I will for the most part avoid metaethical questions about, e.g., the metaphysical status of the property of enhancing someone's well-being. I assume that it is a real property, and I assume that there is some complex natural property that it supervenes on. My main task is to figure out what this natural property can reasonably be taken to be.

Nonetheless, one might want to know what methods we are to use in seeking answers to normative ethical questions. Strictly speaking, this is a metaethical question. It is, after all, a question in moral epistemology. However, I think it is important to take a moment to reflect on the methods I will be using here, and to acknowledge and try to address (however superficially) some of the doubts that people who are not in the habit of doing normative ethics might be inclined to have about the standard methods of normative ethicists. Thus while my dissertation is primarily about normative ethics, I begin with a brief digression into metaethics.

1.2 The Method of Reflective Equilibrium

The sort of methodology I will be employing in the chapters that follow is the well-known method of Reflective Equilibrium. This phrase was originally made famous by Rawls,[1] but Shelly Kagan's description of the practice of moral philosophers is both sufficiently clear and concise:

> in defending a moral theory, we must see how well that theory fits in with a wide variety of judgments that we are inclined to make about many different matters. We have opinions about cases, about principles, about the nature of morality, about what counts as an adequate explanation, and more. Some of these opinions are fairly specific, others are more general; some are arrived at rather "intuitively" and spontaneously, others only after considerable reflection; some are extremely difficult to give up, others are more easily abandoned. We try to find the moral theory that provides the best overall fit with this eclectic set of beliefs. But if – as seems overwhelmingly likely – no theory can actually accommodate all of the relevant initial beliefs, we revise the set: we alter our beliefs, and reevaluate our theories, until we arrive as best we can at a theory that seems on balance to be more plausible than any of its rivals. Ultimately, then, defending a normative theory is a matter of arguing that it provides the best overall fit with our various considered judgments. (Kagan, p. 15)

Thus, to put it very roughly, the method of Reflective Equilibrium consists in seeking to provide the best possible theoretical systematization of our *pre-theoretical beliefs* or *intuitive judgments* about a given moral topic, M. An important part of the procedure is to test our theories about M against our intuitive judgments about M, and then revise our theories so that they better match our intuitive judgments. But as Kagan points out, this is not all the method consists in. It also typically involves revising our intuitive judgments somewhat so that they better match our best theories about M.[2]

[1] Rawls, 1999, pp. 40-46. (Also see Daniels, 1979.)

[2] Rawls stresses this point himself: 'When a person is presented with an intuitively appealing account of his sense of justice (one, say, which embodies various reasonable and natural presumptions), he may well revise his judgments to conform to its principles even though the theory does not fit his existing judgments exactly. He is especially likely to do this if he can find an explanation for the deviations which undermines his confidence in his original judgments and if the conception presented yields a judgment which he finds he can now accept. From the standpoint of moral philosophy, the best account of a person's sense of justice is not the one which fits his judgments prior to his examining any conception of justice but rather the one which matches his judgments in reflective equilibrium.' (Rawls, 1999, ch. 1, sec. 9, p. 42)

Thus, the way in which moral theories meet the body of evidence that consists of our intuitive judgments is complex, according to the method of Reflective Equilibrium. How exactly are we to respond when we find that a given theory of ours conflicts with some of our intuitive judgments? It depends, since our intuitive judgments may differ when it comes to the degree to which we are committed to them. Some intuitions might be central and virtually impossible for us to give up; others might be much less important to us and therefore easier to give up.[3] So if we find that the moral theory we are investigating conflicts with certain very central intuitions that we are deeply committed to, then we would typically have two choices: revise the theory in order to make it cohere better with these intuitions, or else abandon the theory. On the other hand, if we find that the theory conflicts with relatively peripheral intuitions that we are not deeply committed to, then we might be justified in simply rejecting or, better yet, explaining away the offending intuitions. Thus the proper response to a given conflict between a theory and a certain set of intuitions will depend in large part on the centrality of the intuitions in question.

Given these intricacies, we can expect an investigation of well-being that employs the method of Reflective Equilibrium to look something like this. We would begin by considering some typical intuitive judgments that we are inclined to make about how well off particular people are in particular circumstances, and then somehow come up with a theory that purports to systematize these intuitions. We would then proceed to test this theory against other of our intuitive judgments about well-being. In all likelihood, we will discover that the theory is not a very good overall match with the body of our intuitive judgments. So we would attempt to revise the theory so as to better fit these intuitive judgments. This cycle of testing and revision would continue until we arrive at some fairly sophisticated theory that fits reasonably well with our intuitions and has a decent

[3] As a result of this, a theory's degree of coherence with our intuitive judgments must be understood in some way that takes the centrality, or weight, of our various intuitive judgments into consideration. As a rough first pass, we could perhaps say that the more of our intuitions that a given theory coheres with, and the weightier the intuitions it coheres with are, the better fit there will be between the theory and our intuitions.

However, it is going to be quite tricky to calculate precisely the degree to which a theory coheres with our intuitions because (among other things) there is no limit on the number of intuitions that we could in principle have. After all, there are infinitely many possible cases for us to have intuitions about, and so there are presumably also infinitely many intuitions that we might have. This means it would be too simple to take it that the degree of coherence between theory T and our intuitions just equals the centrality-adjusted percentage of our intuitions that T is able to accommodate.

amount of explanatory power. Because of its plausibility, we might now be inclined to believe this theory's implications even in the face of further clashes with intuition. Accordingly, we might find it best to revise the intuitive judgments we are inclined to make rather than modify the theory. Thus we would be led into a process of give and take between theory and intuitions, where we revise our intuitions somewhat so as to better fit our theory, then revise our theory somewhat so as to better fit our intuitions, then revise our intuitions even further, and so on. Eventually we will reach a point where the degree of fit between theory and intuition cannot be enhanced any further by revising either the theory or our intuitions. Then we will have reached the point of *reflective equilibrium*. The theory that we in this way end up in reflective equilibrium with is as good a systematization of our intuitive judgments as can be hoped for, and this theory may be taken to be true.[4]

This gives us a rough understanding of the method that moral philosophers typically use to investigate a topic like well-being. However, some people (especially if they are not in the habit of doing moral philosophy in this way) might have some doubts about the method of Reflective Equilibrium. In particular, why think that a theoretical systematization of our intuitive moral judgments that is arrived at via this method is likely to be *true*? Why think that our intuitive judgments about particular cases can count as *evidence* for or against theories in normative ethics in the first place?[5]

[4] Of course, as Kagan points out, our views about other, more general matters – like what constitutes a good explanation, or our views about human psychology – may also be relevant in assessing the relative plausibility of competing theories. A theory that coheres as well as possible with all these more general beliefs as well is said to be in *wide* reflective equilibrium with our beliefs.

[5] One might think that this latter question is misguided. For one might think that if our intuitive judgments can be revised on the basis of a given theory we accept, then there is no way that intuitive judgments can be *evidence*. After all, bodies of evidence in other areas of inquiry (e.g. observations in science) are not revisable in light of theory in this way, are they?

In fact, there is a good case to be made that evidence in science, e.g. observation, is revisable in light of theory in *precisely* this way. For instance, if we have a plausible and well-confirmed theory of optics, and we proceed to make some observations that conflict with this theory, then we might well be justified in explaining away these observations as aberrations or noise (e.g. by saying 'the lenses in that particular telescope are out of alignment'), rather than taking our optical theory to be refuted by these observations. Thus it seems that under some conditions, we would indeed be epistemically justified in revising our observations in order to preserve our best theory. It might be hard to say exactly what the appropriate conditions are here, but whatever they are, presumably there is an analogous set of conditions under which our moral intuitions may legitimately revised in light of our best moral theories. (For more on the whether moral intuitions play a role in moral theorizing that is analogous to the role that observations play in scientific theorizing, see Cummins, 1998, and Harman, 1977, pp. 3-6.)

Strictly speaking, the answers to these questions will depend on the position that one takes in metaethics, and of course, metaethics is not my primary concern here. But I think these questions are good ones, and I would feel uncomfortable using this method without at least briefly trying to see whether there are any plausible answers to these questions on offer. Fortunately, it seems that there are some such answers. In the next section, I'll briefly discuss four ways to understand what it is that we are doing when we use the method of Reflective Equilibrium to do normative ethics. Each one suggests an answer to the questions of why theories arrived at via this method can be expected to be true, and why our intuitive moral judgments can be expected to count as evidence. My hope is that by showing that there are at least some plausible answers to the doubts one might have about the method of Reflective Equilibrium, I will have done enough to justify my going ahead and using this method myself.

1.3 Four Accounts of the Method of Reflective Equilibrium

There are, it seems to me, four main ways to understand what is going on when we use the method of Reflective Equilibrium in settling on a theory in normative ethics. Each account starts with an idea about how to understand the subject matter of theories of normative ethical theories – i.e. what they are supposed to be theories *about* – and then proceeds to explain why our moral intuitions count as evidence for or against these theories. What is a moral intuition? For simplicity's sake, I'm going to say that you have *a particular-case intuition* about some moral property, P, if it's the case that, when a given case or scenario has been described to you in sufficient detail, it seems to you (but not because of, say, inference or memory) that this case or scenario is an instance of P.[6] So what each of the four accounts I'll discuss in this section is trying to explain is why we should think that a theory that systematizes these particular-case intuitions of ours is likely to be a *true* theory.

[6] Compare this to Bealer's account of intuition. (Cf. Bealer, 1999, p. 30) As the passages from Kagan and Sumner suggest, there may be other types of intuitions, in addition to particular case intuitions. For instance, one would be having an intuition about a certain generalization if one considers that generalization and it seems to one that it is true. For simplicity, however, I will focus primarily on particular-case intuitions in what follows.

Because the main focus of this dissertation is normative ethics, my discussion of these four proposals must remain superficial. I will not be able to fully canvas the arguments for and against each proposal, and so I will not be in a position to decide between them. To do that would require a lot of serious work in metaethics. Here I will confine myself to merely sketching each proposal and pointing out the main challenges it faces.

1.3.1 Direct Awareness

According to what I'll call the Direct Awareness account, we are directly aware of substantive moral truths through our particular-case moral intuitions. Our particular-case moral intuitions count as evidence for or against normative ethical theories because each of these intuitions will (under certain privileged circumstances) constitute a direct awareness of the moral truth. Therefore they are capable of conferring *a priori* justification. What theories in normative ethics are theories *about*, on this proposal, is the moral reality that we're directly acquainted with through intuition. And this is why a theory that puts us into reflective equilibrium with our particular-case moral intuitions is likely to be *true*.

This sort of approach to defending intuition-based methods has been developed in great detail by a number of contemporary philosophers. Perhaps most importantly, George Bealer offers this kind of defense of the standard intuition-based method as it is used in philosophy in general.[7] Michael Huemer argues on similar grounds that the standard intuition-based method is capable of issuing in justified moral theories.[8] I'm not going to discuss the specifics of Bealer's or Huemer's accounts here. I will merely sketch the basics of the sort of view they defend, as it applies to moral theorizing.

According to the Direct Awareness account, what a moral theory is theory *about* is the conditions under which a given moral property would be instantiated. Moral properties are construed as universals, i.e. as necessarily existent abstract entities that different things (whether they be objects, states of affairs, or what have you) can have in common. It is possible for us, via the faculty of moral intuition, to have direct knowledge of when a given moral property, P, is instantiated. Just as we can see, via the faculty of

[7] Cf. Bealer, 1996, 1998, and 1999.

[8] Cf. Huemer, 2005. (In fact, Huemer relies on the work of Bealer in giving his own account of moral epistemology.)

vision, when certain physical objects are right in front of us, so too can we ‘see,’ via the faculty of moral intuition, when a given object or scenario instantiates P. In order for one to be able to have this sort of particular-case intuition about P, however, one must have *grasped* P. One’s grasp of P may be more or less determinate. If one’s grasp of P is less than fully determinate, then one’s intuitions about whether P is instantiated in various cases might be wrong. But if one has a fully determinate grasp of P, then this will guarantee that one’s intuitions about whether P is instantiated by some object or scenario (which has been described to one in sufficient detail) *will be true* – provided, of course that, one is not asleep, unconscious, drugged, or in some other kind of deficient cognitive conditions. Thus, our particular-case intuitions about some moral property that we determinately grasp will be a reliable guide to the truth about what things have this property. This, then, explains why our intuitions about moral properties that we determinately grasp will count as evidence. Because such intuitions would be instances of a *direct awareness* of the moral facts, these intuitions would be *a priori* justified. Accordingly, it would make good sense to test our normative ethical theories against our particular-case moral intuitions.

This is just a superficial sketch of how Direct Awareness account would attempt to justify the practice of testing our normative ethical theories (e.g. of well-being, or moral rightness, or what have you) against our particular-case moral intuitions. However, even this superficial sketch should be enough to show that the Direct Awareness view is open to a number of challenges. For one thing, one might question the picture of moral properties as independently existing universals that the Direct Awareness account seems to presuppose. Moreover, one might question whether we really have a faculty of moral intuition by which we gain direct knowledge of when various moral properties are instantiated. Or else one might question whether our grasp of various moral properties really is fully determinate. But perhaps a sophisticated defender of the Direct Awareness proposal would be able to answer these worries.[9]

However, I think there is a certain problem for the view that is even more worrying than these preliminary problems. In particular, one might think that the notion of determinately grasping a moral property that the Direct Awareness account employs ends

[9] Huemer, for instance, addresses the ‘no such faculty’ objection. (Cf. Huemer, 2005, pp. 107-115)

up making that account question-begging.[10] After all, the Direct Awareness account says that our particular-case intuitions about some moral property P are going to be reliably tied to the truth, and hence carry epistemic weight, if and only if we determinately grasp P. However, 'determinately grasping P' is simply defined as 'grasping P in such a way that one's particular case intuitions about P are guaranteed to be true (under sufficiently good cognitive conditions)'.[11] And this, of course, would make the Direct Awareness explanation of why our intuitions have epistemic weight turn out to be trivial. For it amounts to saying that our intuitions about P have epistemic weight if and only if we grasp P in such a way that our intuitions about P have epistemic weight.

Thus to make the Direct Awareness account successful, we would need an account of what it is to determinately grasp a moral property that does not lead to this kind of circularity. I am currently not able to see how such an account would go. So, for the time being at least, I have doubts about the prospects for getting the Direct Awareness defense of the method of Reflective Equilibrium in normative ethics to succeed.

1.3.2 Systematizing Private Moral Views

Another way to understand the method of Reflective Equilibrium would be to say that what one is doing when using the method is simply seeking to systematize one's *own* pre-theoretical moral views. According to this account, which I'll call the Private Moral Views account, a theory arrived at via the method of Reflective Equilibrium is a theory that organizes and makes coherent your own pre-existing moral views. The reason your intuitive judgments about particular cases count as evidence, then, is that they *express* the thing that normative theories are theories about (viz. your own pre-theoretical moral views).

David Lewis, at least in places, seems to endorse a proposal of this sort when it comes to philosophical theorizing generally. For instance, he writes:

> Our "intuitions" are simply opinions; our philosophical theories are the same. Some are commonsensical, some are sophisticated; some are particular, some are general; some are more firmly held, some less. But they are all opinions, and a reasonable goal for a philosopher is to try to bring them into equilibrium. Our common task is to find out what equilibria there are that can withstand examination, but it remains for each of us to come to rest at one or another of them.[12]

[10] I have argued that this is the case specifically when it comes to Bealer's views. (Cf. Sarch, 2009)

[11] Bealer's account of determinate concept possession makes this clear. (Cf. Bealer, 1999, p. 41)

[12] Lewis, 1983, p. *x*. (Also see Lewis, 1973, p. 88)

The proposal as applied to moral philosophy, then, would be that what one does when using the method of Reflective Equilibrium in ethics is to seek a theory that to the greatest possible degree harmonizes and coheres with one's own pre-reflective moral beliefs.

What role do your *particular-case intuitions* play in the method of Reflective Equilibrium, according to the Private Moral Views account? The idea is that you have acquired – e.g. through your upbringing, becoming socialized, and the like – all sorts of substantive beliefs about, for instance, what things are good for a person. These substantive beliefs are revealed by your intuitive judgments about whether various particular cases instantiate the property of being welfare enhancing. Thus on the Private Moral Views account, your particular-case intuitions provide access to the thing that substantive theories of well-being are supposed to be theories *about* (viz. your own pre-theoretical views). And this is why your particular-case intuitions about well-being count evidence for or against substantive theories of well-being.

This seems to be a fairly modest account of what moral philosophers do. Accordingly, some might be dissatisfied with the Private Moral Views account.[13] After all, if my pre-theoretical views about well-being are entirely crazy (e.g. because I have become obsessed, say, with seeking honor), then nobody else besides me would have much interest in any theories that purport to systematize these views. So the Private Moral Views account seems to allow one's intuitions to count as evidence only at the expense of making philosophical theories be about something of potentially very limited concern. One might think that the subject matter of moral theories should be taken to be something of broader concern than merely one's own possibly idiosyncratic beliefs.

1.3.3 Systematizing Public Moral Views

The next strategy for defending the method of Reflective Equilibrium in ethics is in many ways similar to the previous strategy, but it is intended to avoid the drawbacks of that strategy. On this account, which I'll call the Public Moral Views account, theories in normative ethics are not supposed to be just a systematization of your *own* (possibly

[13] For a good discussion of such problems as they come up for an analogous account of the methodology typically used in the field of epistemology, see Kornblith, 2002, p. 9.

idiosyncratic) pre-theoretical moral beliefs, but rather of everybody's pre-theoretical moral beliefs – or at least everybody in the relevant community. Our intuitive judgments about particular cases are going count as evidence, on this proposal, again because they *express* the thing that normative theories are theories about, namely the shared moral views of our community.

Something akin to this view has been defended, among others, by Frank Jackson.[14] Jackson himself describes his view by saying that moral philosophers are in the business of 'analyzing our shared moral concepts',[15] but this immediately raises many difficult and controversial issues having to do with what concepts are and what an analysis is. We needn't get into these problems here, however. For the core of Jackson's view is actually quite simple:

> What we are seeking to address is whether free action according to our ordinary conception, or something suitably close to our ordinary conception, exists and is compatible with determinism. (…) But how should we identify our ordinary conception? The only possible answer, I think, is by appeal to what seems to us most obvious and central about free action, determinism… as revealed by our intuitions about possible cases. (…) Thus my intuitions about possible cases reveal my theory of free action… Likewise, your intuitions reveal your theory. To the extent that our intuitions coincide, they reveal our shared theory. To the extent that our intuitions coincide with those of the folk, they reveal the folk theory. (Jackson, 1998, pp. 31-32.)

So Jackson's idea, as applied to ethics, seems to be roughly this. Whenever there is some important moral property, P, that the members of a given community, C, have thought a fair amount about, the people in C are going to have some beliefs about what things instantiate P. In all likelihood, these beliefs are not going to be very precise, coherent or well organized. They will be messy and full of contradictions. So it is the job of the moral philosopher to systematize, in the best way possible, the jumble of conflicting moral views that people in the relevant community possess. The result is going to be a theory saying what it is that makes something instantiate P, according to the shared moral views of the people in C.

What, then, is the role of our particular-case moral intuitions on the Public Moral Views account? Contrary to the picture offered by the Direct Awareness view, one's

[14] Cf. Jackson, 1998.

[15] More specifically, Jackson thinks that the job of a moral philosopher is to determine the 'folk theory' that specifies when it is appropriate to apply certain moral concepts that are shared by 'the folk'. Jackson calls this activity, which moral philosophers are supposed to be engaged in, *conceptual analysis*. Some philosophers, however, think this should not be called conceptual analysis. These philosophers hold that analyzing the concept of F-ness just is figuring out the meaning of the word 'F'.

particular-case intuitions are not caused by one's grasp of any universal. Instead, one's moral intuitions express one's pre-theoretical moral beliefs. Thus they reveal part of what it is that normative ethical theories are supposed to be theories *about*. This is why they carry epistemic weight, on the Public Moral Views account. In particular, the particular-case moral intuitions of any given individual in the relevant community are going to carry epistemic weight to the extent that they are representative of the dominant moral views of that community as a whole.

The Public Moral Views account seems to go some way towards avoiding the sort of problem that threatened the Private Moral Views account. After all, what philosophical theories are *about*, on the Public Moral Views account, are the pre-theoretical moral beliefs that are shared by the members of a given community. So the proper subject matter of a normative theory is always going to be something that enjoys widespread acceptance. Thus such theories will be about something that is of interest to many.[16] Accordingly, while the Private Moral Views account allowed that the subject matter of normative theories might be something idiosyncratic and uninteresting, the Public Moral Views account ensures that the subject matter of normative theories is typically going to be something of broader interest.

Nonetheless, there might still be reason to be dissatisfied with the Public Moral Views account. Your and my particular-case intuitions about well-being are going to count as evidence for or against philosophical theories of well-being to the extent that they are representative of the views that are dominant in our community. After all, these dominant moral views just are what our theories of well-being are supposed to capture, according to the Public Moral Views account. Nonetheless, we might want something *more* from our theories of well-being than just a coherent systematization of the messy majority opinions of the people in our community. In particular, we might want our normative theories to tell us something about how the world is independently of our beliefs about it.[17]

For this reason, many philosophers might not be very interested in discerning the nature of *our* pre-theoretical views about moral phenomena like moral rightness or well-

[16] Barring, of course, counterfactual situations in which the community in question contains just one or a very small number of members.
[17] For similar arguments, see Kornblith, 2002, p. 10

being at all. After all, they might wonder why our views about rightness and well-being are likely to be an accurate representation of moral rightness and well-being *in themselves* (i.e. of some phenomenon that exists in the world, independently of our beliefs about it). But if the Public Moral Views account is correct, then even the best philosophical theories of well-being would not purport to be about moral rightness or well-being *in themselves*. Thus such theories would not be about the thing that many philosophers are in fact most interested in.[18]

Nonetheless, I am willing to admit that this source of dissatisfaction with the Public Morality account can in principal be answered. In particular, if we had some independent reason to think that our shared pre-theoretical views about well-being in fact latch onto the phenomena of rightness and well-being *themselves*, then philosophical theories as they are construed by the Public Moral Views account would indeed help provide answers to the question that philosophers are most interested in, viz. what the natures of rightness and well-being in themselves are. In other words, if we had reason to think that our pre-theoretical views about well-being correspond to the *truth* about rightness and well-being, then the Public Moral Views account might not end up looking so bad. The final proposal I will discuss can be seen as providing a reason of just this sort.

1.3.4 Systematizing Human Beings' Moral Emotions – The Tracking Account

The final account of what we might be doing when we use the method of Reflective Equilibrium in normative ethics I'll call the Tracking Account. On this account, a theory arrived at via the method of Reflective Equilibrium is not supposed to be a systematization of the pre-theoretical moral *beliefs* of any given community, but rather of the moral emotions or gut feelings that human beings in general tend to have in virtue of the way we are psychologically constructed. People's particular-case intuitions about, say, moral rightness or well-being count as evidence for or against normative ethical theories because these intuitions are expressions of the gut feelings that these theories purport to capture. A moral theory arrived at by the method of Reflective Equilibrium is supposed to tell us what natural property human beings' gut feelings tend to track.

[18] What's more, we might want to allow for the possibility that the moral views that happen to be widespread in our community are actually mostly *incorrect*. This seems to be impossible on the Public Moral Views account. So we might have yet another reason to be dissatisfied with it.

The Tracking Account is similar in spirit to many naturalistic accounts of morality (whether realist or anti-realist).[19] I do not want to get bogged down by discussing the wide variety of views of this sort that have been proposed, or by the complex debate about whether or not any such view can ultimately count as a form of realism. Thus I will simply present what I take to be a generic, fairly uncontroversial[20] form of the view. While it is loosely based on the ideas of others, whatever defects it ends up having are my own fault. The proposal is easiest to understand when applied to normative ethics, so I begin by sketching the proposal as it applies to this area of inquiry. Then I explain how it is to be carried over to the topic of well-being.

1.3.4.1 The Tracking Account in ethics

Normative ethics is the attempt to discover, formulate and defend the correct criterion of moral rightness. But what could make various criteria of moral rightness correct or incorrect? In other words, what are theories in ethics supposed to be theories *about*? To see the Tracking Account's answer to this question, start by noting that human beings often have responses towards concrete actions. A particular action is described to you in complete detail and you approve of it; some other action is described to you in complete detail and you disapprove of it. We can call such a response to a particular, fully described act token a 'moral intuition' (though this departs from the standard notion of a philosophical intuition[21]). Thus a moral intuition, on this proposal, is a gut feeling of approval or disapproval that a person has towards a concrete action, performed by a particular person, on a particular occasion, with a determinate set of consequences.

Two further assumptions are required. First, let's assume that these intuitive responses – in a very messy, rough and imperfect way – track *some* complex natural property or other. Call that property 'N.' The psychological machinery that makes human beings respond with approval or disapproval to actions can, for the most part, tell when N

[19] Cf. Boyd, 1988; Brink, 1989 (esp. ch. 1 & 2.); Gibbard, 1990.

[20] I suggest that it is uncontroversial because I try to present the view in a way that is supposed to be neutral with respect to both realism and anti-realism. Some might defend the claim that the Tracking Account counts as a realist view, while others might deny this. I will not take a stand on this question here.

[21] This is not what people like Bealer typically mean by 'an intuition.' They mean something like 'a non-inferrential intellectual seeming'. However, it seems fairly reasonable to regard intuitions *at least in the moral realm* as feelings of approval or disapproval towards concrete actions. This is what the present proposal assumes, anyway.

is instantiated and when it's not. Thus people's responses to actions are (in some imperfect way) responsive to N, in the sense that actions instantiating N would tend to be approved of, while actions lacking N would tend to be disapproved of. Perhaps N is the property of utility maximization; perhaps it is the property of not treating anybody as a mere means; perhaps it is something else entirely. Whatever it is, the assumption is just that there is *some* natural property, N, that people's responses (no matter what society these people live in, no matter what upbringing they have, etc.) towards actions tend to track, or be responsive to.[22] Second, let's assume that the moral facts (the facts about what actions are right or wrong) are made true or false by the facts about whether or not this property, N, is instantiated. More precisely, let's assume that an action is morally right iff it instantiates N. This is simply a view about what the truth-makers of moral claims are.[23]

Given these two assumptions, then, we may characterize the project of normative ethics as that of figuring out what this property N *is*.[24] Thus what the Tracking Account proposes is that the target phenomenon that ethicists are (or should be) interested in investigating is fixed by people's moral intuitions, i.e. by their attitudes of approval and

[22] I think most people would accept this first assumption. Denying it seems implausible. Someone who denied it would be committed to thinking that our moral intuitions are not latching on to any natural property. But that seems wrong. There are natural properties that our moral intuitions are responsive to. Setting the cat on fire just for fun is clearly wrong, while setting it on fire in order to prevent a nuclear holocaust isn't. There is a difference in natural properties between these two cases that grounds our differing intuitive evaluations of these two cases. In other cases, our moral intuitions might be tracking some other natural properties. Presumably, the various natural properties that our moral intuitions track are related in some interesting way, or have something important in common. Thus the assumption here is that there is some basic natural property that our moral intuitions are fundamentally responsive to. This is property N.

[23] I want to emphasize that the assumption being made here is not that the property of moral rightness *just is* the natural property N. It's not that an action's being morally right is *identical* to or *consists in* this action's instantiating the natural property, N, which human beings' responses to actions tend to track. Although some naturalistic philosophers might hold such a view (e.g. Boyd, Brink, Sturgeon), this is not what I am assuming here. All I am assuming is that an action's being morally right *supervenes* on whatever the natural property, N, is that human beings' responses towards actions tend to track. Thus even someone who thinks that moral rightness is a non-natural property could in principle accept the assumption I am making here.

[24] On the Tracking Account, this property N cannot be taken to stand for simply 'the property that is tracked by human beings' moral emotions' or something of the sort. For this is too general to yield an informative criterion of moral rightness. After all, such a criterion is supposed to be a *non-trivial* statement of the form 'An action, a, is morally right iff _____.' But if the Tracking Account is true, then the claim 'An action, a, is morally right iff a instantiates the property that is tracked by human beings' moral emotions' is going to be trivial. Thus this claim cannot be a candidate criterion of moral rightness, if the Tracking Account is true.

disapproval towards particular actions. A normative ethical theory is supposed to be a theory that states what the natural property is that people's moral intuitions are *tracking*.[25]

1.3.4.2 The Tracking Account as applied to well-being

So far, I have merely described the Tracking Account as it applies to normative ethics. But a similar account can be given when it comes to well-being. Here, roughly, is how this would go. Theories of well-being purport to tell us what makes something in itself enhance a person's well-being. What makes theories of this sort be correct or incorrect? It is our responses to various things, according to the Tracking Account. If some scenario involving a given person is described to you in complete detail, then you might have some feeling about whether or not this scenario is good for that person. Alternatively, if a pair of possible scenarios involving a person are described to you in complete detail, then you might have some feeling about which of them is *better* for the person in question. These are examples of what I'll call 'intuitions about well-being.' They are supposed to be gut feelings, pre-theoretical responses to possible cases.[26] Next, the Tracking Account assumes that there is some general and complex natural property, N*, that people's intuitions about well-being are tracking or are responsive to (no matter what society the people come from, what kind of upbringing they have, etc.). And finally, the Tracking Account assumes that claims about well-being are made true or false by the facts about whether or not this property N* is instantiated. More specifically, the assumption is that something enhances a person's well-being iff it instantiates N*. This is a view about what well-being supervenes on, about what the truth-makers for well-being claims are.

Given these assumptions, then, the Tracking Account says that the project we are engaged in when theorizing about well-being is the project of figuring out what this property N* *is*.[27] The target property that we're interested in finding is picked out by

[25] Note that the moral intuitions that are supposed to fix the target of normative ethical theories are not just the intuitions of people living at any particular time, or in any particular society. Rather, the target is fixed by the intuitions of people *in general*, no matter what society they are from or what cultural context they live in.

[26] What are they responses *to*, exactly? This is a question I will address in section 1.4. There I discuss the hard question of what our 'intuitive judgments about well-being' really are judgments *about*.

[27] As before, on the Tracking Account, this property N* cannot simply stand for 'the property that is tracked by human beings' gut feelings about what's good for a person' or something of the sort. For this is

people's intuitions about well-being, which are conceived of as gut feelings or intuitive responses. A philosophical theory about well-being is supposed to be a theory that states what the natural property is that people's intuitions about well-being are *tracking*.

1.3.4.3 Advantages and Disadvantages

I think the Tracking Account, when it comes to theorizing both about moral rightness and well-being, has some attractive features. For one thing, it provides a neat answer to the question of why our intuitive judgments about rightness and well-being count as evidence for or against philosophical theories of these topics. In particular, such intuitions simply pick out the phenomenon that such theories are trying to capture. Thus these intuitions are the raw data that such theories are trying to systematize.

What's more, the Tracking Account seems to offer an explanation of why it is not a waste of time to systematize *our* beliefs about moral rightness and well-being. On the Tracking Account, investigating our beliefs about moral rightness and well-being would help us discern the nature of moral rightness and well-being *in themselves* (specifically, what the supervenience base is for these properties). After all, our beliefs about moral rightness and well-being surely must have been heavily influenced by the psychological machinery that produces the various gut feelings and responses that pick out the supervenience base of moral properties. Thus on the Tracking Account, we could make progress towards learning when rightness or well-being *themselves* are instantiated by figuring out the conditions under which we would be inclined to call various scenarios right or welfare-enhancing. So by systematizing our various beliefs about moral rightness or well-being, we would arguably be learning something about the nature of moral rightness or well-being in themselves (specifically, what makes them be instantiated). As a result, there might be reason to think that the Tracking Account has more going for it than the Public Moral Views account. That account, after all, did not obviously provide any independent reason to think we can learn about moral rightness and well-being in

too general to yield an informative criterion of goodness for a person. After all, such a criterion is supposed to be a *non-trivial* statement of the form 'X is good for a person P iff _____.' But if the Tracking Account is true, then the claim 'X is good for a person P iff X instantiates the property that is tracked by human beings' gut feelings about what's good for a person' is going to be *trivial*. Thus this claim cannot be a candidate criterion of goodness for a person, if the Tracking Account is true.

themselves by systematizing the moral views that happen to be widespread in our community.[28]

However, there might seem to be a major drawback of the present proposal. In particular, one might think that normativity is lost, on the Tracking Account. The proposal construes a theory in normative ethics as a theory about what complex natural property is being tracked by human beings' attitudes of approval or disapproval towards particular actions. Similarly, a theory about well-being is a theory about what complex natural property is being tracked by human beings' gut feelings about what's good for one. But it might seem that such theories are not *normative* theories at all. Instead, they seem merely to purport to describe certain features of human psychology. Accordingly, these theories do not say what a person *ought* to pursue from the perspective of morality or from the perspective of well-being.

Perhaps a proponent of the Tracking Account can respond to this objection.[29] But the issue is very difficult and I cannot pursue the matter any further here. So I conclude simply by pointing out that also the defender of the Tracking Account has some challenges to meet before he can claim victory.

1.3.5 Conclusions

In this section, I have discussed four different accounts of the method of Reflective Equilibrium in moral philosophy. Each one suggested an explanation of why our moral intuitions are likely to count as evidence, and in what sense theories arrived at by the

[28] Of course, the Public Moral Views account and the Tracking Account – at least as I have described them here – are not mutually incompatible. There could a hybrid view on which there is a place both for systematizing public moral views and trying to find out what natural property our gut reactions are tracking. I won't discuss this possible view separately, however.

[29] In particular, the defender of the Tracking Account can perhaps respond as follows. Suppose we have discovered what the property N is that human beings' attitudes of approval and disapproval towards actions are tracking. Since you and I are human beings, we too will tend to approve of actions that instantiate N and disapprove of actions that fail to instantiate N. Thus the facts about whether N is instantiated will tend to *seem* to us to be normative facts. And perhaps this is good enough when it comes to capturing normativity. Similarly, when it comes to well-being. Suppose we have discovered the property N* that is being tracked by human beings' gut feelings about what's good for a person. Since you and I are human beings, we too will tend to think that things that possess N* are good for us and that things that lack N* are bad for us. Thus the facts about whether N* is instantiated will tend to *seem* to us to be normative facts. And again, perhaps this is good enough when it comes to capturing normativity.

Is that enough to allow the Tracking Account to count as a form of realism, however? That question is too hard for me to answer here. (See Sayre-McCord, 1986 and Joyce, 2006, ch. 5, for arguments that views like the Tracking Account cannot count as forms of moral realism.)

method of Reflective Equilibrium are supposed to be true. However, each proposal had at least some problems that it must deal with. But perhaps some of these challenges can be overcome. Because of space limitations, however, I could not fully canvas the arguments for and against each proposal. Thus I am not in a position to decide between them. It is really a job for the metaethicist to tell us which account of the method of Reflective Equilibrium is the one we must adopt.

Nonetheless, I hope the preceding discussion will be sufficient to show that there are many avenues open to someone (like me) who wants to justify his use of the method of Reflective Equilibrium in moral philosophy. I think we have seen four plausible (though perhaps incompatible) ways in which one might try to explain why our intuitions about well-being count as evidence for or against philosophical theories about well-being.

1.4 A Methodological Problem for Theorizing about Welfare in Particular: What Distinguishes Welfare Value from other Types of Value?

In the chapters that follow, then, I will use the method of Reflective Equilibrium to evaluate various theories that purport to capture our intuitive judgments about well-being. Thus I will be looking for the theory of well-being that, to use Sumner's term, is most *descriptively adequate*, i.e. 'the one which is most faithful to our ordinary concept and our ordinary experience.' (Sumner, 1996 p. 11)

However, there is one major remaining problem for the method of Reflective Equilibrium when it comes to theorizing specifically about *well-being*. A life is said to be high in well-being – or, what amounts to the same thing, high in *welfare value* – when the life goes well for the one who leads it. But there are many other kinds of value a life might have as well: moral value, aesthetic value, value for society, and so on. So what distinguishes welfare value from the other kinds of value a life might have? We cannot use the method of Reflective Equilibrium in theorizing about well-being (or welfare) until we have an answer to this question. For we cannot test a theory of well-being against our intuitive judgments unless we know precisely what these judgments are supposed to be judgments *about*. In other words, we need to get clear on what the concept of welfare value *is* before we will be in a position to evaluate theories of welfare by appeal to our intuitions about that concept.

The trouble is that it turns out to be a very difficult task to say exactly what makes welfare value different from other kinds of value. There is little chance of our being able to get clear on what the concept of welfare is simply by appealing to our everyday usage of the words 'welfare' or 'well-being.' As Griffin points out, our job in investigating the nature of welfare cannot simply be

> to describe the everyday use [of the word 'welfare']. It is too shadowy and incomplete for that… (Griffin, 1986, p. 7)

There are, however, other more promising strategies on offer for distinguishing welfare from the other sorts of value. In the remainder of this chapter, I will show why five initially promising strategies for pinning down what is distinctive of the concept of welfare fail. I end by offering my own proposal about what distinguishes welfare value from other kinds of value.

1.4.1 The 'good for' strategy

Roger Crisp explains the first strategy for specifying what is distinctive of welfare value:

> Well-being is a kind of value, sometimes called 'prudential value', to be distinguished from, for example, aesthetic value or moral value. What marks it out is the notion of 'good for'. The serenity of a Vermeer painting, for example, is a kind of goodness, but it is not 'good for' the painting. (Crisp, 2008)

Crisp's suggestion seems to be that welfare value just is being *good for* something. More precisely, the suggestion is this:

(A) X is welfare-valuable iff there is something such that X is good for it.

This proposal about what distinguishes welfare from other kinds of value, however, immediately faces a problem. If (A) were correct, we would be committed to saying that cars, for instance, may have levels of well-being. After all, we say that it's good for your car that you change its oil regularly. (A) entails that changing the oil in your care is welfare-valuable. But that is not a plausible result. When we say that it's good for your car to change the oil in it, we are making a claim about what is required for the car to continue properly performing the function it was designed for. We clearly don't mean that changing the oil in your care increases its welfare level. Cars, as inanimate objects, seem not to have welfare levels at all.

Sumner offers a refined version of Crisp's proposal about what is distinctive of welfare value, and his refined proposal avoids the present problem. Sumner says:

> What distinguishes welfare from all other modes of value is its reference to the proprietor of the life in question: although your life may be going well in many respects, it is prudentially valuable only if it is going well *for you*. This subject-relativity is an essential feature of our ordinary concept of welfare. (…) We have already established that no theory about the nature of welfare can be faithful to our ordinary concept unless it preserves its subject-relative or perspectival character. (Sumner, 1996, p. 42)

Sumner's suggestion seems to be that welfare value just is being good for some living creature. Thus what makes welfare value different from the other kinds of value is that things that increase the welfare-value of a creature's life are *good for* the creature whose life it is. To put it more precisely:

> (A') X is welfare-valuable iff there is some living creature such that X is good for that creature.

The refined proposal faces serious problems, however. We must distinguish between two senses of 'good for a creature.' The phrase might mean either 'appears good to a creature' or 'is a benefit to a creature.' But there is a problem in either case. If (A') is interpreted using the first sense of 'good for a creature,' then objective theories of well-being are going to be ruled out from the get go. Sumner characterizes objective theories of welfare as follows: 'On an objective theory, therefore, something can be (directly and immediately) good for me though I do not regard it favourably, and my life can be going well despite my failing to have any positive attitude toward it.'(Sumner, 1996, p. 38) If what is distinctive of welfare value is that it is the sort of value that attaches to things that appear to be good to a creature, then no substantive theory on which something can increase one's welfare without one's regarding it favorably could be true. Thus if the phrase 'good for' in (A') is taken to mean 'appears good to', no first-order theory of welfare that is objective in Sumner's sense could be true. But this is a big problem, since many philosophers want to defend objective theories of just this sort.

So it seems 'good for' should be understood in the second way mentioned. On this interpretation, being good for a creature is benefiting it, or having a positive impact on it. However, if (A') is understood in this way, a new problem threatens. In particular, it does not distinguish welfare value from the other kinds of value. The reason is that things can be good for a creature in the sense of benefiting it (or having a positive impact on it) in many other ways besides the welfare way. For instance, adding aesthetic value or moral value to a creature's life would also seem to have some positive impact on that creature. To add aesthetic value to a person's life would be to make it more beautiful, and adding this kind of value to his life is to benefit him aesthetically. Similarly, to add moral value to a person's life is to increase the moral worth of her life, and thus we can say that adding this kind of value to her life morally benefits her. Thus even if it is true that in order for X to increase a creature's welfare, X must have a positive impact on that creature, this cannot be what is distinctive of welfare value. A creature's getting the other kinds of value can be *good for* that creature too in various ways. Accordingly, I take it that this first strategy for distinguishing welfare value from the other kinds of value fails.

1.4.2 Hooker's Sympathy Test

Brad Hooker proposes a different way of distinguishing welfare value. His proposal is apparent in the following two comments:

> How sorry we feel for someone is influenced by how badly from the point of view of his own good we think that person's life has gone, that is, by whether we think his life has lacked important prudential goods. (…) if two people's lives have contained the same amounts of pleasure, knowledge, and autonomy, but one has contained significantly more achievement than the other, we feel sorrier for the person whose life has contained less achievement. (Hooker, 1996, 149)
>
> If (1) two people are as much alike as possible except that one's life contains something which the other's does not and (2) we do not feel sorrier for the one whose life lacks this thing, then the explanation is that we do not really think this thing is one of the fundamental categories of prudential value. (Hooker, 1996, p. 150)

Here is the test that Hooker's comments suggest. Consider the lives of S and S*. Suppose these lives are as much alike as possible except that S's life contains X, while S*'s life does not. If we feel sorrier for S* than we do for S, then X benefits a person's welfare. More specifically, we may state Hooker's proposal as follows:

> (B) X positively impacts a person's welfare iff the following conditional is true of X: if there were two people, S and S*, whose lives are as much alike as possible except that S's life contains X while S*'s doesn't, then we would feel sorrier for S* than we would for S.

On this proposal, what distinguishes welfare value from the other sorts of value is that it is the sort of value which is such that we feel sorry for people who lack it.

This proposal has problems as well. For one thing, what if there is no consensus about who, of a given pair of people, to feel sorrier for? We are likely to find many cases like this in which there are two people, S and S*, where some people would feel sorrier for person S, while others would feel sorrier for person S*. Whose feelings get to set the bar for what enhances welfare? It isn't clear. However, Hooker's proposal faces more fundamental problems than this. There are two in particular. While the first might be surmountable, the second is not.

The first problem may be illustrated by an example. Suppose S and S* lead virtually identical lives, and both lives are really great ones. Both S and S* are very happy; they both have meaningful relationships, are successful, experience a lot of pleasure and no pain, and on the whole are extremely content. The only difference between the two lives, in fact, is that S's life contains a certain event in which S gets a certain moderate amount of pleasure from sitting in a hot tub for half an hour, while S*'s life does not contain any such event. Pleasure is a clear example of something that positively impacts welfare. Accordingly, if Hooker's test were correct, we would expect to feel sorrier for S* than we do for S. However, we don't. It seems we would feel sorry neither for S* nor for S. After all, both have outstanding lives. Thus we have another sort of counter-example to Hooker's claim about what distinguishes welfare. For in this case, we have an example of something that clearly has a positive impact on welfare (i.e. a certain episode of pleasure), even though we do not feel sorry for someone who lacks this thing. Thus (B) would be false.[30]

Perhaps this counter-example to (B) can be avoided.[31] Assuming there is a numerical scale that represents the degree to which we feel sorry for a person, it would be natural to assume that the same numerical scale can be extended to encompass whatever the opposite of feeling sorry for a person is –envy, perhaps.[32] Let's call this *the pity-envy*

[30] Another counter-example in the same spirit would involve two people who are equally horrible human beings, but where one of them is very happy and the other is not. We would not feel sorry for either one of these people. After all, they are just plain evil. The only thing we feel towards them is indignation. Since it's not the case that we feel sorrier for the unhappy person, Hooker's test would imply that happiness is not welfare-enhancing. However, that is obviously a mistake. Happiness is a paradigmatic example of something that is welfare enhancing.

[31] Many thanks to Fred Feldman for suggesting this response to me.

[32] I'm not committed to the idea that feeling envy towards someone is the opposite of feeling sorry for someone. Whatever the opposite is, though, that's what goes into this numerical scale.

scale. People who we feel sorry for on balance receive negative numbers on the pity-envy scale, while people who we envy on balance receive positive scores. To find a person's score on this scale, we subtract the degree to which we feel sorry for the person from the degree to which we envy the person. Using this pity-envy scale, then, we can modify (B) in such a way as to avoid my hot tub counter-example:

> (B') X positively impacts a person's welfare iff the following conditional is true of X: if there were two people, S and S*, whose lives are as much alike as possible except that S's life contains X while S*'s doesn't, then S falls higher on the pity-envy scale than S* does.

This would avoid my counter example because the person who does get to enjoy the hot tub for half an hour would indeed fall higher on the pity-scale than his doppelganger who does not get to enjoy the hot tub. Thus (B') – unlike (B) – does not have the counter-intuitive result in the case of the hot tub that pleasure fails to be welfare enhancing. So no counter-example.

Even if this response to my first counter-example to Hooker's proposal succeeds, the proposal faces another, more serious counter-example – this time in the opposite direction. In particular, there are examples that make it clear that even if we would all feel sorrier for someone who lacks a given thing than we would for someone who has that thing, this thing still might not be welfare-enhancing. Suppose we are all zealous supporters of *The Party*. Our whole belief system centers around the idea that the most praiseworthy thing a person can do is to sacrifice themselves and everything that is good for them for the sake of The Party. Now suppose we consider two twins: one who makes the sacrifice and another who simply can't go through with the sacrifice. Since we are such passionately devoted Party members, we feel sorrier for the person who didn't make the sacrifice. Thus (B) entails that making the sacrifice would have a positive impact on one's welfare. What's more, (B') has the same implication. As devoted Party members, we would all rank the twin who *did* manage to go through with the sacrifice higher on the pity-envy scale than the twin who did not manage to do so. Thus on both (B) and (B'), making the sacrifice would have a positive impact on one's welfare.

However, this is obviously not the correct result. After all, the sacrifice in question here *cannot* promote one's welfare, since the sacrifice was stipulated to involve forgoing everything that is good for one. As a result, we have a case that shows that our feeling sorrier for (i.e. giving a lower score on the pity-envy scale to) a person who lacks a given thing does not guarantee that this thing is welfare enhancing. And so we have a counter-example to (B) and (B') alike. The general problem here is that the degree to which we feel sorry for or envy people can in fact be influenced by all sorts of other things than just how much welfare people enjoy. As a result of counter-examples like this, I take it that Hooker's proposal does not succeed in specifying what distinguishes welfare value from the other sorts of value.

1.4.3 Darwall's Rational Care Test

Stephen Darwall proposes that what is distinctive of welfare is that it is the sort of value that we wish for people about whom we care. Here is how he puts his idea:

> what it is for something to be good for someone *just is* for it to be something one should desire for him for his sake, that is, insofar as one cares for him. (…) what it is for something to be good for someone is for it to be something that is rational (makes sense, is warranted or justified) to desire for him insofar as one cares about him. (Darwall, 2002, pp. 8-9)

So Darwall seems to think that something positively impacts a person's welfare just in case it would be rational to desire that thing for a person insofar as one cares about him or her. To put it more precisely:

> (C) X positively impacts a person's welfare iff one's caring for a person would make one desire X for that person (provided one is rational).

On this proposal, when we say that something positively impacts your welfare, we mean that it's good for you in just the same way that what people who care about you would want for you is good for you. Things that have other kinds of value (moral value, aesthetic value, medical value, etc.) would not be good for you in this way. And this is what distinguishes welfare value from the other kinds of value.

Fred Feldman has argued (successfully, I think) that Darwall's idea about what is distinctive of welfare value is mistaken. Feldman (2004, pp.9-10) presents Darwall's idea in terms of a thought experiment in which a parent is looking down into the crib of his new-born baby. With nothing but love in his heart for the child, the parent wishes that his child will have a good life. Darwall's idea, then, may be understood as the thought that a life high in welfare just is what the parent in this thought experiment is wishing for his child. Feldman goes on to point out the problem with this idea about what is distinctive of welfare as follows:

> It is not entirely clear that this thought experiment will always work. Suppose a religious fanatic looks into his child's crib. Suppose he wants the child to have a wonderful life. Suppose he thinks that the best imaginable life for the child is one in which the child becomes a martyr for God. This religious fanatic might be filled with love, and he might be thinking about the Good Life for his child. But it is not clear that he is expressing a hope about what we would normally think of as the child's *welfare*. (Feldman, 2004, p. 10)

The upshot is that the idea stated in (C) does not specify what is distinctive of welfare. Darwall's proposal does not reveal what makes welfare different from the other kinds of value. Since there is in principle no bounds to what we might desire for a person we care about, *anything* – even things that obviously do not promote one's welfare – might count as having a positive impact on one's welfare. Things that are aesthetically, or religiously or morally valuable could easily qualify as possessing welfare value according to (C). Thus Darwall's proposal about what is distinctive of welfare is mistaken.[33]

It might perhaps be possible to construct a better version of Darwall's test. Suppose there is a Greek god who has a human child and the god loves this child above all else. The god wants to give his child the best possible life a human being can get. What kind of life would the god want to grant his child? Let's stipulate that the god has no concern for any other living creature, and so there are no moral constraints on what sort of life the god would desire for his child. Similarly, the god is entirely uninterested in aesthetic matters, and so he has no independent desire to make his child's life be a beautiful or fascinating one. Nor does the god have any religious agenda, political goal, or other cause that he might want the child's life to serve. In short, the god has a single-minded devotion to his child, and this makes the god desire that his child will do or experience whatever would be required for leading the best possible life for the child, no matter what the effects might be on anybody or anything else. Accordingly, we might take it that when we say that X promotes a person's welfare, what we mean is that X is good for a person in the way that what the Greek god in this story wishes for his child would be good for the child.

The original version of the Rational Care test seemed implausible because people who care about you might desire all sorts of things for you (such as a martyr's death) that

[33] Also see Feldman, 2006 (Available online.)

clearly would not promote your welfare. The revised version of the Rational Care Test, however, might avoid this sort of problem. After all, it was stipulated that the Greek god has no desire that his child should have a life that is high in moral value, aesthetic value, religious fervor, and so on. Nonetheless, the revised version of the Rational Care Test still seems unsatisfying. It attempts to pinpoint welfare value simply by building features into the story whose sole purpose is to exclude the other sorts of value. The god was stipulated not to desire that his child's life possess any special moral value, aesthetic value, religious piety or general usefulness. Making these stipulations amounts to rigging the thought-experiment to guarantee that the god will not desire anything but the child's welfare. Thus the revised version of the test seems to be ad hoc.

The idea of revising the test in this way, however, does begin suggest another way to pin down what is distinctive of welfare. In particular, rigging the test to exclude the other kinds of value suggests that we might be able to specify what welfare is simply by appeal to what it is not. Let's go on to investigate this idea directly.

1.4.4 Feldman's strategy of appealing to what welfare is not

The strategy of pinning down what is distinctive of welfare value by appeal to what welfare is *not* is the very strategy that Feldman goes on to use after rejecting Darwall's test. Here is how Feldman puts it:

> let us distinguish among several different things that we might have in mind when we ask whether someone has a good life. A. When we speak of a good life, we might mean a morally god life… B.When we speak of a good life, we might use 'good' in a sense in which it means 'good as a means'… C. Another sort of good life would be the beautiful life. We might want to know what makes a persons life *aesthetically* good. (…) D. Someone might take the question about the Good Life to be equivalent to a question about what sort of life best exemplifies human life. (…) E. Finally, we come to the sense of the phrase that is relevant here. Sometimes, when we speak of the Good Life, we have in mind the concept of a life that is good in itself for the one who lives it. (Feldman, 2004, pp. 8-9)

The suggestion that is implicit in this passage seems to be roughly this: if X is good for some person, P, but X is not extrinsically good for P, morally good for P, good for P merely as a means, aesthetically good for P, or good for P in the perfectionist sense, then X would in itself have a positive impact on P's welfare.[34] Two brief comments on how to improve this suggestion. First, I think this list can be expanded to include some other

[34] For another (somewhat less explicit) attempt to employ this strategy for distinguishing welfare from other kinds of value, see Brink, 1989, p. 218.

kinds value that we want to distinguish welfare value from. For example, something might be good for a person's health, good for a person's social life, and good for a person's finances. There might be other sorts of value that should be included here as well. Second, perfectionist value should not be included on the list, I think. The reason is that this would make the present proposal about what is distinctive of welfare rule out on conceptual grounds first-order theories that say perfection is what makes for welfare. To include perfectionist value on the list here would beg the question against perfectionist first-order theories of welfare.

Perhaps, after making these two slight modifications, we would want to state the present proposal as follows:

> (D) X would (in itself) have a positive impact on the welfare of some person P iff X would be good for P, but X would not be instrumentally, morally, aesthetically, medically, socially, economically (etc.) good for P.

However, it would be a mistake to understand the present proposal as giving both necessary and sufficient conditions for something's having welfare value. After all, things will often in themselves promote people's welfare, while at the same time having other kinds of value as well. Thus it can't be a necessary condition on something's having welfare value that it not have any other kind of value. That would be silly. Accordingly, I think it would be more plausible to take the proposal to be giving just the following sufficient condition:

> (D') If X would be good for some person, P, but X would not be morally, instrumentally, aesthetically, medically, socially, economically (etc.) good for P, then X would (in itself) have a positive impact on P's welfare.

There are a number of reasons to be dissatisfied with this proposal. For one thing, the list of kinds of value that welfare is to be distinguished from is clearly not complete. And even if one continues to add other sorts of value to the list, it might not be easy to tell when the list finally does become complete. Second, it seems that one requirement on a good analysis is that it analyzes the target concept in terms of other concepts that we have a better grasp on than the target concept. But that does not seem to be the case here. The target concept here is the concept of welfare value, and it is analyzed in terms of other sorts of value, such as moral value, aesthetic value, etc. However, it is unlikely that we have a better grasp of moral value, for instance, than we do of welfare value.[35] Thus (D) does not meet this requirement on being a good analysis.

The most important problem with the present proposal, however, is that it is not informative. We want our explanation of what is distinctive of welfare value to help us see *why* welfare is different from the other kinds of value. However, while the present proposal, (D'), tells us that there is *some* such difference, it does not tell us what that difference is. *Why* are these other kinds of value included on the list of things that welfare is to be distinguished from? In virtue of *what* is welfare value different from these other kinds of value? This present proposal does not provide an answer. It offers no explanation of why welfare value is not the same as moral value, or aesthetic value, and so on. Presumably the reason is that there is some positive characteristic of welfare value that these other kinds of value lack. But what is it? Since the present proposal does not tell us, it is uninformative.

1.4.5 Feldman's 'conceptual role' strategy

The final proposal I will argue against here is an attempt to say what the positive characteristic of welfare is that distinguishes it from the other kinds of value. In particular, the proposal is that welfare value is the sort of thing that plays a particular conceptual role, and this makes welfare value different from the other kinds of value in that they cannot play this role. In a forthcoming work, Feldman gives the following description of this strategy, which I'll call the Conceptual Role Strategy:

> Welfare: the sort of value that is necessarily decreased when a person is harmed; (…) the sort of value that we increase in a person when we benefit him. Additionally, when a person is selfishly trying to enhance his own self-interest, the sort of value that he is seeking to enhance is his own welfare. Contrariwise, it is the sort of value an altruistic or benevolent person tries to enhance in others. Welfare is the value that we have in mind when we worry about someone's quality of life, or when we consider whether he has a life worth living. Welfare is the value about which we may be concerned when, at graveside, we reflect on the question whether the deceased "had a good life". (Feldman, forthcoming, ch. 8)

Feldman is suggesting here a number of different conceptual truths about welfare, and together they seem to pick out the conceptual role that the notion of welfare must be able to play:

[35] In fact, philosophers who favor Consequentialism want to understand a certain kind of moral value (i.e. the moral rightness of actions) in terms of welfare.

1) Necessarily, if a person is harmed, then his welfare is decreased.
2) Necessarily, if a person is benefited, then his welfare is increased.
3) Necessarily, if a person selfishly tries to promote his self-interest, what he is trying to enhance is his welfare.
4) Necessarily, if a person tries to act in ways that are altruistic or benevolent, then he is attempting to promote the welfare of others.
5) Necessarily, if we consider the quality of a person's life or ask whether a given person has a life worth living, then what we are considering or asking about is that person's welfare.

Accordingly, one way to formulate the Conceptual Role Strategy would be this:

> (E) X is what intrinsically enhances a person's welfare iff a suitable phrase denoting X can be substituted for 'welfare' in claims 1)-5) without making these claims become false.[36]

On this proposal, what is distinctive of welfare value is that it plays the conceptual role that claims 1)-5) pick out.[37] In order for something to count as a first-order theory of welfare, it must be a theory about whatever it is that plays the conceptual role picked out by claims 1)-5).

I think this is the most promising strategy yet for specifying what is distinctive of welfare. However, it too seems to have serious problems. In particular, because it is possible to adopt *broad conceptions* of harm, benefit, selfishness, altruism, and quality of life, it turns out that (E) does not succeed in specifying what is distinctive of welfare. To see what is meant by a 'broad conception' of harm, benefit, etc., first note that there are many different scales on which a person's life might be evaluated. In addition to evaluating a life based on how much individual welfare it contains, a life can be

[36] A stronger version of this proposal would state that X is what intrinsically enhances a person's welfare iff a suitable phrase denoting X can be substituted for 'welfare' in claims 1)-5) without *changing the meaning of these claims.* I think this intensional version the proposal is unnecessarily strong, however. The extensional version should be sufficient.

[37] I want to point out that we could construct a subtly different version of the Conceptual Role Strategy if we replaced the claims 1)-5) with similar claims in which the conditionals simply go the other way. So instead of 1), we would have the claim 1*): necessarily, if a person's welfare is decreased, then he is harmed. And instead of 2), we would have the claim 2*): necessarily, if a person's welfare is increased, then he is benefited. And so on. However, a version of the Conceptual Role Strategy that employed 1*)-5*) would seem to be substantially weaker than a version that employed 1)-5). As I point out below, there are many kinds of harm/benefit (e.g. financial) that do not necessarily entail *welfare* enhancements/reductions. By contrast, there is no denying that welfare enhancements/reductions do entail harms/benefits of a certain kind at least. Accordingly, claims 1)-5) pick out a more robust conceptual role for welfare than claims 1*)-5*) do.

However, the main reason that I formulate the Conceptual Role Strategy in terms of 1)-5) – instead of 1*)-5*) – is that these are the claims that Feldman actually makes use of in the passage that I repeated above. He is not making claims in which the conditionals go the other way.

evaluated according to how good a story it makes, how healthy a life it is, how musical it is, how religiously devout it is, how financially successful it is, how morally good it is, how useful the life is for some specific purpose (e.g. the advancement of nuclear disarmament), and so on and so forth. What, then, is a broad conception of, say, harm? The *most* broad conception of harm would be one that allows that a person may be harmed by decreasing his score on *any* of these evaluative scales. A less broad conception of harm would be one on which one may be harmed by having one's score lowered only on a certain restricted group of scales – say, all possible scales except for the financial, health, and aesthetic scales. And in general, we may say that a conception of harm is broad just in case there is some *non-welfare scale* such that having one's score decreased on that scale would count as a harm. It should be easy to see how this can be extended to the other key notions in (E), namely benefit, selfishness, altruism and quality of life. For instance, a broad conception of benefit would be one that allows that one can be benefited by having one's score increased on some non-welfare scale.

The fact that broad conceptions are available of harm, benefit, selfishness, altruism and quality of life causes problems for (E). After all, if (E) is to be capable of specifying what is distinctive of welfare, then harm, benefit, selfishness, altruism and quality of life cannot be understood broadly. If they were, (E) would allow all manner of benefits and harms that clearly are not at all relevant to one's level of welfare to count as intrinsic welfare enhancements or reductions. But this would of course be a mistake. Enhancing or reducing welfare involves not raising or lowering the score one's life receives on *just any* evaluative scale, but rather raising or lowering the degree to which one's life goes well specifically in the welfare way. Not just any increase or decrease on *some evaluative scale or other* would count as an intrinsic welfare gain or loss; only an increase or decrease in one's score specifically on the welfare scale would count.

Thus if (E) is to succeed in specifying what is distinctive of welfare value, it cannot employ a broad conception of harm, benefit, selfishness, altruism or quality of life. For by definition, broad conceptions of these notions involve an appeal to evaluative scales other than the welfare scale. So the proponent of (E) must stipulate that harm, benefit and all the rest may not be understood broadly. However, making this stipulation results in (E) becoming circular. For to rule out the broad conceptions of harm, benefit, etc., one must insist that specifically welfare-type harm, welfare-type benefit and so on, are what (E) appeals to. In other words, (E) would have to appeal to modified versions of 1)-5) like the following: “Necessarily, if a person is *welfare-harmed*, then his welfare is decreased” and “Necessarily, if a person is *welfare-benefited*, then his welfare is increased.” But in that case (E) would be specifying what is distinctive of welfare by appeal to the notion of welfare itself. And that is circular.

Thus (E) faces a dilemma. Either harm, benefit, selfishness, altruism and quality of life are to be understood broadly or they are not. If they are to be understood broadly, then (E) will not be able to successfully distinguish welfare value from the other kinds of value. But if they are not to be understood broadly, then (E) will become circular. And so (E) is not a successful proposal about what is distinctive of welfare.

1.4.6 The ‘generic person’s preferences’ strategy

We need an answer to the question of what distinguishes welfare value from other kinds of value in order to be able to evaluate theories of welfare by appeal to our intuitions about that concept, as the standard methodology in moral philosophy requires. In this section, I propose a way of understanding what welfare value is that would allow some of our intuitions and preferences to count as evidence for or against theories of welfare.

There is a common style of philosophical argument that gives a vague indication of what is distinctive of welfare. It goes like this: ‘Theory T implies that life A would be better for a particular person than life B would be. But it's intuitive that if we were in this person’s shoes, we would rather have life B than life A. Thus theory T is false.’ The experience machine argument, Moore’s bestiality argument, Rawl’s grass-counter argument, and many other influential arguments in the literature on welfare, all have this form. In order for this to be a legitimate way to argue against a theory of welfare, we must assume that there is *some* connection between the welfare value of lives and our preferences about which lives we would rather lead. Accordingly, I believe that in trying to say what is distinctive of welfare value, we should start from the basic idea that welfare is related in some special way to people’s preferences about what life they want to have. If it can be shown that what is distinctive of welfare is the particular way in which it is related to people's preferences between lives, then this would explain why the form of argument mentioned above (which permeates the philosophical literature on welfare) is legitimate. That would be a very welcome result. And so this is the strategy I will pursue here.[38]

However, there are a couple challenges to be addressed in spelling out precisely what the connection is between welfare value and our preferences between lives. For one thing, the connection we are seeking cannot be the following very simple one: 'If you'd prefer life A to life B, then this is because it seems to you that A contains more welfare than B.' Clearly we might prefer to lead one life rather than another because it contains more of something *other* than welfare. Perhaps we'd prefer this life because it contains more moral goodness, or more aesthetic value, or because it would result in fewer people needlessly suffering, or whatever. Thus one challenge in spelling out the link between welfare and people's preferences between lives is to ensure that welfare value does not get confused with moral value, aesthetical value, or any of the other non-welfare type of value.

There is a second challenge to be addressed when spelling out the link between welfare and people's preferences between lives. This is the challenge of answering in a non-arbitrary way the question of precisely *whose* preferences are the ones that matter for evaluating candidate theories of welfare. Are *my* preferences the ones that matter? Are yours? Or is it the preferences of the members of a certain privileged group? If we are to identify a plausible link between welfare and our preferences among lives, then this question cannot be answered in an arbitrary manner.

Both these challenges can be met, I think, by appealing to a certain fiction (which is loosely inspired both by Rawls' notion of the original position and ideal observer theories of right action). This is the fiction of a generic person being put in a position to preview

[38] It might be thought that the argument I am proposing has the following logically invalid form: '1) If what's distinctive of welfare is the way in which it is connected to our preferences between lives, then a certain widely used form of argument is legitimate. 2) This widely used form of argument is legitimate. And so 3) what's distinctive of welfare is the way in which it is connected to our preferences between lives.'

However, this is not the way my argument should be construed. Instead, it should be taken to be an inference to the best explanation: '1) A certain widely used form of argument is legitimate. 2) The best explanation of 1) is that what's distinctive of welfare is the way in which it is connected to our preferences between lives. 3) If 1) and 2) are true, then 4). 4) Therefore, what's distinctive of welfare is the way in which it is connected to our preferences between lives.'

and then choose among various lives that he could lead. In particular, the notion of the generic person allows us to meet the second challenge just mentioned, while the choice scenario that the generic person is to be put in can be described in such a way as to meet the first of the two challenges above. Let me explain how I think this will work.

In order to provide a non-arbitrary answer to the question of whose preferences determine what welfare is, one thing to try would be to say that everybody's preferences matter. However, perhaps surprisingly, this too would be arbitrary. After all, why should only *actual* people's preferences get to count in determining what welfare consists in? Other people with different preferences might have existed instead of us, and then *their* preferences would have been the ones that determined what welfare is. The result in that case would have been that welfare ended up being something different than it actually is. Thus we would get a kind of modal arbitrariness that is no more acceptable than any other kind of arbitrariness when it comes to picking the people whose preferences determine what welfare is. Thus I think a better to proceed in forging the link between welfare and people's preferences between lives would be to *abstract away from the particular features specific individuals*. In particular, we need to abstract away from any particular conception of the good life, from anybody's particular projects, desires, goals, commitments or ideologies. Moreover, we need to abstract away from any particular person's character traits, talents or handicaps, and from any particular social or cultural conditions.[39] The fiction of the generic person provides a convenient way to abstract away from these things.

I propose that we should think of the generic person as a disembodied spirit who will be given a physical life on earth, but who knows nothing about what will happen to him in this life. In his[40] disembodied state, the generic person has no body or physical characteristics. He is under a veil of ignorance such that he does not know what sort of body he will eventually receive in his life on earth. Nor does he know what kind of society or cultural conditions he will be born into, or what position in society he will

[39] It should not be surprising that we need to abstract away from such things in specifying what is distinctive of welfare value. After all, many kinds of value – moral, aesthetic, epistemic, and welfare value, too – are abstract notions. The hard part is to say exactly how the abstraction is rooted in facts of a more concrete nature (e.g. empirical facts). Hopefully, what I say here should go some way towards clarifying this.

[40] I use the masculine gender just for convenience of writing. The generic person, as a disembodied spirit, of course has no gender.

enjoy. In fact, he knows nothing about what will happen to him in life: not what experiences he will have, what education he will receive, what childhood events will shape his personality, and so on.

When it comes to mental characteristics, the generic person, in his disembodied state, has no special psychological traits, talents or handicaps. His mind works in the same way that a human mind works: he is capable of having beliefs and memories, making inferences, having visual and auditory perceptions, and so on. But beyond that, the generic person does not possess any notable personality traits. His is like a brand new mind, fresh from the factory, without any of the emotional dispositions, behavioral tendencies or good or bad habits that characterize real people (with one exception soon to be discussed: viz. self-love). What's more, the generic person does not know what sort of mental abilities and personality traits he will develop once his physical life on earth begins. Let us suppose, however, that the generic person, in his disembodied state, is fully rational (i.e. as rational as a real human could be). His mental abilities are not limited by temporal or computational constraints: there is no limit on how many beliefs he can accommodate, and he has an infinite amount of time at his disposal to think, reflect, reason and infer; his memory is perfect.

So far, the generic person has been described in a way that abstracts away from any of the particular character traits, talents or handicaps that real people might have, and from any of the particular social or cultural conditions that real people live might in. But the generic person needs not only to be characterized so that the *workings* of his mind are generic, but also so that the *content* of his mind is generic. That is, we need to abstract away from any particular conception of the good life, from anybody's particular projects, desires, goals, commitments or ideologies. This can be done by stipulating that the generic person, while in his disembodied state, has no desires, wishes, plans or projects concerning anything on earth. He has no concern for what happens to any object or creature in the physical universe. There is no worldly item or event such that the generic person desires it, aspires to it or takes an interest in it. There is no worldly person, creature or cause such that the generic person is loyal to it, obliged to it, or cares about it.

To put the point in a more formal way, let us suppose that there is no 'worldly proposition' such that the generic person desires, wishes or intends that it be true. A ‘worldly proposition’ is any proposition describing an event in the physical universe (of which, recall, the generic person, as a disembodied spirit, is not a part), or which in any other way is about any object in the physical universe. Thus the generic person is stipulated to be indifferent to the goings on of the physical world. Of course, once the disembodied spirit begins his worldly life, he will come to have such desires, wishes, intentions and concerns. (In fact, part of the generic person’s task in selecting a life for himself will be to consider different desires and goals that he could come to have and pick among them.)

Let us make one last stipulation about the content of the generic person’s mind. While a disembodied spirit, the generic person has no moral beliefs to speak of, and no beliefs about what is beautiful, good, or right. He has no substantive conception of the good life. The generic person will eventually be given the task of previewing the possible lives he could lead on earth and then select among them, but before beginning this task, he does not have any evaluative beliefs or ideological commitments that could influence his choice among lives.

So to sum up, I have characterized the generic person as a disembodied spirit whose mind generally works in the same way that a normal human’s does, but who has no notable physical or personality traits, who has no desires or intentions with respect to any ‘worldly proposition,’ who has no moral or evaluative beliefs, and who is under a veil of ignorance with respect to what life he will eventually lead on earth. Characterizing the generic person in this way allows us to meet the first challenge from above, i.e. the challenge of answering in a non-arbitrary way the question of *whose* preferences between lives are determinative of welfare. The answer to this question is ‘the generic person’s preferences.’ It is the generic person’s preferences between lives that determine (in a way soon to be spelled out) what welfare value is.

In order to say exactly how the concept of welfare is linked to the generic person’s preferences between lives, imagine that the generic person as characterized above is given the following task. He is placed in a celestial viewing theater and given a very vivid presentation of every possible life that he could lead down on earth. More specifically, he will be given a viewing of every possible combination of the items in the following four categories:

1) the different personality traits, talents and handicaps a human being could have,
2) the different social and cultural conditions that a human being could live in,
3) the various desires, projects, tastes, motives and commitments that a human being could possess, and
4) the different events that a life could consist of – i.e. the different actions that a human being, once the parameters in 1-3 are pinned down, could perform and the different things that could happen to such a human being.

Once the generic person has previewed in vivid detail all of the possible combinations of these four things – i.e. all of the possible lives that he could lead – he is instructed to rank these possible lives according to his preferences between them. If the generic person would rather receive a given life than another, then the former would be placed higher on the ranking than the latter. Finally, the generic person is to pick the life that he would most prefer to lead down there on earth. This is the life that he will receive.

If the generic person is stipulated to be a disembodied spirit who has no particular concern for any worldly item or event, one might wonder how he could possibly form any preferences between the various lives he could lead. Recall that the generic person has a mind that generally functions in the same way as that of normal human beings. So to answer the question of how he forms preferences between lives, we may point to the fact that the mind of the generic person has this in common with real human beings: a sense of self-love. That is to say, the generic person cares about himself, he is disposed to form desires for the things that he comes to believe will benefit him, and he is disposed to defend himself if he perceives any danger. In whatever way it is that real human beings display self-love, so it is that the generic person is motivated by self love as well.[41] This is the guiding motivational force that allows the generic person to form preferences among the various combinations of personality traits, social conditions, desires and events that he gets to preview in the celestial theater.[42]

[41] By saying that the generic person is motivated by self-love in the same way normal people are, I am trying to prevent my proposal from becoming circular. I am on purpose not taking self-love to be anything like the desire that one's own life go well, or that one flourish, or that one lead a good life, or for any other thing that is equivalent to a life high in welfare. This would make my proposal come out circular. So instead, I take it that there is some independent, empirical way of pinning down the sort of self-love that human beings actually display. And whatever this turns out to be, this is what generic person displays as well. (Darwall, 2002, seems to adopt a similar strategy for avoiding circularity.)

[42] In taking self-love to be the principle on which the generic person forms preferences between lives, I mean to be giving a nod to Darwall. He was right, I think, that caring about a person - in particular, caring for *oneself* – is connected in a deep way to the concept of welfare. However, as we saw in section 1.4.3, Darwall didn't specify his proposal in careful enough a way to distinguish welfare from other types of

A final stipulation about the generic person's choice situation needs to be made. As we will see, this last stipulation will help to address the first of the two challenges I mentioned above (i.e. the challenge of ensuring that welfare value does not get confused with moral value, aesthetic value, or any of the other non-welfare type of value). In particular, the generic person is instructed that when forming his preferences between the lives he could lead, he may not take into consideration the effects on any other creature of his actions in his earthy life or of what happens to him there. He is instructed to form his preferences between lives solely on the basis of his sense of self-love. To ensure the indifference of the generic person (at least while he's in his disembodied state) towards others, he is assured (in a way that leaves no possibility of doubt) that any pain or harm to other creatures that he might cause in his life on earth will be fully and generously compensated in the afterlife of these harmed creatures. Accordingly, the generic person does not have to worry about any of the pain or harm he might cause to other creatures down on earth. Thus his preferences between lives will be formed *solely* on the basis of his sense of self-love.

Now we are in a position to say what is distinctive of welfare value. The different lives that the generic person gets to preview in the celestial theater – i.e. the different combinations of 1) personality traits, 2) social and cultural conditions, 3) desires and projects and 4) actions he might perform and things that might happen to him – will differ with respect to something that makes certain lives more preferable to the generic person than others. This something is welfare value. Thus we may say that welfare value is the kind of value that the generic person would prefer to have more of and that would be maximized by the life he would ultimately pick for himself. My proposal, then, about what is distinctive of welfare is this:

> (F) X is what intrinsically enhances a person's welfare iff X is what the generic person, as I characterized him and when he is placed in the choice scenario I described, would prefer to have more of in his earthly life, and X is what is maximized by the life that this generic person would ultimately choose to lead.

This is my proposal about what is distinctive of welfare value. I think that it meets the two challenges mentioned above. For one thing, (F) does not characterize welfare in a

value a life might display. My device of the generic person and his choice scenario is meant to rectify this defect in Darwall's basic idea, which I have some sympathy for.

way that makes it depend on the preferences of some arbitrarily selected person or group of people. Moreover, I think (F) succeeds in distinguishing (in a non-circular way) welfare value from the other types of value a life may possess. After all, what sort of life would the generic person prefer to lead down on earth? For one thing, he will not prefer to lead a life that is higher in moral or aesthetic value. For the generic person was stipulated to have no views about what is beautiful or about how other people should be treated that could influence his preferences. Moreover, the generic person will not be led out of sympathy or concern for others to choose a life that is good for others but not maximally good for himself. For the scenario was stipulated to be such that whatever pain or harm the life he chooses might cause to others will be generously compensated to the victims, and so the generic person does not have to consider the effects of the life he chooses on any other creature. His preferences between lives are guaranteed to be determined solely by his self-love. Furthermore, the fact that the generic person was stipulated to lack any worldly commitments, any ideological agenda, and any specific projects guarantees that the generic person won't simply prefer a life because it is better for some political, ideological or religious cause. In general, I am inclined to think that the way in which the generic person was described, as well as the way in which his choice scenario was described, guarantees that (F) does not get welfare value confused with the other types of value a life might instantiate. Thus as far as I can tell, the first challenge mentioned above seems to be met as well.

1.4.7 Conclusions

As I said, the idea of the generic person choosing what sort of life he would prefer to lead is a fiction. But it is a useful fiction. It is a way of abstracting away from any particular person's preferences, evaluative beliefs or substantive conception of the good life. However, because of all this abstraction, it might seem hard to answer the question of what life the generic person would prefer. Nonetheless, I assume that there is a fact of the matter about what sort of life he would prefer under the conditions I described. It is the moral philosopher's job to figure out what this fact of the matter is.

If my proposal, (F), is correct, then the intuitions that various theories of welfare are to be tested against are intuitions about what sort of life the generic person, as described above, would prefer. For if welfare is the sort of value that the generic person would prefer more of in the life that he will lead down on earth, then reflecting on our intuitions about what the generic person would prefer will help us figure out what theory of welfare is true. This, then, is how the various intuitions about welfare that I talk about in the rest of this dissertation should be understood. They should be seen as being (or amounting to) intuitions about what sort of life the generic person would prefer. If my proposal is right, then we are free to go ahead and employ the standard intuition-based methodology in seeking the true theory of welfare. Having said my piece about all these methodological questions, then, let me go on to do just that.

CHAPTER 2

PUTTING TAXONOMY TO WORK

In this chapter, I develop a taxonomy of theories of well-being.[1] This is useful because on my taxonomy, theories that share certain fundamental flaws are grouped together. Thus my taxonomy provides a convenient way to offer blanket arguments against whole categories of theories. My taxonomy divides all theories of well-being into three main categories, and in this chapter I will argue that the true theory of well-being, whatever it is, must fall under the second category in my taxonomy. By clearing the playing field early on in this way, I will be able to focus my investigation on the theories that have the most going for them. The thesis that I will argue for in this chapter amounts to the claim that well-being must have at least some objective component (though, as we'll see, I think the terminology of 'subjective theories' and 'objective theories' should be avoided due to its ambiguity).

2.1 The Need for a New Taxonomy

Traditionally, theories of well-being are divided into three main categories: Hedonistic Theories, Desire Satisfaction Theories and Objective List Theories.[2] Parfit, for instance, gives the following rough picture of these three types of theory:

> On *Hedonistic Theories*, what would be best for someone is what would make his life happiest [where happiness, usually, is understood in terms of pleasure and the absence of pain]. On *Desire-Fulfillment Theories*, what would be best for someone is what, throughout his life, would best fulfill his desires.

[1] Thanks to Pete Graham for helpful conversations that led to the development of this taxonomy.
[2] See, for example, Parfit, 1984, pp. 493-502; Kagan, 1998, pp. 29-41; and Crisp, 2008

> On *Objective List Theories*, certain things are good or bad for us, whether or not we want to have the good things, or to avoid the bad things. (Parfit, 1984, p. 493)

For each type of theory, there are countless versions. Under the heading of "Hedonism," you find (among others) Bentham's Sensory Hedonism, Mill's Qualified Hedonism and Feldman's Desert-Adjusted Hedonism. Among the Desire Satisfaction Theories, you encounter Actual Desire Satisfactionism and Ideal Desire Satisfactionism, as well as Unrestricted Desire Satisfactionism and Restricted Desire Satisfactionism. The Objective List Theories comprise Aristotelian Perfectionism, Nietzschean Perfectionism and a host of other theories that appeal to hodge-podge lists of goods that do not seem to be constructed on the basis of any one principle.

This way of carving up the theoretical landscape is very messy. For one thing, many important and promising theories of well-being do not fall neatly into just one of these categories. For instance, is Sumner's Authentic Happiness theory of well-being properly classified as a Desire-Satisfaction view or a Hedonistic view? Sumner's insistence on happiness having both a 'cognitive and an affective component'[3] seems to suggest that the theory incorporates elements of both Desire Satisfactionism and Hedonism. Moreover, certain philosophers have attempted to blend Hedonism or Desire Satisfactionism with certain objective elements. Parfit (1984) and Kagan (1998) debate the merits versions of Desire Satisfactionism that place objective restrictions on the desires whose satisfaction would count as welfare enhancing, while Mill (2001, ch. 2) and Feldman (2004) advocate certain objective restrictions on how the contributions to welfare of various kinds of pleasures and pains are to be determined. Where are we to place such versions of Desire Satisfaction and Hedonism, given the objective restrictions they incorporate?

Matters are made worse by the fact that the traditional tri-partite division of theories is not exhaustive. Many philosophers working on well-being today favor various kinds of hybrid theories on which there is not just one fundamental bearer of welfare value (like pleasure or desire satisfaction or perfection) but several.[4] These hybrid theories do not fall clearly under any of the three main headings in the traditional taxonomy. Given the

[3] Cf. Sumner, 1996, p. 146

[4] See, for example, Adams, 1999, pp. 93-94; Brink, 1989, ch. 8; Parfit, 1984, p. 502; Scanlon, 1998, ch. 3.

limitations of the traditional tri-partite division of theories of well-being, we would do well to seek a more systematic way of dividing up the terrain.

2.2 Subjective vs. Objective Theories of Well-Being

In my view, a better way to organize theories of well-being would be to group them, roughly speaking, according to whether they are 'subjective' or 'objective'. This distinction provides a systematic and exhaustive way of carving up the theoretical landscape. In section 2.3, I will present my own taxonomy, which proceeds roughly along these lines (though to prevent confusion, I will use different terminology). Nonetheless, many extant attempts to spell out the distinction between subjective and objective theories are problematic. The taxonomy that I will go on to present is designed to avoid these problems. Thus to understand the motivation for the various features of my taxonomy, let me begin by discussing the problems for previous attempts to distinguish between subjective and objective theories of well-being.

2.2.1 Kagan and Parfit – It's not just desires

Shelly Kagan and Derek Parfit give similar accounts of what distinguishes an 'objective' theory from a 'subjective' one. As Kagan puts it, the objective theories are the ones according to which

> being well off is a matter of having certain goods in one's life, goods that are simply worth having, objectively speaking. Similarly, there may be certain objective bads or evils, the having of which simply leaves one worse off. (…) And the goods and evils themselves have intrinsic value or disvalue independently of our desires (actual or ideal)… (Kagan, 1998, p. 39)

In a similar vein, Parfit takes objective theories to be the ones on which

> certain things are good or bad for us, whether or not we want to have the good things, or to avoid the bad things. (Parfit, 1984, p. 493)

These two passages suggest the following account of the distinction between subjective and objective theories of well-being:

> $O_{K\&P}$) A theory of welfare, T, is **objective** iff T implies that states of affairs can positively (or negatively) impact a person's welfare even though that person does not desire that they obtain (or that they not obtain).
>
> $S_{K\&P}$) A theory of welfare, T, is **subjective** iff T is not objective.

This way of cashing out the distinction between subjective and objective theories of well-being does not capture the intuitive sense of the terms ‘subjective’ and ‘objective.’ For $O_{K\&P}$) and $S_{K\&P}$) would classify as objective certain theories that are *obviously* subjective. Take Simple Sensory Hedonism (SSH). This is roughly the theory that one is well-off to the extent that one experiences more sensory pleasure and less sensory pain. If any theory of well-being deserves to be called a subjective theory, this one does. However, it counts as an objective theory according to $O_{K\&P}$) and $S_{K\&P}$). After all, SSH implies that some states of affairs can positively impact one’s welfare even though one does not desire that they obtain. Suppose, for example, that there is a religious person who has no desire whatsoever to experience any sexual pleasure in life. We could even suppose that the person has an explicit desire to *not* experience any such pleasure. Let’s imagine that one night, despite his desires to the contrary, this person gives in to temptation and has a sexual experience that is very pleasurable. SSH would imply that the state of affairs consisting of this person’s feeling pleasure because of the sexual encounter enhances this person’s welfare *even though the person does not desire in the least that it obtain.* Thus according to $O_{K\&P}$), SSH would count as an objective theory. But this seems wrong. Since SSH makes one’s well-being be entirely dependent on what is in the subject’s head, it should be regarded as a subjective theory.[5] So $O_{K\&P}$) cannot be the right way to cash out the notion of objectivity with respect to welfare. By extension, $S_{K\&P}$) can’t be right either.

The problem (or one of them at least) with this attempt to cash out the objective-subjective distinction is that it appeals only to what one desires. But there are many theories of well-being, some of which are objective and some of which are subjective, which do not mention desires at all. Thus in order to successfully characterize the distinction between subjective and objective theories, one must appeal to something broader than merely the notion of what one desires.

[5] For this reason, I am inclined to count as subjective-seeming theories both the *Knowledge Theory*, which states that the more true beliefs you possess the higher your welfare, and the *Belief Theory*, which states that the more beliefs, true *or* false, that you possess the higher your welfare. These theories would count as entirely response dependent, on my definition of that term. However, they would count as objective according to $O_{K\&P}$) and $S_{K\&P}$). So much the worse for $O_{K\&P}$) and $S_{K\&P}$), it seems to me.

Could this problem be avoided by appealing to the broader notion of *attitudes* instead? This would involve substituting the talk of desires in $O_{K\&P}$) with talk of attitudes in something like the following way:

> $O_{K\&P}$') A theory of welfare, T, is **objective** iff T implies that states of affairs can positively (or negatively) impact a person's welfare even though that person does not have any *pro-attitude towards it* (or have any con-attitude).[6]

The definition of subjective theories, of course, can remain the same.

This modification will not avoid the problem, however. After all, sensory pleasures are not *attitudes*. What makes a mental state an attitude, on the standard account, is that it has *propositional* content. Desires, beliefs, hopes, wishes, intentions, and so on are all attitudes because they are states that have propositions as their content. However, sensory pleasures do not have propositional content. They have phenomenal content, perhaps, but they are not states that have propositions as their content. And so they do not count as attitudes. The upshot of this is that according to $O_{K\&P}$') a theory like Simple Sensory Hedonism would still count as an objective theory. Because sensory pleasures are not attitudes, SSH allows that a state of affairs may enhance one's welfare no matter what attitudes one has towards it. And so the problem in question remains even for $O_{K\&P}$').

In order to avoid this problem, we need to appeal to a notion that is also broader than that of attitudes. In particular, we need a notion that covers *both* propositional attitudes *and* non-propositional reactions like sensory pleasures and pains. Towards this end, I propose to follow the lead of Michael Huemer and make use of the notion of *psychological responses*.[7] I will use term 'psychological responses' as a generic way to refer both to the various ways in which one might react psychologically to things that happen to one and to the various mental states that one might acquire as a result of what happens to one. Thus the term 'responses' is supposed to cover both propositional attitudes, like desires, beliefs, intentions, wishes, hopes, etc., and non-propositional psychological reactions, like sensory pleasures, visual perceptions and other experiences one might have.

[6] This view is very close to the denial of a view known as internalism about a person's good. See ch. 3 of this dissertation for a critical discussion.

[7] Cf. Huemer, 2005, pp. 1-13. Note that I do not claim to be using the term in exactly the same way that he does. I mean to use it just in the way that I explain above.

Now, there is a second problem that afflicts Kagan's and Parfit's proposal. However, it is a problem that also afflicts the proposal offered by Michael Huemer, and it is more convenient to bring out the problem by discussing his proposal.

2.2.2 Huemer – It's not 'responses towards'

Huemer offers an account of what makes a given property be subjective as opposed to objective. He does not specifically talk about the difference between subjective and objective theories, but his account of the notion of a subjective property clearly suggests a way to distinguish subjective from objective theories. Here is how Huemer explains his account of what makes a *property* be subjective:

> *F*-ness is subjective = Whether something is *F* constitutively depends [as opposed to causally depends] at least in part on the psychological attitude or response that observers have or would have towards that thing. (Huemer, 2005, p. 2)

Huemer then goes on to say that the objective properties just are the ones that are not subjective. Given this account of subjective and objective *properties*, it is reasonable to suppose that Huemer would endorse something like the following account of the distinction between subjective and objective *theories of welfare*:

> S_H) A theory of welfare, T, is **subjective** iff T implies that what makes something (e.g. an object or a state of affairs) be welfare enhancing is its possession of a subjective property.
>
> O_H) A theory of welfare, T, is **objective** iff T implies that what makes something be welfare enhancing is its possession of an objective property.

The idea here is this. According to the subjective theories, things like drinking a glass of beer or looking at a beautiful painting would enhance my welfare only if I am affected in the right way. For instance, drinking the beer might enhance my welfare only if it is enjoyable to me. The property of being enjoyable to me is subjective, according to Huemer's account, because whether or not a thing has that property depends on what my psychological responses are to that thing. Thus a subjective theory of welfare is one that implies that the welfare value of a given thing for me depends on what my psychological responses towards it are. The objective theories, by contrast, are the ones that deny this.

Because Huemer's account proceeds in terms of psychological responses, it seems to avoid the problem that damaged Kagan and Parfit's proposal. However, S_H) and O_H) face a different sort of problem. In particular, it is a technical, metaphysical problem.[8] On Huemer's account, subjective theories are the ones according to which something enhances one's welfare if and only if one has the right psychological responses towards that thing. But this is what causes the problems. We cannot take it that what makes a theory of welfare subjective is that it says that something can be welfare enhancing only if one has the right responses *towards that thing*. For a great many theories of welfare say that certain states of affairs are welfare enhancing no matter what responses one has *towards those very states of affairs*.

As examples, let's look at Desire Satisfactionism and Simple Sensory Hedonism. On Desire Satisfactionism, there is one and only one kind of thing that is welfare enhancing – namely, episodes of desire satisfaction, or to put it another way, states of affairs in which some desire of the agent's is satisfied.[9] But whether or not one of *these* states of affairs would be welfare enhancing does *not* depend on any of your responses, even your desires. Imagine a state of affairs – call it '$S_{\text{satisfied}}$' – that consists of your getting your desire to drink a Carlsberg beer satisfied. Now compare this to the state of affairs consisting just of your drinking a Carlsberg – call it 'S_{Carl}'. Obviously, the latter state of affairs is not of the sort that can by itself enhance your welfare, according to Desire Satisfactionism. By contrast, on Desire Satisfactionism, $S_{\text{satisfied}}$ would indeed enhance your welfare if actual *even if you do not desire that* $S_{\text{satisfied}}$ *obtains* – that is, even if you do not desire any episode of desire satisfaction. (Maybe you've never heard of an episode of desire satisfaction before.) What this example shows is that according to O_H), Desire Satisfactionism would count as an objective theory. After all, according to Desire Satisfactionism, the question of whether a given state of affairs is welfare enhancing does not depend at all on what desires one has *towards* that very state of affairs. According to Desire Satisfactionism, a state of affairs like $S_{\text{satisfied}}$, which consists of your getting one of your desires satisfied, would enhance your welfare no matter what your desires are towards *that very* state of affairs. Furthermore, according to Desire Satisfactionism, no matter what your attitudes towards drinking a Carlsberg are, the state of affairs S_{Carl} (i.e. your drinking a beer) would not by itself be capable of enhancing your welfare. Thus O_H)

[8] Note that this problem is shared by $O_{K\&P}$) and $S_{K\&P}$).

[9] This claim is crucial for how I conceive of the taxonomy I develop in this chapter. I explicitly defend this claim in ch. 3, of this dissertation.

would imply that Desire Satisfactionism is an objective theory. For according to Desire Satisfactionism, the welfare values of certain states of affairs do not depend on one's psychological responses *towards those very states of affairs*. But the result that Desire Satisfactionism is an objective theory does not seem plausible. Desire Satisfactionism is a clear example of a subjective theory. So S_H) and O_H) are in trouble.[10]

S_H) and O_H) encounter a similar problem when it comes to Simple Sensory Hedonism (SSH) as well. The states of affairs that SSH implies are welfare enhancing are the ones consisting of your experiencing an episode of sensory pleasure. So if you're drinking a beer and it feels good, then what enhances your welfare, according to SSH, is not the state of affairs consisting of your drinking a beer, but the state of affairs consisting of your experiencing sensory pleasure from drinking the beer. Call the former state of affairs 'S_{beer}' and the latter state of affairs '$S_{beer\text{-}pleasure}$'. According to SSH, $S_{beer\text{-}pleasure}$ will enhance your welfare no matter what your psychological responses to $S_{beer\text{-}pleasure}$ *itself* are. Moreover, according to SSH, S_{beer} does *not* by itself enhance your welfare even if you *do* receive pleasure from the drinking. For S_{beer} is simply not the kind of state of affairs that can be welfare enhancing according to SSH. Thus on SSH, whether a given state of affairs has the property of being welfare enhancing does not depend at all on what your responses *towards* that very state of affairs are. And so O_H) also implies that SSH is an objective theory of welfare. But clearly this too is an absurd result. SSH is a paradigmatic example of a subjective theory.

These problems arise because Huemer's account implies that a theory is subjective iff that theory says a state of affairs is welfare enhancing as long as one has the right responses towards that very state of affairs. But as we have just seen, the subjective

[10] Note that in chapter 3, I present an argument based on considerations very much like these against a view known as Internalism about a Person's Good. (See in particular my argument against two-tier internalism.) This makes me wonder if a satisfactory version of Internalism about a Person's Good can be formulated by appeal to the idea of one's responses *in* a state of affairs, rather than *towards* a state of affairs. In particular, the view might be formulated as something like this: 'X can be good for a person, P, only if X is a state of affairs such that P has certain positive responses in X.' Thus apples and beers can't be good for people, according to this version of internalism, but states of affairs can (in particular, states of affairs in which one has positive responses to various things). This is clearly not a satisfactory formulation of internalism yet, however. For one thing, internalism is usually taken to require only that *it's possible* for you to care about something in order for that thing to be good for you. I don't see how to work this into the sort of formulation I'm proposing. In any case, this might be an avenue for the internalist to pursue.

What's more, I want to emphasize that the issues discussed in this chapter concerning the subjective-objective distinction are clearly closely related to the issue of internalism discussed in chapter 3.

theories are not the ones that take it that for a state of affairs to enhance your welfare, you need to have the right responses towards *that very state of affairs*. Rather, what seems to matter is whether the theory takes it that for a state of affairs to enhance your welfare, you need to have the right responses *in that state of* affairs. Thus to avoid the present problems, we need to draw the distinction between subjective and objective theories in a way that accommodates this. We can do this if we say that the subjective theories are the ones according to which a given state of affairs, S, would enhance one's welfare provided one has the right responses *in S*.[11] As we will see, the taxonomy I present in section 2.3 is formulated in just this way. Thus my taxonomy correctly classes Desire Satisfactionism and Simple Sensory Hedonism as theories of the 'subjective' type.[12]

[11] If you prefer, you could remain neutral on the question of whether, according to subjective theories, it is *responses in* or *responses towards* a state of affairs that determines its welfare value. To remain neutral on this, all we have to do is to take it that the subjective theories are the ones according to which a given state of affairs, S, would enhance one's welfare provided one has the right responses *in or towards S*. However, I don't think that this extra bit of complexity is necessary.

[12] Wayne Sumner, too, discusses the question of what characterizes a subjective theory of welfare. (Wayne Sumner, 1995, pp. 764-790) For ease of exposition, I opted not to discuss Sumner's article in the body of the text. For one thing, it is not clear what Sumner's view is, however. He says different things that suggest different accounts. Second, his proposals suffer from the same sort of problems we have already encountered. Discussing Sumner would not allow me to address any new issues in addition to the ones I already talk about in the body of the text.

One proposal of Sumner's about what makes a theory of welfare be subjective is this: 'something can make me better-off on this sort of account [i.e. a subjective theory] only if I have a positive attitude (of one sort or another) toward it.' (Sumner, 1995, p. 767) But this account clearly suffers from the problems discussed here in section 2.2.1 and 2.2.2. For one thing, Sumner's talk of having a positive *attitude* is not broad enough. It leaves out theories that make welfare depend on non-propositional positive responses (like sensory pleasure). More importantly, Sumner formulates his account of subjective theories in terms of the requirement that in order for X to be good for one, one must have a positive attitude *towards it*. But as we saw in 2.2, this is implausible because according to all sorts of clearly subjective theories, various things will be intrinsically good for you even though you would not have any positive response directly towards it. For instance, Hedonism implies that pleasure is good for you even if you do not explicitly desire or approve of the pleasure you get. This account of what makes a subjective theory simply has to go.

However, another comment of Sumner's suggests a different sort of account of what makes a theory subjective. It seems to be better in certain respects: 'a theory treats welfare as subjective if it makes it depend, at least in part, on some attitude or concern on the part of the welfare subject. More precisely, a subjective theory will map the polarity of welfare onto the polarity of attitudes, so that being well-off will depend (in some way or other) on having a favorable attitude toward one's life (or some of ingredients)…' (Sumner, 1995, p. 767) This certainly sounds better. It mentions concerns in addition to attitudes. And it doesn't mention any requirement to the effect that one must have a positive response *towards* X in order for X to be good for one. However, the proposal rather vague. What does it mean for a theory to make one's welfare *depend* on one's attidues or concerns? To give a more precise formulation of what makes a theory be subjective, our account must specify *how it makes the welfare value of a state of affairs for a person depend on that person's psychological responses.* My definition of an 'entirely response dependent theory', which I present in section 2.3, does exactly this. Thus my taxonomy is better than Sumner's suggestions because my taxonomy is both plausible and precise.

2.2.3 Huemer again – How to carve up the landscape?

Even if the Huemer account were to be modified in such a way as to avoid the previous problem, there would still be something contentious about his account. On Huemer's account, a subjective theory would, very roughly, be one that makes the welfare value of states of affairs depend at least *in part* on one's responses. But is this the right way to carve up the theoretical landscape? Some might think it is more plausible to say that the subjective theories are the ones that make the welfare value of states of affairs depend *solely* on one's responses. The question of which of these two options to pick has consequences for which theories will fall on which side of the battle lines.

'So what?', one might ask. Does it matter where one draws the battle lines? Perhaps not. Huemer might just have in mind one, but not the only, acceptable way of carving up the landscape. However, I think there is one consideration that suggests that it is more plausible to say that the subjective theories are the ones that make welfare depend *solely* on one's responses (rather than just *in part*, as Huemer does).

Many philosophers are inclined to favor so-called 'subjective' theories of welfare (whatever that means) because they think that these theories do not face a certain prima facie explanatory challenge that the 'objective' theories do face. I don't have to explain this challenge in detail here, but the basic thought is that there is something vaguely mysterious to the idea that certain things can be good for a person no matter what one's responses are, and any acceptable 'objective' theory must discharge the burden of explaining how this is possible.[13]

Now suppose we follow Huemer and take it that the major distinction in the welfare landscape should be drawn between a) those theories that make the welfare values of states of affairs depend at least *in part* on one's responses and b) those theories that do not. If we do this, then we end up with many *subjective* theories that face this prima facie explanatory challenge as well. For example, consider a hybrid theory of welfare according to which one's level of welfare is determined by the amount of pleasure one feels plus the amount of success one has in one's career. According to this theory, welfare has a subjective component, involving pleasure, and an objective component, involving professional success. The theory makes welfare depend at least in part on one's

[13] For an in depth treatment of this issue, see chapter 3.

responses, and thus it would qualify as a subjective theory on Huemer's way of cutting up the terrain.

However, this hybrid theory inherits the prima facie explanatory challenge that might worry some people about objective theories. In particular, this hybrid theory would seem to have some explaining to do when it comes to the question of how the objective facts it says welfare is partially dependent on (i.e. the facts about professional success) could be relevant to one's good no matter what one's responses are. So on Huemer's way of understanding the subjective-objective distinction, some subjective theories face the very same prima facie explanatory challenge that worries some people about the objective theories. This would be bothersome to those philosophers who are inclined to say that they favor subjective theories because these theories do not face the explanatory challenge that casts doubt on the objective theories.

But this messy situation does not arise if one takes the important distinction to be between a) those theories that make welfare *entirely* dependent on responses and b) those theories that do not. On this way of cutting up the terrain, all the theories that face the prima facie explanatory challenge mentioned above, which might worry subjectivist-friendly philosophers, *will fall into the same group* – that is, the group of theories that makes welfare depend at least in part on something other than one's responses. By contrast, the entirely response-dependent group will contain no theories that face this prima facie explanatory challenge. And so the terrain of theories of welfare would be divided up in a much tidier way. Thus there seems to be a small advantage to cutting up the terrain not as Huemer does (i.e. by distinguishing between theories that make welfare *somewhat* response dependent and those that don't), but rather by distinguishing between theories that make welfare *entirely* response dependent and those that don't.

However, I grant that this argument against cutting up the terrain as Huemer does might remain controversial. So what I propose to do is to sidestep the whole issue. Instead of taking it that there are two main types of theories of welfare, the subjective ones and the objective ones, I will take it that there are three main types. You have the theories that make welfare depend entirely on one's responses, those that make welfare depend only in part on one's responses, and those that make welfare depend not at all on one's responses. Opting for a binary distinction, between subjective and objective, theories leaves open the difficult question of precisely where to draw the battle lines between subjectivists and objectivists. But a three-way distinction of the sort I propose allows us to circumvent this tricky issue. It gives a more accurate picture of the theoretical landscape.

2.3 A Better Taxonomy

In light of the problems we have just seen with other attempts to spell out a distinction between subjective and objective theories of well-being, I favor a three-way distinction between *entirely response-dependent theories*, *partly response-independent theories* and *entirely response-independent theories*. This taxonomy classifies theories of well-being according to whether or not they take it that one's psychological responses are all or part of what determines the welfare value of a state of affairs for one. Roughly speaking, this taxonomy maps onto the distinction between the theories that intuitively seem to be 'subjective', those that seem to be partly 'objective' and those that seem to be entirely 'objective.' However, (as we have seen) the terms 'subjective' and 'objective' are used in different ways by many different people. So to avoid confusion , I opt not to use this terminology in spelling out my taxonomy. But more importantly, I argue that my taxonomy avoids the problems encountered in the previous section.

Let me begin by explaining the technical terms that my taxonomy employs. For starters, my taxonomy centrally involves a certain term of art, viz. 'psychological responses'. The way I use this term, it refers to the ways in which one might react psychologically to things that happen to one and the various mental states one might acquire as a result of what happens to one. Thus (as noted above) the term covers both *propositional attitudes* – like desires, beliefs, intentions, wishes, hopes, etc. – and *non-propositional psychological reactions* – like sensory pleasures, visual perceptions and other experiences one might have.

Second, for reasons explained in section 2.2.2, my taxonomy is defined in terms of what a person's psychological responses are *in* a state of affairs. I should point out that the phrase 'P's psychological responses in S' is not supposed to mean simply the ways in which P reacts to *the features of S*. In addition to this, this phrase is also supposed to denote all the attitudes and non-propositional responses (like pleasures and perceptions) such that P has them, or tokens them, in S. So suppose S is a state of affairs consisting of three components: a) you drink a Carlsberg right now, b) you derive a certain amount, x, of sensory pleasure from your drinking Carlsberg right now, and c) you believe that √2 is an irrational number between 1 and 1.5. Items b) and c) here count as your 'responses in S'. Item b) is a response that you have *towards* another feature of S, namely your drinking a Carlsberg. Item c), too, is a response that you have *in* S, but it is not a response towards any feature of S. I intend that both kinds of responses – responses that are had toward features of S and responses that are not – be covered by the term 'psychological responses in S'.[14]

With this terminology in hand, I can now present the taxonomy itself. The entirely response dependent theories, roughly speaking, are the ones that imply that the degree to which a person would be benefited by obtaining a given end is determined solely by what that person's psychological responses are. More precisely, the idea is this. Consider a theory of welfare, T. Now consider a possible state of affairs, S, (in particular, a non-evaluative one). Suppose S were made actual. Now suppose that according to T, in order to figure out how much P's welfare is intrinsically enhanced or decreased by S, *all* you need are facts about either a) the actual strengths[15] of P's responses in S, or b) what the strengths of P's responses would be in S if P were idealized in some way. If this is the case according to theory T, then T will be an entirely response dependent theory. To capture this, we can adopt the following definition:

> D1. T is **entirely response dependent** iff T implies that there is a function *solely* from **a)** a possible state of affairs, S, and **b)** the strengths of P's psychological responses in S (or what they would be if P were idealized in certain ways), to **c)** the degree to which S would, if actual, intrinsically impact P's welfare.

[14] Perhaps we could give the following mereological gloss on the notion of having a response *in* a state of affairs:

> P has response R *in* state of affairs S =df. S is a state of affairs that is composed at least in part of the state of affairs of P's having response R.

(Something roughly along these lines is hinted at in Dale Dorsey, 'The Hedonist's Dilemma', manuscript, p. 4. In particular, see his discussion of the possible objection to his taxonomy, at the end of section 1.)

[15] For present purposes, we can take the strength of a pleasure, like the strength of a desire, to be its intensity times its duration.

Thus the entirely response dependent theories imply that the facts about the strengths of P's psychological responses in S are the only thing needed to figure how much a given state of affairs, if actual, would intrinsically impact P's welfare.

By contrast, the entirely response *in*dependent theories are the ones according to which one's psychological responses are not at all relevant to determining the welfare value of a state of affairs for you. So consider a theory of welfare, T, and a possible (non-evaluative) state of affairs, S. Suppose S were made actual. Now suppose that according to T, in order to figure out how much P's welfare is intrinsically enhanced or decreased by S, you do not need *any* facts about the strengths of P's responses in S (or what they would be if P were idealized in some way). If this is the case according to theory T, then T will be an entirely response independent theory. To capture this, we can adopt the following definition:

> D2. T is **entirely response independent** iff T implies that there is a function just from **a)** a possible state of affairs, S, and **b)** something *other than* P's psychological responses in S or their strengths, to **c)** the degree to which S would, if actual, intrinsically impact P's welfare.

Thus the entirely response independent theories imply that the degree to which a state of affairs would (if actual) intrinsically impact P's welfare is *not at all* determined by the facts about P's psychological responses in S.

Finally, we have the partly response independent theories. These are the theories according to which one's responses matter to welfare, but are not the *only* thing that matter. Consider a theory of welfare, T, and a possible (non-evaluative) state of affairs, S. Suppose S were made actual. Now suppose that according to T, in order to figure out how much P's welfare is intrinsically enhanced or decreased by S, you need *both* a) the facts about the strengths of P's responses in S (or what they would be if P were suitably idealized), *and* b) some *additional* facts that do not concern the strengths of P's responses in S. If this is the case according to theory T, then T will be a partly response independent theory. To capture this, we can adopt the following definition:

> D3. T is **partly response independent** iff T implies that there is a function just from **a)** a possible state of affairs, S, **b)** the strengths of P's psychological responses in S (or what they would be if P were idealized in certain ways), and **c)** something *other than* P's psychological responses in S or their strengths, to **d)** the degree to which S would, if actual, intrinsically impact P's welfare.

Thus the partly response independent theories imply that the degree to which S (if actual) would intrinsically impact P's welfare is determined *in part but not entirely* by the facts about P's psychological responses; something else in addition to the facts about P's psychological responses are required to determine how good S would be for P.

It should be clear that this taxonomy is exhaustive. Any theory of well-being will belong to one of these three types. After all, the function from a possible state of affairs, S, to the welfare value that S (if actual) would have for you will need to be supplemented either *only* by the facts about your responses in S, or else *not at all* by these facts, or else by these facts together with some other facts that don't concern you responses. The exhaustiveness of this taxonomy makes it superior to the traditional tri-partite taxonomy of theories of well-being discussed in section 2.1.

To fully explain this taxonomy, I need to say something about which theories fall under which category. I also need to explain the importance of the parenthetical phrase 'if actual' in these definitions, but I will get to that in a moment (when I discuss the entirely response dependent theories).

For ease of exposition, begin with the class of the entirely response *independent* theories. There are not many philosophers who defend theories of this sort. But many toy theories of well-being, often used for illustrative purposes, would count as entirely response independent. Consider, for instance, a theory on which one's level well-being directly corresponds to the number of dollars in one's bank account. Or consider a theory on which one's level of well-being is determined solely by the amount of political power one enjoys, or by the number of friends one has. These theories would be entirely response independent because they imply that the facts about what responses one has in a given state of affairs has no relevance to the question of how much that state of affairs (if actual) would intrinsically enhance or decrease your welfare. According to the Money Theory, the Power Theory and the Friends Theory, the degree to which a state of affairs would (if actual) intrinsically impact one's welfare is *not at all* determined by the facts about one's psychological responses in S. And in general, paradigmatic examples of Objective List Theories would tend to count as entirely response independent theories.[16]

When it comes to the class of entirely response dependent theories, they are meant to comprise the theories of well-being that have a predominantly 'subjective' feel. One important group of these theories are the ones commonly referred to as *mental state theories of well-being*.[17] I propose that we define the mental state theories as follows:

> D4. T is a **mental state theory of welfare** iff T implies that there can be no difference between person A's level of welfare and person B's level of welfare without there being a difference between the mental states of A and the mental states of B.[18]

Simple forms of Hedonism, like Bentham's Sensory Hedonism and (un-adjusted) Intrinsic Attitudinal Hedonism,[19] count as mental state theories on this definition. So does Heathwood's Subjective Desire Satisfactionism.[20] For on all these theories, a difference in the levels of well-being that two people have requires a difference in their mental states. In other words, they imply that mental duplicates are going to be duplicates with respect to well-being as well. All such theories are going to be entirely response dependent theories. After all, on mental state theories one's level of well-being supervenes on one's mental states, and so on mental state theories, it will be possible, just given the facts about your responses, to read off how much you would be benefited by any given state of affairs were it to obtain. For all that is needed to determine the welfare value of a state of affairs for you is a specification of what your mental states are in that state of affairs. Thus all the mental state theories are entirely response dependent.

However, the class of the entirely response dependent theories is not *exhausted* by the mental state theories. There are some theories that are not mental state theories, but that nonetheless are entirely response dependent, on my definition. For example, consider Actual Desire Satisfactionism (ADS). According to ADS, a person is well-off to the extent that his or her actual intrinsic desires are satisfied. Episodes of desire satisfaction

[16] One notable exception, though, is the Knowledge Theory of well-being, according to which your welfare is determined by only one thing: the number of true beliefs you possess.

[17] See, for instance, Kagan, 1998, ch. 2.

[18] Notice that the mental state theories are defined in terms of levels of welfare, while the entirely response dependent theories are defined in terms of the welfare value of states of affairs. This difference matters, and we'll see why in a moment.

[19] Cf. Feldman, 2004, chapter 4

[20] Cf. Heathwood, ms

are the items that enhance welfare according to ADS. Now, to see that ADS is not a mental state theory, suppose that I am currently drinking a Carlsberg and have a desire for this to happen. Moreover, suppose that my identical twin has a desire for a Carlsberg that is equally strong as my desire, but although he *believes* he is currently drinking a Carlsberg, he is in fact hallucinating the whole thing. Thus my desire is satisfied while his is not. ADS implies that my level of well-being is higher than my twin's. But this is not because of any difference in our mental states. We're supposing that my twin and I are mental duplicates. Rather, the difference in well-being between us is the result of a non-mental difference in the conditions of our lives. Thus ADS is not a mental state theory.

Even though ADS is not a mental state theory, it does still count as an entirely response dependent theory on my definition. And this is where we see the importance of the phrase 'if actual' in the definition D1. (Promissory note filled.) The reason that ADS counts as an entirely response dependent theory is that it implies that the welfare function – i.e. the function from a state of affairs, S, to the welfare value that S, *if actual*, would have for you – needs to be supplemented by only one thing: viz. the facts about the strengths of the desires you have in S. Given a state of affairs, all that we need in order to determine how much you would be benefited or hurt by that state of affairs if it were actual is a report of the strengths of the desires that you have in that state of affairs.[21]

To see this, consider, for example, the state of affairs consisting of your getting your desire for a Carlsberg satisfied. As before, call this state of affairs '$S_{satisfied}$.' According to ADS, all we need to do to figure out what the welfare value of $S_{satisfied}$ would be for you, if it were actual, is to look at how strong your desires are in that state of affairs – more specifically, how strong your desire for a Carlsberg is. Thus supposing that $S_{satisfied}$ were actual, the facts about the strengths of your responses suffice, according to ADS, for determining the welfare value of $S_{satisfied}$ if actual. By contrast, consider a different state of affairs, namely the state of affairs consisting just of your drinking a Carlsberg. As before, call this second state of affairs 'S_{Carl}'. Note that this state of affairs is such that

[21] A more sophisticated version of Desire Satisfactionism is Actual Concurrent Desire Satisfactionism (ACDS). The welfare value of a state of affairs, S, for you would be a function solely of the strength of your desire for S at the time that S actually obtains. And so ACDS, too, would be an entirely response dependent theory.

you have no desires in it. According to ADS, S_{Carl} is not capable of enhancing your welfare. The sort of thing that enhances welfare on ADS is episodes of desire satisfaction. Thus S_{Carl} would have no welfare value for you, even if actual. But this too can be read off from the facts about the strengths of your responses in S_{Carl}. In particular, since S_{Carl} is not a state of affairs in which you have any desires – after all, it just consists of your drinking a Carlsberg (not your drinking *and desiring* one) – we can read off that, according to ADS, S_{Carl} would have zero welfare value for you even if actual.

Thus it should be clear that ADS will count as an entirely response dependent theory. States of affairs are either such that they contain some desires or they are not. And in either case ADS implies that given a state of affairs, S, all you need in order to figure out the welfare value of S for a person is a specification of that person's responses in S. Thus according to ADS, the function from a state of affairs, S, to the welfare value that S would have for you if actual needs to be supplemented by only one thing: viz. the facts about the strengths of the desires you have in S. And so ADS is an entirely response dependent theory.

Similar considerations show that other versions of desire satisfactionism are also entirely response dependent theories. For instance, according to Ideal Desire Satisfactionism (IDS), to determine how much you would be benefited by obtaining a given end, all that is needed is a report of how strongly you *would* desire to obtain that end if you were placed in ideal conditions of one kind or another (say, if you were fully informed, or if you underwent cognitive psychotherapy). The desires you would have if you were placed in ideal conditions may be called your 'ideal desires.' What enhances your welfare on IDS are episodes of ideal desire satisfaction – i.e. states of affairs consisting of a) the fact that you would desire X under ideal conditions, and b) the fact that X obtains. Thus on IDS, given a state of affairs, S, we would need only one thing in order to determine the welfare value that S would have for you if actual: viz. a specification of what your ideal desires are in S. On the one hand, S might be a state of affairs consisting in part of your having certain ideal desires. In this case, we can obviously read off from the facts about the strengths of the ideal desires you have in S what impact S would have on your welfare if S were actual. On the other hand, S might not be a state of affairs consisting of your having any ideal desires. And in this case too we can read off what impact S would have on your welfare provided it were actual. After all, since you have no ideal desires in S, S would have zero impact on your welfare if actual. Thus IDS, too, is an entirely response dependent theory. For according to IDS, the function from a state of affairs, S, to the welfare value that S would have for you if actual needs to be supplemented by only one thing: viz. the facts about the strengths of the ideal desires you have in S.

I have been claiming that theories like Desire Satisfactionism although not mental state theories, are still entirely response dependent theories. Could one object to this by pointing out that on such theories, to determine the welfare value of a state of affairs for you, we would also have to know something about something else besides the strengths of your responses are – in particular, whether the object of your desires actually obtains? That, after all, seems to be a fact about the external world, not a fact about your responses. So according to Desire Satisfactionism, won't the welfare value that a state of affairs, if actual, would have for you not be solely a function of your responses? And in that case wouldn't such a theory not qualify as an entirely response dependent theory?

It would be a mistake to think this. It is true that according to Desire Satisfactionism, in order to determine what your *actual level of well-being* is at a time, we need to know whether the state of affairs you desire is actual. However, Desire Satisfactionism is still an entirely response dependent theory. I defined the entirely response dependent theories as the ones on which the function from a state of affairs to the welfare value of that state of affairs, if actual, for a person will have to be supplemented by only one thing: viz. the strengths of that person's responses in that state of affairs. This is indeed the case for Desire Satisfactionism. After all, given any possible state of affairs, all we need to know in order to figure out how much that state of affairs would impact your welfare, if actual, is the strengths of your desires in that state of affairs. Of course, just being able to figure out how much your welfare *would* be enhanced or decreased by various states of affairs is not going to be sufficient to determine your *actual* level of welfare. For that we would also need to know which states of affairs are actual. But this doesn't mean that Desire Satisfactionism isn't an entirely response dependent theory. It just means that it isn't a mental state theory.

The difference is subtle but important. Mental state theories are the ones on which a specification of your responses (at a time) would be sufficient to determine your actual level of welfare (at that time). The entirely response dependent theories, by contrast, are the ones that imply merely that if we are given a particular possible state of affairs, then a specification of your responses in that state of affairs is all that is needed to determine the impact that this state of affairs *would* have on your welfare *if it were actual*. Of course, according to an entirely response dependent theory, a specification of your responses will not guarantee that we can figure out what your actual level is. This might be the case on some entirely response dependent theories (e.g. Sensory Hedonism), but not all (e.g. Desire Satisfactionism). Thus the class of entirely response dependent theories is going to include more than just the mental state theories.

So far I have discussed some theories that count as entirely response independent and some theories that count as entirely response dependent. But what about the *partly* response independent theories? Which theories belong in this category? Let's look at three examples. First, consider a simple hybrid theory of well-being according to which your well-being depends on two things: the amount of sensory pleasure minus pain you experience and the number of meaningful relationships you have with other human beings.[22] This hybrid theory is not an entirely response dependent theory. For on this theory, the degree to which you would be benefited by a given state of affairs, if actual, cannot be read off solely from the facts about your psychological responses. Given a state of affairs, S, and a complete specification of your responses in S, it would still not be possible to say how much you would be intrinsically benefited by S. According to this hybrid theory, something else in addition to the facts about your psychological responses is determinative of what the welfare value of S would be for you if actual: namely, the number of meaningful relationships you have with other human beings in S. Thus this hybrid theory would count as a partly response independent theory. After all, it implies that the degree to which S (if actual) would intrinsically impact your welfare is determined *in part but not entirely* by the strengths of your psychological responses in S; something else in addition to this is required to determine how good S would be for you.

[22] Note that this theory is not a mental state theory since, according to it, a difference in well-being doesn't *require* a difference in mental states.

Next, consider a more plausible theory of well-being: Feldman's Desert-Adjusted Intrinsic Attitudinal Hedonism (DAIAH). This is the theory, roughly, that you are well off to the extent that you take attitudinal pleasure in things that are more pleasure-worthy.[23] For starters, note that DAIAH is not a mental state theory. Because DAIAH makes the welfare value of an episode of pleasure depend not only on its intensity and duration but also on the pleasure-worthiness of its object (an objective fact about the episode of pleasure), there could be two people who are mental duplicates but who still have different levels of well-being. For instance, suppose that my twin, who is a mental duplicate of me, and I are both experiencing a certain large amount of attitudinal pleasure right now. I am taking pleasure in the fact that my child has won an Olympic gold medal, while my twin, although he feels just as much pleasure as I do, is merely hallucinating that his child has won an Olympic gold medal. Presumably, its really being the case that one's child has won a gold medal is something much more pleasure-worthy than its merely *seeming* to one as if one's child has won a gold medal.[24] Thus my episode of pleasure would enhance my welfare more than my twin's episode of pleasure would enhance his, even though the two episodes are qualitatively identical.[25] So DAIAH implies that two people who have precisely the same mental states may nonetheless differ with respect to well-being. The upshot is that DAIAH is not a mental state theory. (Another version of Hedonism that Feldman discusses, Truth-Adjusted Intrinsic Attitudinal Hedonism[26], would for similar reasons not be a mental state theory either.)[27]

[23] See Feldman, 2004, p. 121. (For a more complete statement of the view, see chapter 4 of this dissertation).

[24] See Feldman, 2004, pp. 121-122

[25] I am assuming here that the qualitative indistinguishabilty of two people's mental states is sufficient for saying that those two people have the *same* mental states. This might not be quite right. Strictly speaking two people cannot literally be in one and the same mental state. If there are two people, then there must be two numerically distinct mental states. So it would be more accurate to say that if two people have qualitatively indistinguishable mental states, then they are both in the same *type* of mental state, but the tokens are different. Accordingly, I could still say that the two people, in virtue of being in qualitatively indistinguishable mental states, are *mental duplicates*. This would be sufficient for present purposes.

[26] Cf. 2004, pp. 112-114

[27] It should be noted that Mill's Qualitative Hedonism, under the standard interpretation, would *indeed* count as a mental state theory. On Mill's theory (cf. Mill, 2001, ch. 2), a person is well-off roughly to the extent that he has the experience of pleasures that are of a spiritual or intellectual nature. Under the standard interpretation of Mill (see, e.g. Feldman 2004, p. 73), this means that a person's well-being is determined solely by the qualitative features of the pleasures and pains that he or she experiences – i.e. *what it is like* to experience these pleasures and pains. Thus, for instance, Mill would be committed to saying that a brain in a vat who experiences pleasure from a *hallucination* of reading 'The Tell-Tale Heart' would have the same level of well-being as a real human being who experiences the same amount of

More importantly, DAIAH is a partly response independent theory of well-being. Why? According to DAIAH, the welfare value of a given state of affairs, S, for a person, P, is not simply a function of the strengths of P's psychological responses in S. In order to determine the welfare value of S on DAIAH, it is not enough to know the facts about how *much* attitudinal pleasure P feels in S; one must also know the facts about the pleasure-worthiness of the object of the attitudinal pleasures P experiences in S. But the facts about pleasure-worthiness do not in any way depend on one's attitudes or other psychological responses. Accordingly, the welfare value that S would have for P depends on something else *in addition* to the strengths of P's responses in S. And so DAIAH is not an entirely response dependent theory. Rather, it is a partly response independent theory. According to DAIAH, the function from a state of affairs, S, to the welfare value that S would have for you if actual needs to be supplemented by something more than just the facts about the strengths of the responses you have in S. In particular, this function must be supplemented by the facts about the pleasure-worthiness of the objects of your attitudinal pleasures in S.

Finally, consider Aristotelian Perfectionism. As I understand it, Perfectionism is the view that one is well off to the extent that one's life resembles the perfectly good human life minus the degree to which S's life resembles the perfectly bad human life. Aristotelian Perfectionism, then, is the general Perfectionist claim combined with the Aristotelian view of what constitutes the perfectly good human life: namely, that a life resembles the perfectly good human life to the extent that the one whose life it is realizes the properties (whatever they are) that are constitutive of human nature.[28]

Why is Aristotelian Perfectionism a partly response independent theory? For one thing, responses matter to welfare somewhat according to Aristotelian Perfectionism. Insofar as many of the properties that are constitutive of human nature are mental properties like intelligence, courage, wit, etc., the facts about what responses one has in a given state of affairs (one's beliefs, one's intentions, etc.) are going to play *some* part in

pleasure from *really* reading 'The Tell-Tale Heart.' As a result, Mill's Qualitative Hedonism, would be a mental state theory because it implies that there can be no difference between two peoples' levels of well-being without there also being a difference in their mental states. This establishes that Mill's theory is entirely response dependent too.

[28] And presumably, this can be extended to include the claim that a life resembles the perfectly *bad* human life to the extent that the one whose life it is "goes wrong" with respect to these same properties.

determining the welfare value of that state of affairs for one. Nonetheless, the welfare value of a given state of affairs, S, for a person, P, cannot be read off solely from the facts about P's psychological responses in S. After all, Aristotelian Perfectionism has it that realizing the properties that are constitutive of human nature (whatever they are) is going to benefit a person's well-being irrespective of what one's responses are. For example, my coming to more perfectly display the intellectual virtues would enhance my welfare even if I didn't enjoy[29] it, desire it or have any other positive response towards it. Thus Aristotelian Perfectionism implies that there is no function *just* from the facts about a person's responses and a given state of affairs to the welfare value that the state of affairs would have for that person. The welfare benefit that a person would receive from a given state of affairs (if actual) is *also* going to be determined by some facts *in addition* to the facts about this person's responses: viz. general facts of the form 'Person P instantiates characteristically human property p to degree x at time t'. Thus Aristotelian Perfectionism in going to count as a partly response independent theory.

Having now explained the three basic types of theory in my taxonomy, let me give a graphic representation of where various theories of well-being are to be located in the taxonomy:

[29] I believe Aristotle does say some things to the effect that enjoyment is going to be a necessary by-product of virtuous activity, but this does not conflict with the claim I am making here. After all, the enjoyment that one gets from virtuous activity will presumably have a positive impact on my welfare that is independent of the welfare enhancement I get from acquiring the virtues. Perhaps my welfare would be enhanced *more* if I also enjoyed my virtuousness. But becoming more virtuous by itself would presumably enhance my welfare too.

Entirely Response Dependent	*Partly Response Independent*	*Entirely Response Independent*
• Sensory Hedonism • Mill's Qualitative Hedonism • Intrinsic Attitudinal Hedonism • Actual Desire Satisfactionism • Ideal Desire Satisfactionism • Self-regarding Desire Satisfactionism (see Parfit) • Global Desire Satisfactionism (see Parfit) • Heathwood's Subjective Desire Satisfactionism • Sumner's Authentic Happiness Theory(?) • Raibley's Hybrid Theory • The Knowledge Theory	• Some hybrid theories (like the Pleasure-and-Friendship Theory) • Feldman's Desert-Adjusted Intrinsic Attitudinal Hedonism • Truth-Adjusted Attitudinal Hedonism • Aristotelian Perfectionism • Adams's theory • Parfit's theory • Scanlon's theory • Darwall's theory • My own theory	• Some Objective List Theories • The Money-in-the-Bank Theory • The Friendship Theory • The Power Theory

Figure 1: A New Taxonomy

I have explained why most of these theories are placed where they are in the taxonomy. But some of them I have not. I haven't said anything about Adams', Parfit's, Scanlon's or Darwall's respective theories. Nor have I had occasion to present Heathwood's Subjective Desire Satisfaction, Global Desire Satisfactionism, or Sumner's Authentic Happiness theory,[30] or Raibley's Hybrid Theory. Assuming that the reader is

[30] I do want to take a second to argue for this one. Sumner's theory has a decidedly subjective feel, but one might think that Sumner's Autonomy Requirement makes his theory count as a partly response independent theory. This might be seen as a problem for my taxonomy. In particular, it would undermine my claim that the subjective-seeming theories will all be in the entirely response dependent category.

However, I don't think Sumner's theory actually counts as a partly response independent theory. For reasons I will explain later, here is what I take Sumner's theory to be, fully stated:

S's level of welfare at t equals X iff

1) S is happy (i.e. satisfied with his life) at t to degree X,
2) S's happiness is informed, and
3) S's happiness is autonomous.

familiar with these theories, I will not take the time now to explain why these theories are categorized as they are. However, I will have occasion to discuss most of these theories at length in later chapters of my dissertation, and when I do, I will explain why each one has been placed where it is in the taxonomy. With this promissory note, I conclude my presentation of my favored taxonomy of theories of well-being.

2.4 Putting the Taxonomy to Work

I am not interested in taxonomy for its own sake. Categorizing theories is great fun, to be sure, but it would be just a frivolous exercise unless the taxonomy gets us somewhere. I think my taxonomy gets us somewhere. In particular, I will use it to argue that the theories in two of the three main categories in my taxonomy are problematic and can be dispensed with. The entirely response independent theories all share a common defect, while the entirely response dependent theories suffer from other problems. Not only do the mental state theories (an important sub-class of the entirely response dependent theories) all suffer from a particular problem, but the whole category of entirely response dependent theories seem to conflict with widely shared intuitions. And so I will conclude that we have good reason to believe that the true theory of well-being is a partly response independent theory.

S is happy (satisfied with his life) to degree X iff 1) S judges that S is satisfied to degree Y with S's life (i.e. that it is meeting S's expectations), 2) S feels to degree Z that S's life is "rewarding or worthwhile", and 3) X equals the average of Y and Z. Moreover, S's happiness is *informed* iff there is nothing that S doesn't know that would alter S's degree of satisfaction with his life were he to come to know it. Finally, S's happiness is *autonomous* iff S's happiness is based on values that are S's own, reflectively endorsed, not arrived at by unnatural external pressures (brainwashing, indoctrination, etc.), or something of this sort.

Given this statement of Sumner's theory, it seems to be an entirely response dependant theory. After all, it implies that the degree to which any state of affairs, S, would benefit a person, P, is solely a function of the facts about P's psychological responses in S. In particular, the value of S for P depends how happy P is in S, and whether this happiness is based on full information and autonomous (i.e. reflectively endorsed, not caused by external pressure). Whether P's happiness is fully informed and autonomous, I think, is a fact about the responses that P would have if he were made ideal in certain respects. Thus Sumner's theory will count as an entirely response dependant theory for the same reasons that Ideal Desire Satisfactionism counts as an entirely response dependent theory. For according to Sumner's theory, the function from a state of affairs, S, to the welfare value that S would have for you if actual needs to be supplemented by only one thing: viz. the facts about the strengths of the responses you would have in S provided you were made more ideal with respect to information and autonomy. Thus Sumner's theory counts as an entirely response dependent theory.

2.4.1 Against the Entirely Response Independent theories

Let us begin with the least plausible category of theories of well-being: the entirely response independent theories. All the theories in this group suffer from the same fundamental defect: they cannot account for the obvious fact that enjoyment is intrinsically beneficial to a person's well-being. That is, all entirely response independent theories conflict with

> *Intuition A*: Enjoyment of one form or another, whether it be sensory pleasure or attitudinal pleasure or happiness otherwise understood, in itself is good for a person (i.e. has a positive impact on that person's level of well-being).

This is an intuition with such centrality and strength as to be beyond question, in my view. It can be supported by all sorts of examples. Let us take the most generic one. For any two people, Abe and Babe, who are identical in every respect of their lives except that Abe enjoys his life more than Babe does, most everyone would agree that things are going better for Abe than for Babe. Intuition A seems to provide the best explanation for this. And so we have good reason to suppose that Intuition A is true. Moreover, the idea that enjoyment of one form or another intrinsically benefits a person is the basic insight that underlies all forms of Hedonism. Thus Intuition A has a long and respectable philosophical pedigree.

However, the thought that enjoyment of one form or another intrinsically benefits a person is one that no entirely response independent theory of well-being can accommodate. What characterizes these theories is that they imply that the degree to which a given state of affairs would (if actual) intrinsically impact a person's welfare is *not at all* determined by the that person's psychological responses in that state of affairs. So consider one state of affairs, S1, consisting of your going skiing in Vermont and hating every minute of it. Now consider another state of affairs, S2, that is identical to S1 except when it comes to the responses you have. In S2, unlike in S1, you greatly enjoy your skiing. Because the entirely response independent theories do not allow one's responses to play any role in determining the welfare value of a state of affairs for a person, they must imply that S1 and S2 would have the same intrinsic impact on your welfare. After all, S1 and S2 are identical in all respects that do not pertain to your responses. However, this implication conflicts with Intuition A. If one accepts Intuition

A, i.e. the claim that enjoyment is intrinsically good for a person, then one must allow that S1 and S2 would *not* have the same intrinsic impact on your welfare. But any theory that allows this would not be an entirely response independent theory. And so we see that all entirely response independent theories are incapable of accommodating Intuition A.[31]

Given the strength of Intuition A, the entirely response independent theories of well-being should be abandoned. Kagan, Sumner and others mention the requirement that the true theory of well-being must be *descriptively adequate* [32] – that is, consistent with our most deeply held intuitions about the concept of well-being. It should be clear from the foregoing discussion that no entirely response independent theory has any chance of being descriptively adequate.

2.4.2 Against the Mental State Theories

Entirely response dependent theories of well-being, by contrast, can account for Intuition A. Thus they might seem to have more going for them. However, they face other problems. First I will argue against one important sub-class of the entirely response dependent theories, viz. the mental state theories, which seem to be especially problematic. Then I will go on to present an argument against the whole category of entirely response dependent theories.

The problem with the mental state theories is that none of them can accommodate:

> *Intuition B*: Two people who are mentally indistinguishable in every respect (i.e. who are mental duplicates) can still have different levels of well-being.

This intuition, too, has a great deal of strength and prevalence. It is supported by famous thought experiments. Experience machine scenarios,[33] for instance, illustrate that mental duplicates might differ with respect to welfare. Suppose Abe leads a life of luxury and

[31] Perhaps one will object that this argument is too far-reaching. Not only does it impale the entirely response independent theories, you might think that an entirely response *dependent* theory like Desire Satisfactionism can't accommodate Intuition A *either*. However, while it's true that Desire Satisfactionism doesn't allow that specifically *pleasure* has intrinsic welfare value, it does have the resources to explain why a state of affairs like S1 is much worse than a state of affairs like S2. Thus Desire Satisfactionism is in much better shape than any of the entirely response independent theories. What's more, since Desire Satisfactionism allows that one's responses can impact one's welfare, it seems to be in a fairly good position to account for the intuitive appeal of Intuition A. At least, it has many more resources with which to accommodate Intuition A than the entirely response independent theories do.

[32] See Kagan, 1998, ch. 1; Sumner, 1996, ch 2.

[33] Nozick, 1974, pp. 42-45

success in the real world. Abe's twin, Babe, leads an experientially indistinguishable life, but inside the experience machine. Most would agree that Abe's life is the better one, even though Abe and Babe are mentally indistinguishable. If one agrees with this judgment (and I, for one, do agree), then one must accept Intuition B. Take another example, similar in spirit but this time not involving the experience machine. Compare the life of Caesar (the Roman emperor) and the life of a delusional person, Seesar, who roams the streets thinking he is Caesar. By an astounding coincidence, Seesar happens to hallucinate every single one of Caesar's actual experiences, though in reality Seesar just humiliates himself. Caesar and Seesar are mentally indistinguishable. The have precisely the same experiences. But there is great intuitive support for the claim that Caesar's life is much better for him than Seesar's life is for him. Accepting this judgment about this case commits one to accepting Intuition B.

No mental state theory of well-being, however, is capable of providing these results. On any mental state theory, a difference with respect to well-being between two people requires a difference in their mental states. And so any theory with the implication that Abe is better off than Babe, or that Caesar is better off than Seesar, would not be a mental state theory. Thus no mental state theory can accommodate Intuition B. Given the strength and prevalence of Intuition B, the mental state theories of well-being should be abandoned as well. Just like the entirely response independent theories, the mental state theories also have no chance of being descriptively adequate. Accordingly, they too may be dispensed with.

The upshot is that the true theory of well-being must either be an entirely response dependent theory that is not a mental state theory, or else it must be a partly response independent theory. These are the only theories capable of being descriptively adequate. After all, these are the only theories with the resources to accommodate both Intuition A and Intuition B.

2.4.3 Against the whole category of Entirely Response Dependent theories

There is good reason to think that the true theory of welfare is not to be found among the entirely response dependent theories, however. That is because this category of theories as a whole suffers from a certain sort of problem. As before, the problem is that they do not seem to be descriptively adequate. A range of widely shared intuitions support two basic claims that would conflict with all the entirely response dependent theories. I call these the Objectivist Claims.

> *OC1:* It is possible for there to be states of affairs that do not in any way involve or depend on a particular person's actual or counterfactual responses, but that nonetheless would *negatively affect* that person's well-being. (That is, either these states of affairs negatively affect well-being directly, or else they do so indirectly, by modifying the direct impact of other states of affairs on one's well-being.)
>
> *OC2:* It is possible for there to be states of affairs that do not in any way involve or depend on a particular person's actual or counterfactual responses, but that nonetheless would *positively affect* that person's well-being. (That is, either these sates of affairs positively affect well-being directly, or else they do so indirectly, by modifying the direct impact of other states of affairs on one's well-being.)

The phrase in parentheses here is supposed to indicate that there are two main ways in which each of the Objectivist Claims could be true. Most straightforwardly, they would be true if some states of affairs that do not involve anybody's responses could impact well-being independently of one's responses. More precisely, if it were possible for some states of affairs not involving the responses of a person, P, to directly enhance or decrease P's level of well-being *no matter what P's responses are*, then OC1 and OC2 would be true. Second (and perhaps more plausibly) OC1 and OC2 would also be true if there were some non-response-based states of affairs that could *alter* or *modify* the degree to which other states of affairs that do involve responses would directly increase or decrease one's well-being. For example, suppose the facts about the pleasure-worthiness of various objects (facts which do not depend on any particular person's responses) were capable of modifying the degree to which episodes of pleasure directly enhance well-being.[34] If this were the case then, the pleasure-worthiness facts would affect well-being in the sense that altering the pleasure-worthiness facts would either increase or decrease the degree to which a given episode of pleasure *directly* enhances a person's well-being. This, too, would imply that there are some states of affairs not involving responses that nonetheless negatively affect or positively affect well-being (albeit in an indirect way). And so the Objectivist Claims would be true.

[34] It should be clear that this is the way in which DAIAH works.

The Objectivist Claims are likely to be more controversial than Intuition A and Intuition B. But I think many will find them to be plausible enough to cast serious doubt on the entirely response dependent theories as a group.[35] I will give the case for the Objectivist Claims in a moment. First, however, note that the entirely response dependent theories cannot accommodate these claims. What characterizes the entirely response dependent theories is that given some state of affairs, the degree to which one would be benefited by it (if actual) is determined *solely* by the facts about one's responses in that state of affairs – i.e. by which responses one has (would have), how strong these responses are (would be). However, both Objectivist Claims imply that some facts *other* than the facts about one's responses are part of what determines how much certain states of affairs would impact one's welfare.

Begin with OC1. If there were states of affairs not involving your responses that could negatively affect your well-being – either independently of whatever positive responses (like pleasures or desires) you might have, or by decreasing the well-being contribution that your positive responses would have – then this would mean your well-being is determined in part by something *other* than just your responses. It would mean that given a state of affairs, S, to figure out the welfare value of S for a person, we would need to know something over and above the strengths of the responses that this person has in S. Similarly for OC2. If some states of affairs not involving your responses could positively affect your well-being – e.g. independently of your negative responses (like pains or aversions), or by decreasing the well-being reduction that these responses would have – then this would mean your well-being is determined in part by something other than your responses. Again, it would mean that given a state of affairs, S, in order to figure out the welfare value of S for a person, we would need to know something more than just the strengths of the responses that this person has in S. Thus both OC1 and OC2 would imply that the facts about your responses are not all that determine the degree to which any given state of affairs would impact your well-being. But there is no entirely response dependent theory that is compatible with this being the case. For entirely response dependent theories make the impact of a given state of affairs on one's well-

[35] At the very least, these they are plausible enough to generate a demand that defenders of entirely response dependent theories tell some plausible story to explain away the intuitive support for these claims.

being depend exclusively on the strengths of one's responses (or how strong they would be if one were made more ideal somehow).

This matters, however, only if the Objectivist Claims are likely to be true. Are they compelling? I think so. I will focus primarily on OC1. For if one accepts it, then one should not have much trouble accepting O2 as well. The first objectivist claim is supported by all sorts of well-known examples. To make my point, I will focus on two cases offered, among others, by Fred Feldman. First consider the case of Porky,[36] which in turn is based on a passage in Moore.[37] Feldman presents the case as follows:

> Imagine a person – we can call him 'Porky' – who spends all his time in the pigsty, engaging in the most obscene sexual activities imaginable. I stipulate that Porky derives great pleasure from these activities and the feelings they stimulate. Let us imagine that Porky happily carries on like this for many years. Imagine also that Porky has no human friends, has no other sources of pleasure, and has no interesting knowledge. Let us also that Porky somehow avoids pains – he is never injured by the pigs, he does not come down with any barnyard diseases, he does not suffer from loneliness or boredom. (Feldman, 2004, p. 40)

As Feldman presents the case of Porky, it is an objection to Sensory Hedonism in particular. It is supposed to show that some sensory pleasures do not enhance one's well-being, and might even have an overtly negative impact on it. However, the case of Porky can be expanded so that it threatens not just Hedonism, but entirely response dependent theories in general. In particular, we can add the following stipulations to the case. Imagine further that Porky has a strong desire to for his life in the pigsty, that he would continue to have these desires even in ideal conditions (e.g. with full information and rationality, and after undergoing extensive therapy), and in general that all his responses towards his lifestyle are positive and unalterable.[38] My suggestion is that if one finds the case of Porky as described by Feldman to provide a plausible objection to Hedonism, then one should find the expanded Porky case to provide a plausible objection to the entirely response dependent theories.

In this expanded version of the case, is Porky's life of porcine delight a good one for him? What makes the Bestiality Objection to Hedonism so memorable and forceful is the widely shared sense that Porky's life is *not* a valuable one for him, on balance, despite its pleasantness. Even if one admits that Porky's pleasure is indeed of some benefit to him,

[36] Feldman, 2004, pp. 38-41, pp. 118-119

[37] Moore, 1903, sect. 56

[38] To impale Sumner's theory as well, perhaps we should add that Porky's responses towards life in the pigsty are autonomous, i.e. not the result of brainwashing or other external intrusions.

Porky’s beastly lifestyle itself seems to negatively affect his well-being too. Thus the intuitive thing to say seems to be that Porky’s life is on balance not a good one for him, and might even have an overall negative value for him. If this is the intuitive judgment about the original version of the Porky case (which threatens just Hedonism), why should our estimate of the value of Porky’s life for him change just because we add some more details about what Porky’s desires are and would be? I see no reason (no non-theory-driven reason, that is) why it should.

If I am right, then the expanded Porky case speaks in favor of OC1. If Porky’s un-kosher life is on balance not a good one for him despite all his pleasure, all his satisfied desires and his other positive responses, then this means that there are some states of affairs *not involving responses* that negatively affect well-being. This could be true for one of two reasons. On the one hand, perhaps it is because Porky’s pig-loving is a sort of behavior that would have a directly negative impact on one’s well-being no matter what one’s responses are. On the other hand, perhaps it is because there are some facts that do not involve or depend on anybody’s responses (like the pleasure-worthiness facts) that serve to decrease whatever benefits Porky might receive from his pleasant experiences, his desire satisfactions or his other positive responses. I am not ready to take a stand on which one of these alternatives is the right lesson to draw from Porky’s case. But whichever it is, the case clearly supports the idea that there are some non-response-based facts that are able to negatively affect well-being. Thus the expanded Porky case supports OC1. This would refute the entirely response dependent theories.

The second case of Feldman’s that supports OC1 and casts doubt on the entirely response dependent theories is that of Max the Masochist.[39] Here is my adaptation of it. All that Max wants for himself is sickness, misery, humiliation and early death. Suppose that these desires would persist even under ideal conditions. No amount of information, rationality or psychotherapy could dislodge Max’s masochistic desires. Moreover, suppose he gets what he wants. He becomes extremely sick, and his sadistic doctors play cruel jokes on him until the very end. Instead of giving him a simple treatment that would

[39] I’m not sure where (or even if) this case appears in Feldman’s published works. However, Jason Raibley (2007, p. 253) and Ben Bradley (2008) discuss cases like this to point out a paradox for several influential theories of well-being. This is not the purpose for which I use the case here, however. Moreover, the Max the Masochist case is very similar to a case discussed by both Williams 1982, pp. 105-106, and Parfit 1997, p. 112.

improve his condition, they experiment on him with drugs that rot his body and mind into a horrible, utterly pathetic state of waste and confusion. This doesn't bother Max. Instead, he judges that his life is going just right, that everything is going according to plan. This is, after all, just what he always wanted.

This case, too, supports the first Objectivist Claim. As the case is described, Max has a number of different positive responses towards what is happening to him. Max has an actual desire for a wretched life, and what's more he has an ideal desire for it, since his desire would endure even if he were given full information, rationality, psychotherapy and so on. Furthermore, Max makes positive judgments about the way things are going for him. In spite of these positive responses, however, it should be clear that Max's life is quite a bad one for him. This shows that other things besides one's judgments or one's desires (whether actual or ideal) help determine the degree to which certain states of affairs impact one's well-being.

Could a supporter of the entirely response dependent theories respond by saying that 'this other thing' is pain? That is, could someone who is sympathetic to the entirely response dependent theories argue with respect to the case of Max that we take his life to be so bad, even though his desires are satisfied and his judgments are positive, *because he feels so much pain*? This response will not ultimately work. For we can simply add a few more details to the case to block the reply. Let us stipulate that Max's life is not painful or sad – although it is wretched, pathetic, and, yes, even miserable (at least in the non-psychological sense of being shameful or worthy of pity). Does this additional stipulation alter our overall evaluation of Max's life? It seems not. For Max's life is still awful. He is abused, humiliated, pathetic, wretched and befuddled. Thus in the case as described, Max's life seems to be a bad one for him despite his complete lack of pain *and* despite all the positive responses he has towards the things happening to him. Thus the case of Max the Masochist provides yet more support for OC1.

In general, what lends OC1 its plausibility is the thought that things like mental decay, complete powerlessness, being utterly contemptible, being subjected to massive humiliation or exploitation, and so on, are all things that would have some kind of negative effect on a person's welfare – either directly (i.e. independently of whatever one's responses are) or indirectly, by modifying the benefits that do depend on one's responses. Even if one desired these things, or would desire them under ideal conditions, or would take pleasure in them, they would still in one way or another negatively affect one's well-being. After all, most people would prefer, all other things being equal, a life that contains less of the items mentioned above. Suppose one is given the choice between two lives that contain exactly the same amount of pleasure and pain, desire satisfaction and frustration. They are, moreover, identical in terms of all other responses too (e.g. beliefs, judgments, etc.). The only difference between them is that the one life contains significantly more humiliation, exploitation, decay, failure and general pathetic-ness than the other. Which one would you pick? Most people, I suspect, would prefer the life that contains less of these things. And this is at least some evidence[40] that these things, in one way or another, exert some negative influence on one's well-being over and above the contributions to well-being that are a function of one's responses.

When it comes to the second Objectivist Claim, I think that one should have no problem accepting it if one is already willing to accept the first Objectivist Claim. If you are willing to allow that some states of affairs not involving responses can have a negative effect on well-being, then it would be arbitrary not to allow that some non-response-based states of affairs can have a positive influence on well-being. Just as humiliation, exploitation, pathetic-ness and decay seem to negatively influence one's well-being even if one is pleased by or desires these things (or would do so), it seems that things like great achievements, the attainment of excellence, and the realization of beauty can similarly exert some kind of *positive* influence[41] on a person's well-being even if he has some negative response, like aversion or displeasure, towards these things. Perhaps having one's desire not to be great and excellent frustrated involves some decrease in well-being; perhaps taking displeasure in one's greatness and excellence does so as well. But it seems hard to deny that greatness and excellence would also in themselves exert at least some kind of positive influence on one's well-being. Perhaps this positive influence

[40] I have already argued for this assumption in chapter 1 of the dissertation. Preferences between lives do not provide a foolproof test for something's having an impact on one's welfare (since things that are not related to welfare may have an impact on one's preferences between lives), but it does, I think, still provide some evidence in favor of such an impact.

[41] I hope it is clear from the foregoing discussion that by 'positive influence' I don't *necessarily* mean a direct raise in one's level of well-being. That is just one of the two things this phrase might mean. In addition, it might also mean an upward modification to the well-being contributions of other response-based states of affairs (like being pleased or having a satisfied desire).

is direct and involves a change to one's well-being that is entirely independent of one's responses. Or perhaps the positive influence is indirect and occurs by modifying the direct well-being contributions states of affairs involving responses. Whichever it is, OC2 would be vindicated. States of affairs that are constituted by one's greatness or excellence, but which do not in any way involve one's responses, would nonetheless be capable of exerting some kind of positive influence on one's well-being. At the very least, one would be hard pressed to deny that this is the case if one already accepts OC1, which we have already seen there is significant intuitive support for.

Insofar as the two Objectivist Claims are compelling, there is reason to think that the correct theory of well-being is not of the entirely response dependent sort. Since there is intuitive support for thinking that some states of affairs not involving anybody's responses can exert some positive or negative influence on well-being, the entirely response dependent theories have got little chance of being descriptively adequate. For the these theories are characterized by their making the welfare value of a given state of affairs be *solely* a function of the facts about one's responses. Thus unless the entirely response dependent theories have some way of explaining away the intuitive support for OC1 and OC2, we must look to the partly response independent theories to find the correct theory of well-being.

Partly response independent theories, after all, are able to account for OC1 and OC2. For the theories in this category allow that certain things *besides* the facts about one's responses help determine the degree to which a given state of affairs would benefit a person's well-being. There are many ways a theory in this category might account for these intuitions. For example, consider a theory that is hybridized in such a way that one's well-being is equal to the sum of 1) the well-being contributions of a certain group of one's responses (say, one's pleasures) and 2) the well-being contributions of certain objective facts (say, the intrinsic value of one's accomplishments, or the number of meaningful friendships one has with other people). Such a theory would capture the idea behind the Objectivist Claims, namely the idea that some states of affairs that do not involve or depend on anybody's responses can nonetheless exert some positive or negative influence on one's well-being. Moreover, this same idea can be captured by another sort of partly response independent theory: namely, a monistic theory that takes the degree to which certain responses impact one's well-being to be *modified* by certain non-response-based facts. For instance, Desert-Adjusted Intrinsic Attitudinal Hedonism allows the impact of one's (attitudinal) pleasures to be modified by the non-response-based facts about the pleasure-worthiness of the objects of this pleasure. So this theory too is compatible with the idea that some states of affairs (namely those involving pleasure-worthiness) can influence well-being in a way that goes over and above the welfare contributions of one's responses. Since the partly response independent theories can account for OC1 and OC2 in ways like this, there seems to be good reason to prefer theories of this sort over the entirely response dependent theories.

2.4.4 Attempts to Explain Away the Problematic Intuitions

The fat lady hasn't sung yet, however. Might supporters of the entirely response dependent theories have strategies at their disposal for explaining away the intuitive support for OC1 and OC2? I am able to think of two such strategies.

First, those who sympathize with the entirely response dependent theories might claim that the intuitions underlying OC1 are illusions created by the common feeling that the *unstable* absence of pain – i.e. an absence of pain that is not likely to endure in the long run – is undesirable and to be avoided. Taking this line would allow one to say, about the Porky case, that even though it was stipulated that Porky doesn't get sick or lonely and feels no pain from his escapades in the mud, this is a mere fluke. Similarly, in the case of Max, it was stipulated that although he was pathetic, exploited, humiliated and utterly wretched, he experienced no sadness or distress. But this too is just a lucky accident. If cases like Porky's or Max's were to recur over and over again, it would be highly unlikely that the person in Porky's or Max's shoes would feel no pain or distress. The common desire to avoid situations in which the absence of pain is *unstable* in this way explains why it seems to us that Porky's life and Max's life are bad ones for them. Thus the supporter of the entirely response theories may account for our judgments about Porky and Max in a way that does not appeal to anything but the impact that responses would have on well-being. There is no need to assume that anything *besides* responses determines the degree to which states of affairs would be good or bad for people. So the proponent of the entirely response dependent theories has explained away the intuitions behind OC1.[42]

A second response on behalf of the entirely response dependent theories would be to say that in the cases of Porky and Max, we have the feeling that there are other sorts of lives easily available to Max or Porky that would be much more satisfying. If Porky or Max just got an education, say, they would lead lives that are by far superior to their actual lives, either in terms of overall pleasure, or overall desire satisfaction, or whatever your favorite response is. When reflecting on the cases of Porky and Max, we subconsciously compare the lives of Porky and Max as described with these other, clearly superior lives that are easily available to them. And this explains why we have the intuition that Porky's life and Max's life are bad for them. It's not that we really think that Porky or Max are leading lives with *negative* levels of well-being. Rather, we think their actual lives are bad *compared to* other lives they easily could have led instead. Thus the intuitions about Porky and Max that are needed in order to establish OC 1 are not sustainable on closer inspection.

Does either of these strategies on behalf of the entirely response dependent theories succeed in explaining away the intuitions that Porky's life and Max's life, despite all their positive responses, are really not valuable ones for them? I think not. I see no reason why we can't simply add more details to the cases of Porky and Max so as to block the entirely response dependent theorist's responses. Although it might make the case of Porky less realistic, suppose Porky lives in a world where his lack of pain in *not* unstable. Suppose in Porky's world, there is a guarantee that pigs and mud carry no diseases whatsoever, and so there is no chance that people who behave like Porky would ever feel any pain. Now the absence of pain in Porky's case is no longer unstable. But does adding this fact to the case alter our intuitive judgment that Porky leads a life that is not good for him? I don't see how it could. Next, suppose we stipulate that in Porky's world, it would be fantastically difficult for people in Porky's society to get an education and engage in any of the things that people like you and me usually take to be more worthwhile (e.g. conversation, companionship, music, poetry, philosophy, jet skiing, etc.). Now there is no obviously superior life that Porky could easily lead instead of his actual muddy one. But

[42] Something similar, I presume, could be said about the intuitions underlying OC2.

does adding this stipulation change our intuitive judgment about Porky's level of well-being? Again, I don't see why it should. Similar considerations apply in the case of Max. Suppose we flesh out the story of Max in such a way that his lack of pain is *not* unstable and that there is *no* obviously superior life he could easily lead instead of his actual wretched life. Adding these facts to the case would not seem to change our intuitive judgment about the value of Max's life for him.

Accordingly, I do not know of any way that supporters of the entirely response dependent theories could explain away the intuitions that underlie the Objectivist Claims. And since these two claims are not compatible with the entirely response dependent theories, I think we have good reason to abandon the entirely response dependent theories altogether. Instead, I suggest, we should look among the partly response independent theories to find a descriptively adequate theory of well-being.

CHAPTER 3

THE ARGUMENT FROM MOTIVATION

I am inclined to think that the truth about welfare is likely to be best captured by an objectivist theory (i.e. a theory according to which one's welfare is not determined solely by the various responses that one has to things). However, when I tell people this, the response is often skeptical. I hear one sort of objection particularly often: 'Objective theories imply that things can be good for you even if you don't care about them. But that's crazy. How can something be good for you if you wouldn't be at all moved to go get it? People *want* what's in their self-interest. So the objective theories must be false. They don't preserve the necessary connection that exists between motivation and a person's good.'

This argument is not new. Some philosophers have defended arguments in print along these very lines. Consider for instance the following line of argument offered by Peter Railton:

> It does seem to me to capture an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware. It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him. (Railton, 2003, p. 9)

Railton expresses his discontent with certain theories of a person's good (or of what I here call 'welfare' or 'well-being') on the grounds that they fail to preserve a necessary connection between one's good and what a person would find 'compelling' or 'attractive' or would 'engage' him. Connie Rosati develops and defends a version of this argument as well. In explaining the basic idea behind the argument, she says:

> an individual's good must not be something alien – it must be "made for" or "suited to" her. But something can be made for or suited to an individual, the thought goes, only if a concern for that thing lies within her motivational capacity: what is good for her must connect with what she would find "in some way compelling or attractive…" In this way, there must be a "fit" between an individual and her good. (Rosati, 1996, pp. 298-299)

Rosati goes on to refine this idea considerably, and argues that the true theory of welfare must be consistent with the claim that one's good must be linked to what would motivate one.[1] Dan Haybron, too, is aware of this sort of argument, and explains that 'subjectivism seems to respect the "internalist intuition" that an agent's well-being must connect appropriately with her motivational structure.'(Haybron, forthcoming)

What's more, I suspect that the idea that one's welfare must somehow depend on what is capable of motivating one, i.e. what one finds attractive and compelling, is what lies behind standard objections to the Objective List Theory of welfare. Chris Heathwood, for instance, claims that

> Objective list theories may seem unsatisfactory because they make it possible for a person who hates his life through and through nevertheless to have a good one. (Heathwood, 2006, p. 553)

Shelly Kagan expresses a similar concern:

> As we have seen, the objective theorist holds that possession of the objective goods makes one better off – regardless of whether or not one realizes this. This seems to have the implication that your life could be made better off by the possession of some "good" even though you yourself dislike it and would greatly prefer to be without it: since the good possesses objective value your own opinion on the subject is quite irrelevant. Your life could be going well even though you are unhappy with almost all its central features! Thus (…) the objective theory too seems to lead to unacceptable results. (Kagan, 1998, p. 40)

Of course, neither Heathwood nor Kagan explicitly mention *motivation* in these criticisms of the Objective List Theory. Nonetheless, their criticisms clearly are similar in spirit to the key idea in the argument we are concerned with here, namely the idea that a person's welfare must be somehow connected to what one cares about and would be moved by.

[1] For a more detailed discussion of Rosati's views, see Appendix I.

Thus some philosophers would seem to be sympathetic to the idea that since one's concerns and motives must be what determines one's welfare, the objective (i.e. response independent[2]) theories of welfare in particular are in trouble. So far, however, this argument from motivation remains highly unclear. What exactly is the connection between motivation and welfare that the objectivist theories allegedly cannot preserve? Do the subjective (i.e. response dependent) theories really do a better job of preserving it? Is the argument even capable of giving a reason to prefer one group of *first-order* theories of welfare over another? Or is the argument merely a metaethical argument to the effect that there are no objective facts about what would enhance one's welfare? In this chapter, I aim to sort out this confusion. I will argue that once the confusion is cleared up, it turns out that there is in fact no good argument from motivation to be made against the response independent theories and in favor of the response *de*pendent theories.

There are several ways to understand the argument from motivation. First, one might take it to be a Mackie-style argument against the claim that there are objective facts about welfare. This version of the argument, however, would purport only to establish a meta-ethical conclusion. What would be more interesting to us here (since this dissertation is concerned with normative ethics) would be a version of the argument that purports to disprove certain first-order theories of welfare. Here I will consider one version of the argument thus conceived. In particular, one might think that a widespread view about the motivational capacities of *reasons*, namely Internalism about normative reasons, generates an argument against the response independent theories of welfare. In section 3.2, I will formulate this version of the argument from motivation as precisely and as plausibly as I can. Then I will go on to argue that it in fact provides no reason to prefer the response dependent theories over the response independent theories.

One might also think that an argument from motivation could be constructed by appeal to a more modest view than Internalism about normative reasons, viz. by appeal to Internalism about a *person's good*. This more modest sort of Internalism has been defended by Connie Rosati, Peter Railton and David Velleman, among others. However,

[2] Throughout this chapter, I will use 'response independent theories' to cover both the *entirely* response independent theories and the *partly* response independent theories. I'm using 'objective theories' interchangeably with 'response independent theories'. 'Subjective theories', then, is used interchangeably with entirely 'entirely response dependent theories.'

I think Internalism about a person's good is *false* because of the very same sort of problem that undermines the version of the argument from motivation that I discuss in section 3.2. Therefore, I relegate my discussion of Internalism about a person's good to Appendix I.

3.1 The Argument from Motivation is not Metaethical

There are a number of ways to understand the argument from motivation that was loosely described in the introduction. One natural way to understand it would be as an analog to Mackie's metaethical argument for the claim that there are no objective moral facts. In particular, one might think the argument from motivation is supposed to show that there are no objective facts specifically about what would enhance one's welfare. However, as we'll see, this rests on a confusion about the proper target of the argument.

Mackie's original argument may be summarized as follows:

Mackie's Argument

1) If there are objective moral facts, then they are intrinsically motivating.
2) But no facts are intrinsically motivating.
3) Therefore, it's not the case that there are objective moral facts.[3]

A 'moral fact' should be taken to mean 'a fact to the effect that a particular act token is morally permissible/obligatory/wrong'. The claim that these facts are 'objective' amounts to saying that a proposition expressing such a moral fact is capable of being true or false. What's more, to say that the moral facts are intrinsically motivating is to make the following claim: necessarily if you know that it would be morally right (wrong) of you to φ, then you have at least some motivation (not) to φ, provided you are rational. The argument, then, is that since no facts can be intrinsically motivating in this sense, there can be no objective moral facts.

Now perhaps this provides a model for how to understand the argument from motivation as it applies to welfare. Suppose that by a 'welfare fact' we mean a fact to the effect that if a person, P, were to φ it would bring about a state of affairs that in itself enhances P's welfare. Saying that such facts are 'objective' would be to say that a

[3] See Mackie 1977, ch.1. This formulation of the argument is based on Michael Huemer's interpretation of Mackie. (Cf. Huemer, 2005, p. 139.)

proposition expressing such a welfare fact is capable of being true or false. What's more, to say that such facts are intrinsically motivating would amount to claiming the following: necessarily, if you became aware that it would enhance your welfare to φ, then you would have at least some motivation to φ, provided you are rational. Thus we could summarize the metaethical version of the Motivation Argument as follows:

<u>The Metaethical Motivation Argument</u>

1) If there are objective welfare facts, then they are intrinsically motivating.
2) But no facts are intrinsically motivating.
3) Therefore, it's not the case that there are objective welfare facts.

This argument might be worth discussing in its own right since one might think that the assumption it embodies about the motivational capacities of welfare facts is quite a bit more plausible than the assumption in Mackie's Argument about the motivational capacities of moral facts. That is, one might think it is more plausible to claim that

> a) being aware that some action would enhance your welfare (i.e. is in your interest) necessarily provides some motivation to do it, provided you are rational

than it is to claim that

> b) being aware that some action is morally required necessarily provides some motivation to do it, provided you are rational.

You might think that while rational people are necessarily motivated by considerations of self-interest,[4] they are not necessarily motivated by moral considerations. Thus one might think the Metaethical Motivation Argument has more going for it than Mackie's Argument.

Even if one does think this, however, the Metaethical Motivation Argument is of little importance if one's concern is first-order ethics (as ours is here). The metaethical issues raised by Mackie's argument have already been explored in quite some depth,[5] so what seemed particularly interesting about the argument from motivation sketched in the introduction was that it was supposed to provide a reason to reject the *first-order* theories

[4] That is to say, perhaps one thinks that a concern for one's self-interest is built into the concept of rationality. In that case, it would be a conceptual truth that rational people are necessarily motivated by considerations of self-interest.

[5] See, for instance, Brink 1984, Brink 1989 and Huemer 2005.

of welfare that are response independent[6] (or objectivist). However, the conclusion of the Metaethical Motivation Argument, even if it were true, would have no bearing on the relative merits of any first-order theories of welfare. Even if the correct second-order view is that the welfare facts aren't objective, then there might still be good reason to prefer some response independent theory, at the first-order level, over the entirely response dependent theories. By the same token, even if the welfare facts were objective, then there might still be substantive considerations that favor some entirely response *de*pendent theory of welfare, at the first-order level, over the response independent theories. In short, the metaethical status of welfare facts does not bear on the question of which first-order theory of welfare is preferable.[7] Thus the Metaethical Motivation Argument does not capture what seemed to be especially interesting about the argument from motivation as initially sketched.

3.2 A First-Order Argument from Motivation Based on Internalism about Reasons

How, then, are we to construct a version of the argument from motivation that really does purport to establish a first-order conclusion? Perhaps the most natural strategy would be to appeal to Internalism about normative reasons. This is roughly the view that in order for some fact to provide an agent with a reason to perform a certain action, it has to be possible for the agent to be motivated by a recognition of that fact to perform that action. This sort of Internalism is a popular view, and it might be thought to provide

[6] Again, I'm using 'response independent theories' here to cover both the *entirely* response independent theories and the *partly* response independent theories.

[7] One might object to this by claiming that if there are no objective welfare facts – i.e. if claims like 'X is good for P' can be neither true nor false – then all of the first-order theories of welfare would be false. After all, they purport to specify the truth conditions for claims of this sort.

If this is one's view, however, then I can dispense with the Metaethical Motivation Argument in a different way. In particular, what I could now say is that the Metaethical Motivation Argument, if sound, would constitute a problem for *every* theory of welfare. Thus this argument would not give any reason to prefer response dependent theories of welfare over the response independent ones. If this result is all I can get, then that's good enough for me. My main goal is just to show that there is no good argument from motivation for preferring the response dependent theories over the response independent theories.

(Nonetheless, I do actually think the Metaethical Motivation Argument is *unsound*. In particular, I reject the view – underlying premise 1 in the argument – that welfare facts, if there are any, would have to be intrinsically motivating. The view that the welfare facts are intrinsically motivating is equivalent to the view called Internalism about a person's good, which I discuss at length in Appendix I. I conclude that there is good reason not to think that this view is false. Thus I would reject premise 1 in the Metaethical Motivation Argument.)

reason to reject first-order theories of welfare of the response independent type in favor of theories of welfare of the response *de*pendent type. This is the strategy that I will be considering in the remainder of this paper. I will begin by constructing the most plausible version of the argument that I can, and then I'll go on to explain why I think the argument fails.

3.2.1 Stating the argument

Here, in very rough outline, is how this version of the argument would go. Two basic assumptions are required to start. First, we need the assumption that there is a connection between an action's enhancing one's welfare and there being a normative reason for one to perform that action. More specifically, the assumption is this:

> *Premise 1)*: The fact that P's φ-ing would produce an outcome with the features that enhance P's welfare is a normative reason for P to φ.

Next, we need an assumption to the effect that there is a connection between your having a normative reason to act and its being possible for *a recognition* of this reason to motivate you to act, i.e. that some version of Internalism about reasons is true. More specifically, the claim we need is this:

> *Premise 2)*: If fact F is a normative reason for P to φ, then a recognition of F is capable of motivating P to φ.

These two premises lead to the following preliminary conclusion:

> *Lemma*): A recognition of the fact that P's φ-ing would produce an outcome with the features that enhance P's welfare is capable of motivating P to φ. [1, 2, UI & MP]

One final premise, then, is needed to complete the argument, and that is the claim that the response independent theories of welfare are not compatible with the above lemma. More specifically, the claim is this:

> *Premise 3)*: If the true theory of welfare is response independent, then *it's not the case* that a recognition of the fact that P's φ-ing would produce an outcome with the features that enhance P's welfare is capable of motivating P to φ.

Thus we get the following

> *Conclusion*: It's not the case that the true theory of welfare is response independent. [Lemma, 3, MT]

This preliminary sketch of the argument obviously needs to be made precise, however. It is very unclear as it stands. A more detailed discussion of the argument's premises is in order.

Begin with premise 1), which said roughly that the fact that P's φ-ing would bring about an outcome with the features that enhance P's welfare is a normative reason for P to φ. First note that the claim being made here should not be a very controversial one.[8] After all, a normative reason for an action is roughly a fact or a consideration that counts in favor of performing that action. And many sorts of facts, it seems, could count in favor of an action: the fact that the action maximizes utility, the fact that the action treats nobody merely as a means, the fact that the action is courageous, the fact that the action would benefit your kids. Another sort of fact that naturally would seem to count in favor of an action has to do with the benefits that the action will provide for the one who performs it. The fact that an action will produce a large amount of pleasure for the agent, or the fact that it would satisfy one of the agent's desires, or the more general fact that the action would lead to the agent's welfare being enhanced – all these facts, it seems, could in principle count in favor of performing an action. And why not? The threshold for being a consideration that counts in favor of an action does not seem to be very high.

However, there is a crucial ambiguity in premise 1). On the one hand, premise 1) might mean this:

> a) what is a normative reason for P to φ is the fact that P's φ-ing would bring about an outcome, O, with the following *general* feature: P's welfare is enhanced in O.

On the other hand, premise 1) might mean this:

> b) what is a normative reason for P to φ is the fact that P's φ-ing would produce an outcome with the *specific* features (e.g. pleasure, desire satisfaction, money procurement, etc.) that the true theory of welfare picks out as the ones that in themselves enhance P's welfare.

If one is to make the argument from motivation at all plausible, however, then premise 1) cannot be interpreted along the lines of a). But this is not because a) is *false*. On the contrary, it seems quite plausible to claim that if P's φ-ing would lead to an outcome that

[8] Parfit, for instance, points out that many philosophers endorse the following connection between one class of normative reasons, the prudential ones, and well-being: '(P) We have a prudential reason to act in some way if and only if (S) this way of acting would promote our own well-being.' (See Parfit, 1997, pp. 108-109)

is good for P, then that is a normative reason for P to φ. Rather, it is because interpreting premise 1) along the lines of a) would force an interpretation of premise 3) in the argument that clearly *is* false. If we interpret premise 1) along the lines of a), then to preserve the validity of the argument, premise 3) would have to become the following claim:

> *P3')*: If the true theory of welfare is response independent, then it's not the case that any person P would be motivated to φ provided he recognizes the general fact that his φ-ing would enhance his welfare.

But this claim is just plain false. The response independent theories imply no such thing. Consider a toy response independent theory: viz. the Money Theory, according to which acquiring more money is the only thing that would enhance a person's welfare. Even if the Money Theory were true, we would still expect people in general to be motivated to perform actions *that they believe will lead to outcomes that enhance their welfare.* Even a person who doesn't care about money in the slightest would presumably still *do what he thinks would enhance his welfare*. Thus the truth of the Money Theory – or any other response independent theory of welfare for that matter – would not be a reason to think that people will fail to be motivated to perform the actions that they believe to be welfare enhancing for them. Thus the claim in P3') is false. So interpreting the argument along the lines of a) would do nothing to impugn the response independent theories of welfare.

Instead, we must use b) in formulating the argument. For one thing, premise 1) interpreted along the lines of b) is quite plausible in its own right. After all, it is quite easy for some fact to be a normative reason to act. So why *not* think, as premise 1) on this interpretation has it, that an action's producing an outcome with the specific features that the true theory of welfare picks out as good-making is a normative reason to perform that action? What's more, one might think that the rest of the argument remains plausible if interpreted along the lines of b) as well. In particular, premise 3) on this interpretation would become roughly this claim:

> *P3'')*: If the true theory of welfare is response independent, then it's not the case that an arbitrarily chosen person P would be motivated to φ by a recognition of the fact that P's φ-ing would produce an outcome with the *specific* features (whatever they are) that the true theory of welfare picks out as good-making for P.

This claim might seem plausible as well. After all, suppose the Money Theory is true. Moreover, suppose Jack doesn't care in the slightest about obtaining money. Thus, even though we're supposing the Money Theory to be true, Jack clearly will not have any motivation to do things that he knows will produce outcomes with the features that the true theory of welfare (viz. the Money Theory) picks out as welfare enhancing. After all, Jack does not care about getting more money.

What's more, one might think that the same goes for any response independent theory. Response independent theories make welfare depend on other things besides what one cares about or is moved by. And so it might seem that for any response independent theory, whatever features it picks out as good-making, there could be people who know that by φ-ing they could bring about outcomes with *these specific features* and yet remain completely unmotivated to φ. Thus premise 3) understood along the lines of b) – i.e. P3'') – might seem to be quite plausible.[9] Thus the argument from motivation needs to be interpreted along the lines of b) if it is to have any chance at success.

Now let's move on to consider premise 2). This premise embodies a view, often called Internalism about normative reasons, according to which normative reasons necessarily have to be capable of motivating. Many philosophers discuss, and some endorse, some such view. Christine Korsgaard, for instance, claims that '[p]ractical reason claims, if they are really to present us with reasons for action, must be capable of motivating rational persons.' (Korsgaard, 1986, p. 5) Similarly, Bernard Williams, takes it that '[i]f something can be a reason for action, then it could be someone's reason for acting on a particular occasion, and it would then figure in an explanation of that action.' (Williams, 1982, p. 106)[10]

[9] Below, I will call this into question, however. In particular, I will argue that premise 3) is vacuously true since its consequent is true no matter what.

[10] Many other philosophers are sympathetic to this sort of view, as well, although they do not formulate the point specifically in terms of reasons. Richard Price, for instance, wrote that '[w]hen we are conscious that an action is fit to be done, or that it ought to be done, it is not conceivable that we can remain uninfluenced, or want a motive to action." (Price 1787, reprinted in Brink, 1989, p. 38) More recently, Stevenson writes: 'A person who recognizes X to be 'good' must ipso facto acquire a stronger tendency to act in its favor than he otherwise would have had.' (Stevenson, 1937, p. 13) Gilbert Harman too is sympathetic with this view: 'To think that you ought to do something is to be motivated to do it. To think that it would be wrong to do something is to be motivated not to do it.' (Harman, 1977, p. 33) For an opponent of such views, see Parfit, 1997.

Why would anybody be inclined to accept a view like Internalism about normative reasons in the first place? The thought must be based on the idea that telling someone there are reasons for him to do something needs to be able to get a grip on that person, influence him, persuade him, or somehow affect his thoughts and actions. Suppose you are having a disagreement with a person. You want your interlocutor to φ and you tell him 'Fact F is a reason for you to φ.' But suppose you also know that F is completely incapable of motivating your interlocutor to φ. He wouldn't be motivated to φ even if he believed fact F, were completely rationally, had all true beliefs about the matter at hand, etc. In such circumstances, you have no chance of influencing your interlocutor by citing fact F. So, we might wonder, in what sense could fact F be any reason for your interlocutor to φ? Accordingly, we might want to assume that there must be a connection between some fact's being a reason for a person to act and its being possible for that person to be motivated to act by a recognition of that fact.

There are many ways to formulate the Internalist constraint on normative reasons, however. After all, in order for some fact to be a reason for one to perform an action, in exactly what sense does it have to be *capable* of motivating one to perform that action? For starters, note that not every version of the Internalist constraint on normative reasons will be strong enough to allow us to construct a plausible version of the argument from motivation against certain first-order theories of welfare. Take for example the following version of the constraint:

> Constraint 1: Necessarily, if fact F is a normative reason for one to φ, then there is some metaphysically possible world in which P is moved to φ by the belief that F is true.

However, it should be clear that even entirely response independent theories of welfare like the Money Theory are going to be compatible with this amazingly weak version of the constraint. So we won't be able to construct a version of the argument from motivation by appeal to this version of the constraint. Something stronger is called for. A better candidate, then, might seem to be this:

> Constraint 2: Necessarily, if fact F is a normative reason for P to φ, then P would have at least some motivation to φ if P were aware that F is true.

This version of the constraint is clearly strong enough to allow a version of the motivation argument to be constructed, but it might in fact be *too* strong. In particular, it faces counter examples of a certain sort. Imagine a person who is prone to brain malfunctions and other forms of irrationality. Now suppose there is a fact F that is a normative reason for this person to perform a certain action, but – thanks to one of his signature brain malfunctions – he fails to have any motivation at all to act even though he is fully aware that F is true. This failure of rationality on the part of this person should not disqualify F from being a normative reason for him to act. But according to constraint 2, F would be disqualified. So to avoid this kind of problem case, a clause about rationality should be included in the formulation of the constraint.

Thus we get the following Internalist constraint on normative reasons:

> (INR) Necessarily, if fact F is a normative reason for P to φ, then P would have at least some motivation to φ provided P were *rational* and aware that F is true.

This principle can accommodate the cases that undermined Constraint 2. It would be no counterexample to (INR) if there were a person who because of a brain malfunction (or some other form of irrationality) failed to be motivated in the slightest by the recognition of some fact that is a normative reason for him to act a certain way. After all, such a person would clearly not count as rational.[11] So (INR) is more plausible than Constraint 2.

Of course, one might want to continue to refine (INR) in the face of further technical problems and counter-examples. Some philosophers defend more sophisticated versions of this principle. [12] However, I am not going to discuss any further refinements here.

[11] The issue of what is meant by ‘rational’ is complex. Parfit (1997), among others, distinguishes between procedural rationality and substantive rationality, and we get different versions of the principle depending on which notion is employed. What I intend here, however, is procedural rationality. For if substantive rationality were used instead, the argument from motivation would become question-begging.

[12] For example, Kieran Setiya defends a principle that is supposed to get around some other difficulties. (Cf. Setiya, 2007, pp. 11-14, p. 95) However, the difficulty he discusses does not present any counter-example to (INR), since (INR) merely gives a necessary condition on normative reasons. Setiya describes the case as follows:

> [Bernard] Williams imagines a thirsty person, presented with what seems to be a glass of cool, refreshing water. In fact, the glass contains odourless petrol. If I am in this situation, is the fact that I am thirsty a reason for me to drink the contents of the glass? (...) the answer would seem to be "no." If the glass contains petrol, the fact that I am thirsty is no reason to drink from it at all... The inclination to say otherwise turns on the fact that I have a collection of psychological states - including the belief that the glass contains water - such that the disposition to be moved to drink by them, together with the belief that I am thirsty, is a good disposition of practical thought [i.e. one of the dispositions

(INR) is plausible enough for our purposes. It is a reasonable formulation of the constraint that can be used to construct a plausible version of the argument from motivation. What's more, (INR) is not idiosyncratic or unconventional. This very principle has been defended by a number of contemporary philosophers writing on related topics. For example, Stephen Darwall writes:

> We may give (…) give a general internalist account of reasons to act: p is a reason for S to do A if, and only if, p is a fact about A awareness of which by S, under conditions of *rational consideration*, would lead S to prefer his doing A to his not doing A, other things equal. The motivational aspect of reasons to act is clear on this account: a fact can only be a reason for someone to act if consideration of it, under certain conditions, would motivate him. (Darwall, 1983, p. 81)[13]

Thus we see that Darwall accepts precisely the principle I have called (INR) (though he, of course, is offering a bi-conditional, while my principle is merely a necessary condition).[14]

This formulation of the Internalist constraint on normative reasons allows us to construct a plausible argument from motivation against the entirely response independent theories. To state the argument, let 'f_1-f_n' rigidly designate the features that the true theory of welfare says a state of affairs must possess in order to enhance P's welfare (i.e. the features that this theory picks out as intrinsically good-making for a person).

> constitutive of being rational]. What the example shows is that good practical thought corresponds to reasons only when it involves no false beliefs. (Setiya, 2007, p. 12)

And in order to get around this case, Setiya proposes this principle:

> *Reasons*: The fact that p is a reason for A to x just in case A has a collection of psychological states, C, such that the disposition to be moved to x by C-and-the-belief-that-p is a good disposition of practical thought [i.e. a disposition that is constitutive of being rational], and C contains no false beliefs. (Setiya, 2007, p. 12)

Reasons is basically the same as (INR) except for the clause about 'no false beliefs.' However, the case Setiya describes does not provide a counter-example to (INR). It would be a counter-example to a principle like (INR) that gives a *sufficient* condition for being a reason to act. But (INR) is not threatened by the case because it is merely giving a *necessary* condition for being a reason to act. After all, in Setiya's case, I have no normative reason to drink the contents of the glass, given that it contains petrol. Thus the antecedent of (INR) is false and the principle is true even in this case.

However, perhaps one would want to include a 'no false beliefs' clause in (INR) in order to get around some *other* version of the case Setiya describes that *really would* threaten (INR). I am open to this possibility (even though I can't think of such a case right now). After all, (INR) revised in this way could still be used to construct the sort of argument from motivation that we are concerned with here.

[13] Also see Darwall, 1983, pp. 41-42 and p. 86

[14] David Brink also seems to accept roughly this principle, though , he puts the point in terms of a connection between normative reasons and explanatory reasons, not motivation. Explanatory reasons are, however, usually taken to be closely related to motivation. (Cf. Brink, 1989, pp. 39-40) Also see Parfit's statement of the view he calls 'Internalism' (cf. Parfit 1997, p. 100) and Setiya's principle *Reasons* (cf. Setiya 2007, p. 12).

The Motivational Argument:

1) The following fact is a normative reason for P to φ: i) there is a certain state of affairs, S, that P would bring about were he to φ and ii) S possesses features f_1-f_n.[15]
2) Necessarily, if fact F is a normative reason for P to φ, then P would have at least some motivation to φ provided P were rational and aware that F is true.
3) Therefore, P would have at least some motivation to φ provided P were rational and aware of the fact that i) there is a certain state of affairs, S, that P would bring about were he to φ and ii) S possesses features f_1-f_n. [1,2, UI & MP]
4) If the true theory of welfare is response independent, then 3) is false.
5) Therefore, it's not the case that the true theory of welfare is response independent.

Line 1) here amounts to premise 1) in the preliminary statement of the argument described at the beginning of this section, as interpreted along the lines of b) from before. Line 2) here is just a statement of the Internalist constraint on normative reasons, or (INR). Line 4) corresponds to premise 3) from the preliminary statement of the argument, again interpreted along the lines of b).

So far, though, I have not said much to *support* line 4). However, there does seem to be some plausibility to it. After all, the response independent theories of welfare do seem to conflict with line 3). We've already seen that this is the case for one entirely response independent theory, namely the Money Theory. If this theory is true, then the features that a state of affairs, S, must possess in order to be welfare enhancing for P is the feature

[15] There is a slight problem with line 1), which I don't think it is really necessary to worry about for present purposes. But here is the problem anyway. Line 1) is problematic as stated because there will be cases in which there are two possible actions one could perform, A and B, and while both of them would lead to outcomes that enhance your welfare, A would enhance your welfare significantly more than B would. In this situation, it is by no means clear that there is a normative reason to for you to perform action B, even though B would enhance your welfare somewhat. Thus we get counter-examples to line 1) as stated.

This problem, however, can be avoided. In particular, we need to modify line 1) in such a way that it says that there is a normative reason to perform specifically the actions that produce outcomes that *maximize* welfare enhancement. The formulation I propose is this:

> 1') The following fact is a normative reason for P to φ: i) there is a certain state of affairs, S, that P would bring about were he to φ and ii) there is no other state of affairs that P could bring about instead that possesses f_1-f_n to a greater degree than S does is a normative reason for P to φ (where 'f_1-f_n' rigidly designates the features that the true theory of welfare says a state of affairs must possess in order to enhance P's welfare).

This formulation of line 1) is more plausible than the formulation I use in the text. For 1') does not imply, as 1) did, that there is a normative reason for you to perform absolutely *any* action that would bring about a state of affairs that enhances your welfare even a little; rather 1') implies that there is a normative reason for you to perform just those actions that would maximize your welfare. Thus 1') does not suffer from the sort of counter-example that 1) does.

of P's obtaining more money in S. However, even if the Money Theory were true, there clearly could still be people who would be fail to be motivated to perform the actions that they believe would get them more money. Thus if the Money Theory were true, line 3) would be false. That is, even if the Money Theory were true, it would still not be the case that an arbitrarily chosen person P would have at least some motivation to φ provided P were rational and aware of the fact that i) there is a certain state of affairs, S, that P would bring about were he to φ and ii) S possesses features the features that the Money Theory picks out as good-making for P.

The same goes for *partly* response independent theories of welfare as well. Consider Feldman's theory Desert-Adjusted Intrinsic Attitudinal Hedonism (which he calls 'DAIAH'), for example.[16] DAIAH is the theory, roughly, that a person is well off to the extent that she is pleased by objects that are worthy of having pleasure taken in them. This is a partly response independent theory because it implies that the welfare value of a state of affairs for a person, P, is not *solely* a function of P's responses (in particular, the episodes of pleasure P experiences) but also of the *pleasure-worthiness* of the objects of P's episodes of pleasure. DAIAH, too, conflicts with line 3). For if DAIAH is true, there will still be cases in which a person would not have any motivation to φ *even though* he is aware of the fact that his φ-ing would bring about a state of affairs that possesses the features that DAIAH picks out as good-making. For suppose there is a person who cares neither about experiencing pleasure nor about the worthiness of the objects of his pleasures. (Perhaps he cares only about getting more money.) Suppose this person comes to recognize the fact that if he were to φ, then he would bring about a state of affairs in which he takes lots of pleasure in some very pleasure-worthy things. Why think that such a person would have any motivation to φ? After all, he doesn't care in the slightest about taking pleasure in pleasure-worthy things. Thus if DAIAH is true, the claim made in line 3) would turn out to be false as well.

And so it goes, one might think, for all the response independent theories of welfare. Thus one might be inclined to believe line 4) in the argument.

[16] Cf. Feldman, 2004, p. 120

3.2.2 The problem with the argument

Now that something has been said in support of all the lines in the Motivational Argument, we can go on to ask whether it really is a good argument. Does it give reason to reject the response independent theories of welfare? I will argue that it does not. Whatever attractiveness the Argument from Motivation might seem to have on its face is, I suspect, merely the result of a failure to appreciate just how hard it is for a theory of welfare to be compatible with the Internalist constraint on normative reasons, i.e. with (INR). In order for a theory of welfare, T, to be compatible with this constraint, the following would have to be the case: the features that T picks out as welfare enhancing are such that if you were to recognize that your φ-ing would bring about a state of affairs that possesses precisely *these features*, then you would have some motivation to φ (provided you are rational). But not only do response independent theories like the Money Theory and DAIAH fail to meet this condition, I will argue that the response dependent theories do as well. Thus I claim that the Argument from Motivation fails because *no theory of welfare is compatible with the Internalist constraint on normative reasons* (and in particular with line 3 of the argument).

To show this, I will run through several paradigmatic examples of response dependent theories and argue that they are all inconsistent with the Internalist constraint (i.e. with INR), and I hope that this will provide reason to think that no theory of welfare is consistent with the constraint. As a result, my view is that line 4) in the argument is vacuously true. For no matter whether the antecedent of line 4) is true or false, its consequent is always true. Thus the Argument from Motivation does not provide a reason to reject the response independent theories of welfare in favor of the response dependent theories.

Begin with Sensory Hedonism. This is a paradigmatic example of an entirely response dependent theory. According to it, your welfare is directly determined by the amount of pleasure and pain you experience. However, even if Sensory Hedonism were true, there would be no guarantee that you would be motivated (on pain of irrationality) to φ if you were to recognize that your φ-ing would bring about a state of affairs that possesses the features that this theory picks out as welfare enhancing – i.e. pleasantness.

To see this, consider Feldman's example of Stoicus.[17] Suppose Stoicus experienced a lot of sensory pleasure as a young man, but it always got him into a lot of trouble. After much soul-searching, he decides that he doesn't ever want to experience sensory pleasure again. He studies and meditates, and becomes a practiced ascetic. He becomes so practiced, in fact, that he is never again tempted by the prospect of experiencing sensory pleasure. Stoicus seems to be fully rational.[18] He is intelligent and cool and calculated and reflected. He thinks before he acts and is never overcome by passion. He never flies off the handle irrationally. Now suppose that Stoicus learns that a certain action he could perform would lead him to feel a great deal of sensory pleasure. One night, a nun in the adjoining convent sneaks into his room and suggests that they break a couple of the rules. Stoicus knows he would feel a lot of pleasure as a result of doing so. Nonetheless, Stoicus has no motivation whatsoever to perform the action that the nun is suggesting. Even if Sensory Hedonism were true, a case like this is surely possible. While Stoicus might still be motivated by his conception of the good (i.e. tranquility and meditation), he happens not to believe Sensory Hedonism. And so he is not motivated in the slightest by the prospect of sensory pleasure.

What this case shows is that Sensory Hedonism is not compatible with the claim made in line 3) of the Argument from Motivation. Stoicus knows that if he were to perform a certain action together with the nun, it would bring about a state of affairs that has the features that Sensory Hedonism picks out as the good-making ones. But despite the fact that Stoicus is fully rational, he has no motivation whatsoever to perform the action in question. Thus if Sensory Hedonism were true, the claim made in line 3) of the argument would be false. That is, even if Hedonism were true, there would be no guarantee that a person would be motivated to act by the recognition that he could do something that produces an outcome with the features that this theory picks out as welfare enhancing (i.e. pleasantness).

A similar point applies when it comes to Desire Satisfactionism, another paradigmatic example of an entirely response independent theory. Desire Satisfactionism is roughly the theory that what intrinsically enhances your welfare is getting the things that you desire.

[17] Feldman, 2004, pp. 49-50.

[18] Procedurally rational, that is.

According to this theory, the fundamental bearers of intrinsic welfare value are states of desire satisfaction – i.e. complex states consisting of a person's desiring that p and p's being true. Even if Desire Satisfactionism were true, however, there would be no guarantee that a person would be motivated to act by the recognition that he could do something that produces an outcome with the features that this theory picks out as welfare enhancing (i.e. being an episode of desire satisfaction).

To see this, consider the case of Stoicus Jr. He is an intelligent person who is cool and calculated and reflected. He never flies off the handle irrationally. What's more, he has trained himself never to be motivated by the prospect of getting his desires satisfied. Stoicus Jr. *has* all sorts of desires, to be sure, but whenever he realizes that a given action he can perform would lead to one of these desires being satisfied, he cools off and he loses his motivation to perform that action.[19] Thus Stoicus Jr. is never motivated in the slightest by the prospect of getting his desires satisfied. He does not care about receiving episodes of desire satisfaction, and when he thinks he can bring about an episode of desire satisfaction for himself, he has no inclination to do so.[20]

What this case seems to show is that Desire Satisfactionism is not compatible with the claim made in line 3) of the Argument from Motivation either. Consider an occasion on which Stoicus Jr. knows that he would receive an episode of desire satisfaction if he were to φ. That is, if he were to φ, he would bring about a state of affairs that has the features that Desire Satisfactionism picks out as the good-making ones. But despite the fact that Stoicus Jr. is fully rational, he has no motivation whatsoever to φ. Thus if Desire Satisfactionism were true, the claim made in line 3) of the argument would be false. Even if Desire Satisfactionism were true, there would be no guarantee that one would be motivated to act by the recognition that one could do something that produces an outcome with the features that this theory picks out as welfare enhancing (i.e. being an episode of desire satisfaction).

To this, some might object that I have misunderstood Desire Satisfactionism. I took it that Desire Satisfactionism entails that the bearers of intrinsic welfare value are episodes of desire satisfaction (i.e. complex states of one's desiring that p is true and p's really

[19] Thus, it's only when he acts on instinct, without thinking, that Stoicus Jr. ever gets anything done.

[20] Moreover, it's not that Stoicus Jr. has a desire *not* to get his desires satisfied. For that would lead to paradoxes that I would rather steer clear of.

being true). But perhaps one thinks that Desire Satisfactionism should be understood in a different way. In particular, one might want to take the theory to state that the bearers of intrinsic welfare value are the *things* that you desire (provided you obtain them). If one understands Desire Satisfactionism in this way, then Desire Satisfactionism might not be inconsistent with the claim made in line 3) of the Argument from Motivation. After all, while Stoicus Jr. would not ever be motivated to obtain *episodes of desire satisfaction* (i.e. under that description), he would sometimes desire certain *things* that he *is* motivated to obtain. Sometimes he would desire to sleep, and he might be motivated to do so. Sometimes he would desire to meditate, and he might be motivated to do so. Thus if Desire Satisfactionism is understood so that what is good for a person is the *things* that are desired (as opposed to episodes of desire satisfaction), then Stoicus Jr. would not be a case of a rational person who fails to be motivated by the prospect of obtaining that which Desire Satisfactionism (on this new understanding) picks out as the good. Thus Desire Satisfactionism (on this new understanding) would not be inconsistent with the claim made in line 3) of the Argument from Motivation.

However, this new way of understanding Desire Satisfactionism is highly implausible. Clearly we must take the bearers of intrinsic welfare value to be episodes of desire satisfaction. After all, if one takes it that the bearers of intrinsic welfare value are the *things* that are desired and obtained, then Desire Satisfactionism would conflict with the widely accepted axiological assumption that the intrinsic value of something can depend solely on its intrinsic features.[21] Suppose Desire Satisfactionism were taken to state that the bearers of intrinsic welfare value are the *things* you desire (as opposed to episodes of desire satisfaction). In that case, when you desire an apple and get it, the

[21] See, for example, Feldman 2004, p. 73 and Bradely 2009, p. 19. Bradley offers the following argument for the principle (which he calls SUP) that the intrinsic value of something must depend solely on its intrinsic properties: 'SUP is a requirement of any acceptable theory of well-being. This is because, as noted above, the value atoms should be *instantiations of the fundamental good- or bad-making properties* – the properties that are fundamentally and completely responsible for how well a world (or a life, or …) goes. Suppose SUP were false. Then there could be two properties, F and G, such that the only intrinsically good states of affairs are those involving the instantiation of F alone, but whose values are determined by whether there are any instantiations of G. But if that were true, F would fail to be a fundamental good- or bad-making property, for instantiations of F would fail to completely determine what value there is. The fundamental good- or bad-making property would involve both F and G, contrary to our assumption. Once we are committed to the project of finding the fundamental good- and bad-making properties – the fundamental project of axiology, and of the theory of well-being – we are immediately committed to SUP…' (Bradley, p. 19)

thing that would be good for you is the apple. But in that case, the intrinsic value of the apple for you would *not* depend only on the *intrinsic* features on the apple. It would also depend on its relation to you and your desires. But this is an extrinsic feature of the apple. And so the axiological principle that intrinsic value must depend on intrinsic features would be violated. Thus Desire Satisfactionism cannot be taken to state that the bearers of intrinsic welfare value are the *things* that are desired. Instead, the theory must be taken to state that the bearers of welfare value are episodes of desire satisfaction. On the theory so understood, what is good for one is getting what one wants (not the things that one wants). Thus the only plausible interpretation of Desire Satisfactionism would be the one that I have already argued is incompatible with line 3) in the Argument from Motivation.

Thus we have two paradigmatic examples of entirely response dependent theories that are incompatible with the claim made in line 3) of the Argument from Motivation, and so with the Internalist constraint on normative reasons itself. In fact, it seems likely that *no theory of welfare* has any real chance of being compatible with this claim. Just as the case of Stoicus showed Hedonism to be incompatible with the claim that a rational person would always be motivated to φ by the recognition that φ-ing would bring about an outcome with the features that Hedonism picks out as good-making (i.e. pleasantness), so too could an analogous case be constructed for any theory of welfare. It should be obvious how to construct a Stoicus-like case for any other theory of welfare showing that this theory's good-making features do not necessarily provide motivation to act: just imagine a rational person who would not be at all motivated to act so as to produce outcomes possessing the features that the theory in question picks out as good-making.

Based on these considerations, my view is that no theory of welfare, response independent or response *de*pendent, is compatible with the claim made in line 3), and by extension with the Internalist constraint on normative reasons that was used to derive line 3). The upshot of this is that line 4) in the argument is vacuously true. The antecedent of this line supposes that the true theory of welfare is response independent, while the consequent is the claim that line 3) is false. But no matter whether this antecedent is true or false (i.e. no matter whether the true theory of welfare is response independent or response dependent), line 3) is false. Thus the consequent of line 4) is always true, no matter what the truth value of the antecedent of line 4) is. Accordingly, the Argument from Motivation is unconvincing. The Argument from Motivation does not seem to provide a reason to reject the response independent theories of welfare and prefer the response dependent theories instead.

What's more, if this evaluation of the Argument from Motivation is correct, then it would be a problem for philosophers like Darwall who accept the Internalist constraint on normative reasons (i.e. INR). The reason is that (INR), together with the plausible assumption embodied in line 1), entailed a claim – viz. line 3) – that no theory of welfare seems to be compatible with. However, one might think that there clearly has to be *some* theory of welfare that is true. And so this would spell doom for (INR). In particular, the argument would go like this: i) (INR) entails, together with the plausible assumption in line 1), that no theory of welfare is true. ii) But it's not the case that no theory of welfare is true. iii) Therefore, it's not the case that (INR) is true. The truth of this conclusion would surely be an unwelcome result for supporters of Internalism about normative reasons.

3.2.3 Conclusions

Some might want to reject the response independent (objective) theories of welfare because of their perceived inability to account for the necessary connection between welfare and motivation. In section 3.2, I tried to formulate a version of this Argument from Motivation that is based on a common Internalist view about the connection between normative reasons and motivation. Unlike the argument considered in section 3.1, this Argument from Motivation really did purport to establish a conclusion about first-order theories of welfare. However, I argued that the argument does not succeed. For no first-order theory of welfare is compatible with the Internalist constraint on normative reasons that is required in order to get the argument to work. Thus the Argument from Motivation does not provide reason to reject the response independent theories of welfare in favor of the response dependent ones. This is a welcome result for those who want to be free to endorse some kind of response independent theory of welfare.[22]

[22] One might also think that a version of the argument from motivation could be constructed by appeal to a more modest view than Internalism about normative reasons, viz. by appeal to Internalism about a *person's good*. However, I think Internalism about a person's good is *false* for the very same reason that the Argument from Motivation I have just been discussing (based on Internalism about normative reasons) fails. For more on the problems with Internalism about a person's good, see Appendix I.

CHAPTER 4

ADJUSTED ENJOYMENT THEORIES OF WELL-BEING

In Chapter 2, I argued that the correct theory of well-being is in all likelihood to be found among the partly response independent theories, and that we should therefore focus our investigation on the theories of this type. There seem to be four main sorts of theory that fall into the partly response independent category: Objectively-Adjusted Enjoyment Theories, Objectively-Adjusted Desire Satisfactionism, Aristotelian Perfectionism and certain Hybrid Theories. Of course, these four types of theory do not *exhaust* the theories in the partly response independent category. Nonetheless, these four seem to represent the most influential ones that belong in this group. In the chapters that follow, I will investigate all of them except Aristotelian Perfectionism. Given the well-known problems for teleological notions like human essence and the characteristically human functions,[1] and given how crucial some such notion is to Aristotelian Perfectionism, I am pessimistic about the prospects of the theories of this sort. Thus I have opted not to devote a chapter to them.[2] By contrast, I think that certain objectively-modified versions of the Enjoyment theory and the Desire Satisfaction theory are promising and require serious investigation. Chapter 4 (i.e. this one) focuses on the Enjoyment Theories, while chapters 5-7 focus on Desire Satisfaction theories. Chapter 8 deals with certain Hybrid Theories.

[1] See for instance, Kitcher, 1999. (Also see Hurka, 1993, ch. 1 & 2, which Kitcher is criticizing.)
[2] Full disclosure: this was in part due to external time constraints, as well.

The present chapter will focus on those theories I call the Adjusted Enjoyment Theories. What characterizes these theories as a class is that they all take it that a) a certain kind of mental state directly impacts welfare – namely enjoyment of one kind or another (whether it be some kind of pleasure, satisfaction, or what have you), and b) the magnitudes of the contributions to well-being that are made by these episodes of enjoyment are to be *adjusted* by some variable or other.[3] There are two sub-categories of Adjusted Enjoyment Theories, however: those that take the magnitudes of the welfare contributions of episodes of enjoyment to be adjusted by some *mental* variable (like the phenomenological qualities of the enjoyment), and those theories that take the magnitudes of these contributions to be adjusted by some *extra-mental* variable (like the pleasure-worthiness or merit of the enjoyment). The former group I call the *Subjectively-Adjusted Enjoyment Theories* (SAETs), while I call the latter group the *Objectively-Adjusted Enjoyment Theories* (OAETs).

Note that the Subjectively-Adjusted Enjoyment Theories of welfare – one example of which would be a possible interpretation of Mill's qualified hedonism (discussed below) – will all count as mental state theories. For according to the SAETs, a person's actual level of well-being will supervene on his/her mental states. In other words, these theories imply that there can be no difference between the welfare levels of two people without there also being some difference in their mental states.

By contrast, the Objectively-Adjusted Enjoyment Theories will all count as partly response independent theories. Why is this? Given that they take enjoyment to impact welfare, these theories clearly do not count as *entirely* response independent theories. Moreover, since they take there to be some extra-mental factor or other, like pleasure-worthiness, that affects the contributions to welfare made by episodes of enjoyment, the OAETs do not count entirely response dependent theories. Thus the OAETs are going to be *partly* response independent theories.

Like any mental state theory, the SAETs are going to fall prey to powerful objections like the Experience Machine Argument. As we saw in Chapter 2, it seems unlikely that

[3] In this chapter I will only be discussing the monistic theories of welfare that fit this description. It is possible for there to be a hybrid, or pluralistic, theory of welfare that fits this description too, provided that the theory takes it that some kind of adjusted enjoyment is one of the components of welfare. However, I reserve discussion of the hybrid theories of welfare for a later chapter.

any mental state theory could meet the descriptive adequacy requirement, that is, capture the main features of our everyday evaluative experience. The OAETs, by contrast, since they belong to the category of the partly response independent theories, have much more going for them. They do not, for instance, fall prey to Experience Machine Arguments.

In fact, the Objectively-Adjusted Enjoyment Theories currently seem to be in the ascendancy. A number of prominent moral philosophers have defended theories of welfare of this type in the past few decades. In particular, Robert Adams, Steven Darwall, Fred Feldman and Derek Parfit all seem to endorse Objectively-Adjusted Enjoyment Theories of welfare. These philosophers' theories have many strengths when compared with more traditional versions of Enjoyment Theory (i.e. Sensory Hedonism, or Mill's theory). What's more, the problem that many might consider to be the biggest challenge to the OAETs – namely, the absence of a good account of the factor that adjusts the values of episodes of enjoyment – can be solved, I think. In fact, here I will discuss some attempts to solve it. Nonetheless, I will ultimately argue that the OAETs defended by Adams, Darwall, Feldman and Parfit all face a common problem, and thus cannot in fact represent the whole truth about the nature of well-being.

4.1 Mill's Qualified Hedonism

The version of Hedonism that Mill endorsed seems to be the first major attempt to defend an Adjusted Enjoyment Theory of well-being. Thus it is fitting that our discussion of the Adjusted Enjoyment Theories begins with him. However, there are several different ways to understand Mill's attempt to defend an Adjusted Enjoyment Theory, and all of them face some problem or other.

The version of Hedonism that Mill wants to defend rests on a distinction between pleasures of higher quality (associated with the intellectual faculties) and pleasures of lower quality (associated with the body). Mill takes it that pleasures of the former sort, all things being equal, have more welfare value than pleasures of the latter sort. As he puts it:

> It is quite compatible with the principle of utility to recognize the fact that some kinds of pleasure are more desirable and more valuable than others. It would be absurd that, while in estimating all other things quality is considered as well as quantity, the estimation of pleasure should be supposed to depend on quantity [i.e. intensity and duration] alone. (Mill, 2001, ch. 2, p. 8)

This suggests that Mill held the view that the intrinsic (welfare) value of an episode of pleasure should be enhanced if it is a pleasure of the higher type, but decreased if it is a pleasure of the lower type. To help with the calculations, we may suppose that episodes of pleasure may be given *quality ratings* depending on how 'high' or 'low' they are. The intrinsic (welfare) value of an episode of pleasure, then, is to be calculated by multiplying the pleasure's intensity by its duration by its quality.

Thus we end up with the following rough picture of how Mill thinks the welfare value contained in a life should be determined. We should begin by locating all the episodes of pleasure contained in the life, and for each one determine its intrinsic value by multiplying its intensity by its duration by its quality rating. Then we should add up the intrinsic values of all these episodes of pleasure. This is the total amount of pleasure contained in the life. The next step is to calculate in a similar fashion the total amount of pain that is contained in the life. Finally, we are to find the total welfare value of the life by subtracting the total amount of pain it contains from the total amount of pleasure it contains.

Even with this rough procedure in place for determining the welfare value of a life, an important question remains: What exactly is it that makes one pleasure 'higher' or 'lower' than another? In Mill's view, 'what makes one pleasure more valuable than another… except its being greater in amount' is captured by the following test:

> Of two pleasures, if there be one to which all or almost all who have experience of both give a decided preference, irrespective of any feeling of moral obligation to prefer it, that is the more desirable [i.e. valuable] pleasure. (Mill, 2001, p. 8)

This test, involving the preferences of those who have the requisite experiences, has been roundly criticized.[4] Perhaps most troublingly, it is by no means clear, as Mill claims, that all or even most of those people who have experienced both the 'higher' pleasures of the intellect and the 'lower' pleasures of the senses would 'decidedly prefer' the former to the latter. Personally, I have doubts about the veracity of the claim that virtually everyone would prefer the pleasures of art and poetry to those of beer and sex, even holding fixed the intensities and durations of the pleasures involved. What's more, one might even be inclined to doubt that most people would have *any preference at all* if asked to choose

[4] See for example, G.E. Moore's criticism of Mill's Hedonism. (Moore, 1903, ch. 3, § 47-53)

between an episode of pleasure from drinking beer with a certain intensity and duration, on the one hand, and an episode of pleasure from listening to poetry that is equally intense and lasts the same amount of time, on the other.

However, even if one does not find Mill's test for differentiating between higher and lower pleasures to be plausible, one might still suppose that there is *some* distinction or other that he was trying to get at. And insofar as one is interested in understanding Mill's theory, one will want to know what that distinction is supposed to be based on. To put it another way, if we accept just for the sake of argument that higher pleasures are to be distinguished from the lower pleasures by appeal to the preferences of people who have had the relevant experiences, which *features* of the pleasures is it that these judges are supposed to be forming their preferences on the basis of? Mill is not explicit about this, and as a result it becomes possible to interpret Mill's theory in a couple of different ways.

One natural suggestion is that the distinction Mill has in mind is to be drawn on the basis of what *causes* the various pleasures. In particular, one might think that the term 'higher pleasure' applies to episodes of sensory pleasure whose proximate cause is intellectual activity of one kind or another (like reflection, contemplation of a work of art, or reading a great piece of literature), while 'lower pleasure' applies to episodes of sensory pleasure whose proximate cause is some bodily function (like eating, drinking or sex). This might seem to account for Mill's separating 'pleasures of the intellect, of the feelings and imagination, and of the moral sentiments' from the pleasures of 'mere sensation', and for the contrast he draws between 'mental' and 'bodily pleasures.' (Mill, 2001, p. 8)

However, Feldman (among others) has pointed out that this way of understanding the basis for Mill's distinction between the higher and the lower pleasures is highly problematic. As he puts the point:

> if the difference between a higher pleasure and a lower pleasure is simply a matter of *source* or *cause*, then this cannot ground a difference in intrinsic value. Intrinsic values cannot be affected in this way by extrinsic features. (Feldman, 2004, p. 73)

Feldman's point here is that if Mill were going to try to base the distinction between higher and lower pleasures on what the causes of these pleasures are, then Mill's Hedonism would conflict with a widely accepted axiological principle, namely the principle that 'intrinsic values depend upon intrinsic features.' (PGL, p. 73) Why is this?

Although there is disagreement about how to understand what sensory pleasures are, the two main accounts both yield the result that the cause of a sensory pleasure is an *extrinsic* feature of the pleasure. On the one hand, some philosophers have taken sensory pleasures to be characterized by some phenomenological property that they have in common (i.e. they all 'feel a certain way').[5] By contrast, others have taken it roughly that sensory pleasures are experiences that are 'apprehended as desirable.'[6] Either way, the physiological causes of an episode of sensory pleasure would not be an intrinsic feature of it – it would not be part of *what it is* to have a sensory pleasure. This is problematic because episodes of sensory pleasure are precisely what is of *intrinsic* value according to Mill's Hedonism. After all, on Mill's theory, sensory pleasure is not merely productive of well-being; it is what *in and of itself* positively impacts one's well-being. Thus distinguishing between higher and lower pleasures on the basis of their causes would conflict with a widely shared assumption about the concept of intrinsic value – namely, that the intrinsic value of a thing cannot depend on its extrinsic features, but must be determined solely by its intrinsic features.[7]

The upshot is that it is not open to Mill to attempt to draw the distinction between higher and lower pleasures by appeal to what the *causes* of these pleasures are. So Mill must seek a different basis for his distinction. The natural solution is to take it that the higher pleasures somehow *feel* different from the lower pleasures, and that the differences in the phenomenological properties of various episodes of pleasure is what grounds assigning different quality ratings to them. If one seeks to draw the distinction between higher and lower pleasures in this way, it seems one would avoid the result that Mill's Hedonism would conflict with the axiological principle that the intrinsic value of something must depend solely on its intrinsic features.

However, this phenomenological-properties strategy encounters serious problems as well. For one thing, it is difficult to know precisely *which* phenomenological properties one could appeal to in order to draw the distinction. I suppose it depends on which

[5] See, for example G.E. Moore, 1903, p. 12-13. For a modern proponent of this view, see Brink, 1989, p. 221.

[6] Feldman claims that this view traces back to Sidgwick, 1874, p. 127, although others like Alston, Brandt and Frankena also held such a view. For discussion and criticism of the Sidgwickian view, see Feldman, 1997, pp. 448-466.

7 See Feldman, 2004, pp. 72-73. Also see Moore, 1903, p. 260, Feldman, 1997, p. 136-139 and Bradley, 2009, p.

account of sensory pleasure one prefers. If one takes sensory pleasures to be characterized by their having a certain feel, then one might allow that this particular feel comes in various ‘flavors’ – one intellectual flavor that is challenging and involves mental effort, and another bodily flavor that is requires little mental effort. On the other hand, if one takes sensory pleasures to be experiences that are apprehended as desirable, then one might allow that this ‘apprehending as desirable’ can feel different ways to a person as well. Perhaps one might apprehend some things as desirable in a way that feels *deep* or *profound*, in which case one’s pleasure would be of the higher type, while it might also be possible to apprehend things as desirable in a way that feels *superficial* or *shallow*, in which case one’s pleasure would be of the lower type.

I am not certain that either of these two attempts to draw the distinction between higher and lower pleasures by appeal to phenomenological properties is coherent. However, even if there *does* turn out to be some successful way of appealing to the phenomenological properties of pleasures to separate the higher ones from the lower ones, a serious problem remains. In particular, if the higher pleasures are the ones that display a certain phenomenological property P to a greater degree, while the lower ones are those that display P to a lesser degree, then Mill’s Hedonism would turn out to be a Subjectively-Adjusted Enjoyment Theory (SAET). And so it would suffer from the same defect that undermines all theories of this type.

A SAET, recall, is a theory that takes the magnitudes of the welfare contributions of episodes of enjoyment to be adjusted by some *mental* variable, like the phenomenological qualities of the enjoyment. If Mill’s Hedonism is to be understood in the way we are discussing now, it would qualify as a SAET. For if higher quality ratings are to be assigned to those episodes of pleasure that display a certain phenomenological property P to a greater degree, and lower quality ratings are to be assigned to those episodes of pleasure that display P to a lesser degree, then the intrinsic values of episodes of pleasure would end up being modified by a *purely mental variable* – in particular, the phenomenological properties of the pleasures. So Mill’s Hedonism would be a subjectively-adjusted theory.

But this means that Mill’s Hedonism, like any SAET, will fall prey to Experience Machine Arguments. According to the SAETs, a person’s actual level of well-being will supervene on his/her mental states. According to Mill’s Hedonism on the present understanding of it, a person’s well-being will be determined entirely by the intensities, durations and qualities of one’s pleasures and pains, where quality here is to be understood solely in terms of the phenomenological properties of the pleasures and pains one experiences. Thus on the present version of Mill’s Hedonism, there could be no difference between two people’s welfare levels without there also being some difference in these people’s mental states. This, however, is implausible. For suppose there are two people who are identical with respect to their mental states. In particular, they are both experiencing an episode of pleasure of intensity ten for one hour, and the phenomenological properties of their pleasures are identical too. However, the first person is plugged into the Experience Machine and is hallucinating the whole thing, while the second person is not; he’s having a genuine experience in the real world. The present version of Mill’s Hedonism implies that these two people have the same level of welfare. But this seems not to be a plausible result. Many philosophers (myself included) are inclined to think that experiences that occur in the Experience Machine, all other things being equal, have less intrinsic welfare value than qualitatively identical experiences that occur outside the machine.

Thus even if one were to find some satisfactory way of distinguishing between the higher pleasures and the lower pleasures by appeal to their phenomenological properties, the version of Mill’s Hedonism that results would still run afoul of the Experience Machine Argument. Accordingly, I do not take the phenomenological properties strategy for spelling out Mill’s Hedonism to be a promising avenue. Another interpretation of Mill is needed.

Feldman gives us what we need. He proposes a version of Hedonism that is inspired by Mill’s ideas that does not run into the problems that undermine the two previous interpretations. Instead of understanding Mill’s theory in terms of *sensory pleasure*, Feldman proposes that the theory be understood in terms of *attitudinal pleasure*. We will discuss the nature of attitudinal pleasures in more detail below, but for now the general idea is that an attitudinal pleasure is a pleasure that one takes in some state of affairs that it seems to one obtains. Formulating Mill’s theory in terms of attitudinal pleasure is advantageous because attitudinal pleasures have *objects*. As Feldman explains,

> we can construct [a theory like Mill's] in such a way that the objects of attitudinal pleasures are intrinsic elements in the episodes of the pleasure. Thus, we can avoid conflict with the principle that intrinsic values depend on intrinsic features. (Feldman, 2004, p. 73)

The trick, then, is that one can take the 'higher pleasures' to be the episodes of attitudinal pleasure whose objects are associated with the intellectual faculties, while one can take the 'lower pleasures' to be the episodes of attitudinal pleasure whose objects are associated with eating, drinking, sex and other functions of the body. Because the higher and lower pleasures are distinguished by appeal to features of the objects of the attitudinal pleasures (i.e. their highness or lowness), and because *the object of an attitudinal pleasure is intrinsic to (i.e. part of) that pleasure*, the distinction between higher and lower pleasures would not appeal in any way to extrinsic features. As a result, there would be no conflict with the principle that intrinsic values must depend on intrinsic features. And so we would be able to formulate a version of Mill's Hedonism that avoids the problems of the first interpretation of Mill. Moreover, the attitudinal version of Mill's theory would be an Objectively Adjusted Enjoyment Theory. The highness or lowness of an episode of attitudinal pleasure, after all, would not depend on the mental states of the person in question. So the attitudinal version the theory would avoid the problems of the second interpretation of Mill as well.

A problem remains, however. In particular, there is no evidence whatsoever that Mill had any conception of attitudinal pleasure. Thus even if it is *possible* to formulate an attitudinal version of Mill's Hedonism, this theory is not likely to have been the one that Mill himself had in mind.

It is not clear that this should bother us that much, though, since our primary aim here is not historical. We can investigate the question of how plausible the attitudinal version of Mill's Hedonism, which Feldman proposes, is on its own merits. However, this is a question about which I will not say a whole lot about in this paper. For as we will see, the theory of well-being that Feldman himself defends is a version of attitudinal Hedonism that is quite similar to the attitudinal interpretation of Mill's Hedonism. Not only that, Feldman's theory seems to have a number of advantages over this version of Mill's theory. And what's more, the same problems that seem to undercut Feldman's theory would (if successful) also undercut the attitudinal interpretation of Mill's theory. Thus it will not be necessary to evaluate Mill's theory in its own right.

Since, as we have seen, the Subjectively-Adjusted Enjoyment Theories all run afoul of the Experience Machine Argument and therefore do not have much going for them, let us now go on to investigate the Objectively-Adjusted Enjoyment Theories that have been defended in recent time.

4.2 Prominent Objectively-Adjusted Enjoyment Theories

At least four prominent moral philosophers – Adams, Darwall, Feldman and Parfit – have recently defended what I call Objectively-Adjusted Enjoyment Theories. Feldman's view is worked out in the most detail, while Parfit's is worked out in the least. Adams and Darwall fall somewhere in between. In this section, I will argue that Parfit's view (perhaps only because it is underdescribed) and Darwall's view suffer from various problems, and that these problems suggest that Adams and Feldman offer superior ways to defend theories of the Objectively-Adjusted Enjoyment sort.

4.2.1 Parfit

After a discussion of the advantages and disadvantages of various substantive theories of well-being, Parfit goes on to describe some features that the correct theory of well-being would seem to have to possess:

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

In this passage, Parfit does not present a fully worked out theory of well-being. So the passage is in principle open to several different interpretations. However, a couple features of Parfit's view are quite clear.

For one thing, Parfit's statement that 'pleasure with many other kinds of object has no value' suggests that he thinks the correct theory of well-being would have to be consistent with:

a) A given episode of pleasure enhances a person's well-being, but only if the object of that pleasure is some item on the Objective List (e.g. knowledge, rational activity, love, beauty).

Moreover, Parfit's claim that 'if they are entirely devoid of pleasure, there is no value in knowledge, rational activity, love, or the awareness of beauty' suggests that he thinks the correct theory would also have to be consistent with:

b) The items on the Objective List (e.g. knowledge, rational activity, love, beauty) are such that a person's obtaining them would enhance his well-being, but only if he receives some amount of pleasure from obtaining them.

The problem is that a great many theories of well-being are consistent with these two claims. One theory that is consistent with claims a) and b), for instance, is the theory that one must take pleasure in the items on the Objective List in order for them to enhance your well-being, but that these items would enhance your well-being by the same amount no matter how much pleasure you take in them. On this theory, your appreciating the beauty of the Rembrandt painting enhances your well-being to degree X only if you enjoy the experience, but it would enhance your well-being to degree X *no matter how much* you enjoy the experience.

Nothing Parfit says in the passage above rules out his endorsing such a theory. However, perhaps you will agree with me that it has rather implausible consequences. What's more, there are other theories that are consistent with claims a) and b) and that seem to be a more plausible. One is the theory that a person's life contains well-being to the extent that it contains pleasure that is taken in certain objects (e.g. knowledge, rational activity, love, the awareness of beauty, etc…) and lacks pain. The theory I am suggesting we attribute to Parfit is one according to which it is only the 'Good Pleasures' – or pleasures taken in items on the Objective List of goods – that enhance well-being. So to be more precise, we could spell out Parfit's Theory as follows:

(PT) Here is how to determine the amount of well-being in P's life:

1) Find all the episodes of pleasure in P's life.
2) For each one, multiply its intensity by its duration. This is the *raw value* of the episode of pleasure.
3) For each episode, determine whether it is pleasure taken in some item on the Objective List of goods. Call the ones that are the *Good Pleasures* and the ones that aren't the *Bad Pleasures*.

4) Multiply the raw value of each Good Pleasure by 1, and the raw value of each Bad Pleasure by 0. This gives you the *adjusted values* of the Good Pleasures and the Bad Pleasures, respectively.
5) Add up the adjusted values of all the episodes of pleasure in P's life.
6) From this sum, subtract the total amount of pain contained in P's life, calculated in a similar way.[8]
7) The number you end up with equals the amount of well-being contained in P's life.

(PT) is consistent with claims a) and b). Moreover, it is an Objectively-Adjusted Enjoyment theory because the magnitudes of the contributions to well-being that are made by episodes of enjoyment are to be adjusted by a certain extra-mental variable. In particular, the variable in question concerns whether or not the object of the pleasure is included on the Objective List of goods. (PT) specifies that the contributions to well-being that are made by episodes of enjoyment are to be adjusted in a binary way. The value of pleasures taken in items on the Objective List are to be multiplied by 1 and therefore count towards one's well-being, while the value of pleasures taken in items that are *not* on the list are to be multiplied by 0 and therefore do *not* get to count towards one's well-being.

I think (PT) is the weakest but still fairly plausible theory of well-being that is supported by the above passage of Parfit's. If we were to make the theory more sophisticated – say by, allowing that pleasure taken in some objects (perhaps love and beauty) enhances well-being to a greater degree than pleasure taken in other objects (perhaps knowledge and rational activity) – then we would be going well beyond what Parfit says in the passage above. There is no textual support for attributing to Parfit a more sophisticated Objectively-Adjusted Enjoyment Theory than (PT). Perhaps he really had *in mind* a more sophisticated theory. But for purposes of exposition let us just work with (PT) and investigate *its* plausibility.

In fact, it quickly becomes clear that (PT) could have some odd consequences. Most importantly, (PT) implies that it's not the case that all pleasures enhance welfare, as one might have thought they do. Precisely which pleasures enhance welfare according to (PT)

[8] Parfit doesn't say anything about pain in the passage given above. So I leave this step in the calculation intentionally vague. Perhaps we should take it that there is a list of Good Pains and Bad pains that are to be used in the same was as the list of Good Pleasures and Bad Pleasures. Then again, perhaps not. Reasons for why not will emerge in discussing Feldman's theory.

will depend on what items are included on the Objective List of goods. Only pleasure taken in items on this list, according to (PT), enhance welfare. So what's clear is that if (PT) is to have any plausibility, the Objective List of the goods enjoyment of which enhance welfare would have to be quite an extensive list. (PT) would be highly implausible if the Objective List contained, for instance, only the four items Parfit mentions: knowledge, rational activity, love and the appreciation of beauty. After all, if the list contains only these four items, what should we say about pleasure taken in fairly decent things not on this list: like drinking beer, having one-night stands, going shopping, driving a Ferrari with the top down, or pulling off a practical joke on your colleague? Do these more worldly sorts of pleasure not get to count as welfare enhancing? That seems implausible. So we must expand the list of goods enjoyment of which would enhance welfare. But how much should it be expanded? Wherever one draws the line, it seems one would open up oneself to the charge of arbitrariness.

Accordingly, there seems to be a better way to design an OAET than the binary way that (PT) suggests. To avoid the charge of arbitrariness, seems advisable to allow that *at least in principle* all pleasures can have some positive impact on one's welfare. However, in order to respect the intuition (which presumably underlies Parfit's theory) that some pleasures might be taken in objects so foul or base as to confer little or no welfare benefit, we could make the following assumption: the raw value of pleasures are to be adjusted according to where their objects fall *on a scale* (concerning, for instance, quality or pleasure-worthiness). Doing this would allow that pleasure taken in some objects contributes very much to well-being, while pleasure taken in other objects contributes very little or not at all to well-being. Taking this route allows us to avoid becoming easy targets for the criticism of arbitrariness.

We will consider several theories of just this sort in detail in a moment. For now, however, the conclusion I want to draw from the above considerations is simply that there seem to be other theories of the Objectively-Adjusted Enjoyment type that are more plausible than (PT).

4.2.2 Darwall

Darwall suggests a different sort of Objectively-Adjusted Enjoyment Theory. His view seems to be roughly that a person's life contains well-being to the extent that he that he *appreciates things of value*. As Darwall describes his view:

> The normative claim I shall defend is that the best life for a person (in terms of welfare) is one involving activities that bring her into an appreciative rapport with various forms of agent-neutral value, such as beauty, the worth of living things, and so on. (Darwall, 2002, p. 17)

In order to understand Darwall's view, however, a couple of the terms he uses need to be clarified. For one thing, we need to know what makes something possess 'agent-neutral value.' To have agent-neutral value seems to be a matter of *really* being intrinsically valuable, as opposed to being just apparently so. What makes something have intrinsic value is a difficult axiological question, and Darwall does not seem to offer a completely worked out answer to it. He seems to take it that most people will have an intuitive grasp of the notion. In section 4.3, however, we will discuss this sort of question in more depth.

More importantly, we need to know what it means to be in 'appreciative rapport' with things of value. Does Darwall mean *enjoying* things of value? Does he mean *pursuing* them? One thing seems clear, though. On Darwall's view, it does seem to be the case that in order for you to 'appreciate' things of value, you must at a minimum *know* that they are valuable. Consider the following statement of Darwall's:

> There is a way of appreciatively engaging in valuable activities that involves an experienced rapport to the value as exemplified in particular activities. We come to appreciate the value of the activity through a distinctively evaluative mode of awareness we have towards the activity itself. (Darwall, p. 18)

Insofar as I am able to understand these sentences, I think they show that Darwall endorses the idea that being in appreciative rapport with something of value requires knowing (or correctly judging) that it is of value.

Moreover, I think we should take it that 'being in appreciative rapport' with things of value involves enjoying them (or, as Darwall might prefer to put it, their worth). For one thing, it is not a stretch of language to take it that appreciating things of value involves enjoying them. Enjoying something seems to be one standard ways of 'appreciating it.' But what's more, Darwall himself seems to recognize an important connection between enjoyment and appreciating things of value. For instance, he says with respect to activities that are valuable:

> pleasure is a sign of the activities' value, not its substance. What is pleasurable is, at least partly, the appreciation of merit and worth that these activities themselves involve. And what makes these pleasures loom so large in our welfare is the sense that, through them, we are connecting with things that matter. (Darwall, 2002, p. 95)

He then goes on to say that 'the primary source of prudential value is a *felt appreciation* of valuable activity, and not just belief or knowledge…' (Darwall, 95). Thus it seems to me reasonable to take it that Darwall thinks 'being in appreciative rapport' with something of value involves not only correctly judging this thing to be valuable, but also enjoying or receiving pleasure from it.

Thus I suggest the following characterization of the notion of being in 'appreciative rapport' with things of value:

> P is in **appreciative rapport** with something of value, A =df. 1) P correctly judges that A is valuable and 2) P (in some way or other) enjoys A.

In order to state Darwall's theory in a plausible way, we should also take it that it is possible for one's appreciative rapport with something to last for longer or shorter periods of time, and for it to be more intense or less intense. Thus we may say that

> The **strength** of P's episode of appreciative rapport with something of value, A, equals X =df. 1) during the episode P correctly judges that A is valuable to degree Y, and 2) the intensity times the duration of the enjoyment that P receives from A during the episode equals Z, and 3) X equals the average of Y and Z.

With these definitions in place, I propose the following statement of Darwall's theory:

> (DT) Here is how to determine the amount of well-being in P's life:
> 1) Find all the episodes of P's being in appreciative rapport with things of value that occur during P's life.
> 2) For each one, determine its strength.
> 3) Add up all these strengths.
> 4) From this sum, subtract the total amount of displeasure contained in P's life.
> 5) The number you end up with equals the amount of well-being contained in P's life.

Perhaps one might object that step 4) of (DT) is not supported by much that Darwall says himself. This is true. But I am including step 4) in order to make the theory as plausible as possible. Clearly pain, displeasure and such things must have some negative impact on well-being. But the precise way in which this negative impact is supposed to occur is not clear from Darwall's discussion. So the vaguely formulated step 4) in (DT) will have to suffice.

(DT) counts as an Objectively-Adjusted Enjoyment Theory. After all, enjoyments count towards well-being on this theory, and the degree to which they enhance well-being is to be adjusted by the degree to which the objects of one's enjoyments are intrinsically valuable and one knows it. Thus there is a mind-independent variable that modifies the contributions to well-being that episodes of enjoyment make. In particular, on Darwall's theory, this variable is the intrinsic value of the objects with which one is in appreciative rapport.

As we will see in the next section, the OAETs face some common problems. Darwall's theory suffers from these as well. However, there is an additional problem that makes Darwall's theory even more implausible than the next two versions we will consider. Thus I do not think Darwall's theory represents the best avenue to pursue for those philosophers who are sympathetic to the OAETs.

In particular, the problem is this. First note that one can enjoy things that are objectively valuable while failing to *think of them as valuable*. For instance, suppose I receive a great amount of pleasure from listening to a performance of Brahms' Piano Quintet, even though I never once during the whole concert consciously entertain the thought that this piece of music is beautiful or has intrinsic value. Perhaps I am so wrapped up in the moment that this thought never crosses my mind. According to (DT), the enjoyment I receive from listening to the concert would not enhance my welfare at all. After all, because I fail to judge the experience to be intrinsically valuable, it is not an experience that qualifies as an episode of my being in appreciative rapport with something of value. However, it should be intuitive that the pleasure I receive from listening to the beautiful concert *does* indeed have some positive impact on my well-being. (Moreover, given the spirit of Darwall's theory, this is presumably an intuition with which he would himself agree.) Thus (DT) is problematic. It fails to account for the welfare benefits of pleasure taken in valuable things that are not explicitly thought of *as valuable*. Accordingly, I think there are more plausible OAETs to be found.

4.2.3 Adams

Adams defends the theory, roughly, that a person's life contains well-being to the extent that it contains enjoyment of the excellent. As Adams describes his view:

> I shall argue that the principal thing that can be non-instrumentally good for a person is a life that is hers, and that two criteria (perhaps not the only criteria) for a life being a good one for a person are that she should enjoy it, and that what she enjoys should be, in some objective sense, excellent. Its being more excellent, and her enjoying it more, will both be reasons for thinking it better for her, other things being equal… (Adams, 1999, pp. 93-94)

Adams' view resembles Parfit's in that Adams too suggests that two conditions must be met in order for a person to have a good life: that the life is enjoyed, and that the objects of this enjoyment are (in some objective sense) valuable. Adams provides the following defense of the first condition:

> You may be very virtuous; you may be brilliant, beautiful, successful, rich, and famous; but if you do not enjoy your life, it cannot plausibly be called a good life *for you.* (Adams, 1999, p. 95)[9]

And he goes on to defend the second condition as follows:

> The most controversial part of my thesis about a person's good (…) is that it depends on the excellence of what she enjoys. (…) Few parents would desire for their children a lifetime of narcotic highs, no matter how much they would be enjoyed. We do not regard such pleasures, in any amount or intensity, as an acceptable substitute for friendship, knowledge, or accomplishment. (Adams, 1999, p. 97)[10]

It should be noted, however, that Adams' view is different from (PT) in a very important respect. Unlike (PT), Adams takes it that there is scale on which the objects of enjoyment can be rated, in such a way that the higher the object of the enjoyment falls on this scale the more the enjoyment enhances one's well-being. As Adams puts it, something's 'being more excellent, and her enjoying it more, will both be reasons for thinking it better for her…' (Adams, p. 94). Thus Adams' view does not suffer from the troubling consequences that led us to abandon (PT).

On the basis of the above passages, Adams' view seems to be that the degree to which one is well-off equals the degree to which one enjoys things that are excellent, in such a way both that the more one enjoys something, the more it enhances one's welfare, and the greater the excellence of what one enjoys, the more one's welfare is enhanced. I will assume that Adams' talk of enjoyment can be understood in terms of pleasure – not sensory pleasure, but the sort of *attitudinal pleasure* that was discussed in connection

[9] This could mean one of two things: 1) a life that is objectively good (excellent) but contains no enjoyment has some value but not very much – i.e. it is not a *good* life, or 2) such a life is completely *worthless*. Adams' theory, as I interpret it, implies the latter claim. But I will argue in a later section of this paper this is quite implausible.

[10] Note that Adams is employing the Crib Test here. For a discussion of this test, see chapter 1, concerning Darwall's proposal of what is distinctive of welfare value.

with Mill (and will again be discussed in connection with Feldman).[11] A precise formulation of Adams view, so interpreted, would be this:

(AT) Here is how to determine the amount of well-being in P’s life:

1) Find all the episodes of attitudinal pleasure in P’s life.
2) For each one, find its *raw value* by multiplying its intensity by its duration.
3) For each one, find its *adjusted value* by multiplying its raw value by the degree to which the object of this episode of attitudinal pleasure is excellent.
4) Add up the adjusted values of all the episodes of attitudinal pleasure in P’s life. This is the *total excellence-adjusted pleasure* contained in P’s life.
5) Find all the episodes of attitudinal pain in P’s life.
6) For each one, find its raw value by multiplying its intensity by its duration.
7) Add up the raw values of all the episodes of attitudinal pain in P’s life. This is the *total displeasure* contained in P’s life.[12]
8) Subtract the total displeasure in P’s life from the total excellence-adjusted pleasure contained in P’s life.
9) The number you end up with equals the amount of well-being contained in P’s life.

To understand (AT), we need to know what makes the object of one episode of enjoyment more excellent than the object of another episode of enjoyment. On Adams’ view, as we will see, excellence is a matter of resembling God. However, I will postpone detailed discussion of this issue until section 4.3. After all, this sort of question arises for all the theories discussed in this paper. They all have to deal with the question of how to understand the variable that they say modifies the welfare values of episodes of enjoyment. Thus it will be convenient to discuss questions of this sort all in one place.

For now, note that (AT) counts as an Objectively-Adjusted Enjoyment Theory. After all, episodes of enjoyment count towards well-being on this theory, and the degree to which such episodes enhance well-being is to be adjusted by the degree to which the objects of one’s enjoyments display excellence. And excellence, on any plausible

[11] I am thus taking Adams’ theory to be such that an appropriate name for it would be ‘Excellence-Adjusted Intrinsic Attitudinal Hedonism.’ Adams would presumably not be happy with the label of ‘Hedonism.’ But that’s all it really is: a label.

[12] Why didn’t I include any kind of excellence adjustment for the value of episodes of pain? I will discuss this more when I get to Feldman’s view, but roughly the idea is that it would seem pretty implausible to take it that pain taken in more excellent objects is *worse* than pain taken in baser objects. Suppose that winning Olympic medals is excellent. Suppose I am *pained* by the fact that your kid won an Olympic gold medal in swimming, while my kid only got the silver. Now consider pain taken in something very non-excellent, say, my being tortured for no reason. Suppose the intensities and durations of the two episodes of pain are equal. Would it be plausible to take it that the first episode of pain is worse for me than the second episode? That would be highly implausible, I think. My view is that we should not excellence-adjust pains at all. But more on this later.

understanding of it, would seem to be a clear example of a mind-independent variable. Therefore, according to (AT) there is going to be an extra-mental variable that modifies the contributions to well-being that episodes of enjoyment make. So (AT) will count as an OAET.

I take it that (AT) captures the view Adams defends in chapter 3 of his book. However, I should point out that he hedges a little bit, and says he feels the force of certain intuitions that are not consistent with (AT). To begin with, he asks:

> are enjoyment and excellence constituents of our good independently of each other; or is it the enjoyment only *of* excellence, and excellence only *as* enjoyed, that is good for us, as my phrasing has suggested? (Adams, 1999, p. 100)

In answering this question, Adams first makes a claim that is consistent with (AT), namely that he doubts 'that enjoyment of what is not in any way or degree excellent can be a constituent of our good' (Adams, 1999, p. 100). This is consistent with (AT) because (AT) implies that an episode of enjoyment would count towards one's well-being only if the object of the enjoyment is something that has an excellence rating that is greater than zero. However, Adams then goes on to make a claim that is *not* consistent with (AT):

> Probably [the excellence of what is not enjoyed at all] can [constitute part of one's well-being], for it is plausible to think it remains better for oneself to do what is excellent when no available course of action affords any enjoyment. But a life rich only in that sort of excellence is no life to wish on a friend. (Adams, 1999, p. 101)

Adams seems to be expressing some support for the intuition that if no enjoyment is available to one, a course of action that increases the degree to which one's life contains things that are excellent might nonetheless enhance one's welfare somewhat. This is not consistent with (AT) because (AT) implies that things that are excellent can enhance one's welfare only if they are enjoyed to a degree greater than zero. If something is not enjoyed at all, then no matter how excellent, it cannot be of any benefit to one.

Adams is here acknowledging a potential limitation of a view like (AT), and in section 4.4, we will discuss in greater detail just this kind of limitation to theories like (AT). However, since (AT) is supported by almost everything Adams says *except* this one passage just noted, let us assume that (AT) is the theory that Adams at the end of the day would endorse. Before evaluating the OAETs, let us briefly look at one last one.

4.2.4 Feldman

Feldman, too, defends an Objectively-Adjusted Enjoyment Theory. Feldman's first step is to present a generic form of Attitudinal Hedonism whose special feature is that the fundamental bearers of welfare value are episodes of *attitudinal* pleasure, as opposed to, say, sensory pleasure. He calls this generic theory Intrinsic Attitudinal Hedonism (IAH) and it allows of all sorts of modifications. In particular, the contributions to one's well-being that episodes of attitudinal pleasure make can be adjusted by factors such as the *truth* of the propositions one takes pleasure in, or their *altitude* (or 'loftiness' or 'spirituality') or their *pleasure-worthiness* (or desert). One ends up with different versions of Intrinsic Attitudinal Hedonism depending on what one takes the welfare contributions of episodes of attitudinal pleasure to be modified by. Taking it that the value of attitudinal pleasures should be modified by the truth, altitude or desert of the objects of these pleasures will result, respectively, in the theories known as Truth-Adjusted Intrinsic Attitudinal Hedonism (TAIAH)[13], Altitude-Adjusted Intrinsic Attitudinal Hedonism (AAIAH)[14], and Desert-Adjusted Intrinsic Attitudinal Hedonism (DAIAH)[15]. In this chapter, I focus on just on DAIAH since this seems to be the theory that Feldman thinks has the most going for it. I begin by presenting Intrinsic Attitudinal Hedonism in general, and then move on to discuss what characterizes DAIAH in particular.

Feldman's own statement of IAH goes like this:

> i. Every episode of intrinsic attitudinal pleasure is intrinsically good; every episode of intrinsic attitudinal pain is intrinsically bad.
> ii. The intrinsic *value of an episode* of intrinsic attitudinal pleasure is equal to the amount of pleasure contained in that episode; the intrinsic value of an episode of intrinsic attitudinal pain is equal to – (the amount of pain contained in that episode).
> iii. The intrinsic *value of a life* is entirely determined by the intrinsic values of the episodes of intrinsic attitudinal pleasure and pain contained in the life, in such a way that one life is intrinsically better than another if and only if the net amount of intrinsic attitudinal pleasure in the one is greater than the net amount of that sort of pleasure in the other. (Feldman, 2004, p. 66)

The three components in this statement of the theory answer i) the question of what the fundamental bearers of welfare value are, ii) the question of how to calculate the value of

[13] Cf. Feldman, 2004, pp. 112-114
[14] Cf. Feldman, 2004, p. 73
[15] Cf. Feldman, 2004, p. 120

an episode of intrinsic attitudinal pleasure, and iii) how to calculate the welfare value of an entire life.[16]

To fully understand this theory, the main concept that needs to be explained is that of intrinsic attitudinal pleasure. Attitudinal pleasure is different from sensory pleasure in that the former is pleasure one receives upon considering some state of affairs, while the latter is pleasure that, roughly speaking, one feels somewhere on one's body. Feldman explains the basic idea of what an attitudinal pleasure is as follows: 'A person takes attitudinal pleasure in some state of affairs if he enjoys it, is pleased about it, is glad that it is happening, is delighted by it.' (Feldman, 2004, p. 56) Now what makes it the case that an attitudinal pleasure is specifically of the *intrinsic* sort is that one enjoys or is pleased by the object of the pleasure *for its own sake*, not for the sake of something else. To be more precise, we can say that S takes *intrinsic attitudinal pleasure* in some state of affairs p iff S 'takes pleasure in some state of affairs, p, and there is no other state of affairs, q, such that he takes pleasure in p in virtue of the fact that he takes pleasure in q.' (Feldman, 2004, p. 58)

Several further assumptions about attitudinal pleasures should be noted. First, although attitudinal pleasures are not sensory pleasures, the two kinds of pleasure nonetheless 'go hand in hand.' As Feldman puts it

> As I see it, it is not possible for someone to experience sensory pleasure at a time without also experiencing attitudinal pleasure. That is because I think we can define sensory pleasures as feelings in which the feeler takes intrinsic attitudinal pleasure. (Feldman, 2004, p. 57)

Personally, I have some reservations about this definition of sensory pleasure,[17] but that is not a major concern right now. The key point is that on Feldman's view, having an episode of sensory pleasure entails having an episode of attitudinal pleasure. Thus any pleasure that would qualify as welfare enhancing on SSH would entail the existence of something that is just as welfare enhancing on IAH. (IAH, however, allows that other pleasures are indeed welfare enhancing even though they would not count towards

[16] In stating Parfit's, Darwall's and Adam's theories, I focused exclusively on explaining how to calculate the welfare value of an entire life – i.e. to question iii) here. I did this because I assumed that the answers to questions i) and ii) would be apparent from the answer to question iii).

[17] In particular, it seems to me that a person can take attitudinal pleasure in certain *feelings* – say in one's feelings of happiness, or one's good mood – even though it would not be correct to say that this pleasure is a *sensory pleasure*. To fix the definition of sensory pleasure, I think we might have to add the requirement that the feeling in which one takes attitudinal pleasure be a feeling that has a location somewhere on or in the body.

welfare on SSH.) The second assumption about attitudinal pleasure that should be noted is that attitudinal pleasures imply belief. If you are attitudinally pleased by some state of affairs, this entails that you believe that this state of affairs obtains. Third, attitudinal pleasures are not factive. Being attitudinally pleased by some state of affairs does not imply that this state of affairs *actually does* obtain. Fourth and finally, it is possible to take pleasure in future or past states of affairs, not just present ones. (In general, it seems possible to be attitudinally pleased by any state of affairs such that it is possible for one to believe that it obtains.)

Why make use of attitudinal pleasures in one's theory of welfare at all? Feldman explains the advantage of employing the notion of attitudinal pleasure as follows:

> Attitudinal pleasures, unlike sensory pleasures, have objects. (…) Attitudinal pleasure is always pleasure taken in some state of affairs. This feature of attitudinal pleasures makes it possible for IAH to take many forms, depending upon restrictions that we may place on the sorts of objects in which attitudinal pleasure is taken. (Feldman, 2004, p. 71)

So because attitudinal pleasures have objects, while sensory pleasures are brute experiences, it is straightforward to adjust the value of episodes of attitudinal pleasure, while this is not the case for sensory pleasures. In particular, the value of episodes of attitudinal pleasure can be adjusted by the features of their objects. But the value of episodes of sensory pleasure cannot be adjusted in this way because they don't have objects. Granted, their values could perhaps be adjusted by appeal to their phenomenological properties. But there might be all sorts of interesting differences between two phenomenologically identical sensory pleasures – e.g. one might be merely the result of a hallucination while the other is not. Thus formulating one's theory of welfare in terms of the notion of attitudinal pleasure makes possible a range of interesting kinds of adjustment that would not be available if one's theory of welfare merely appealed to the notion of sensory pleasure.

The type of adjustment that I will focus on here is desert-adjustment or, what comes to the same thing, adjustment for pleasure-worthiness. Feldman introduces the notion of desert-adjustment in response to a powerful objection to traditional non-adjusted forms of Hedonism called the Bestiality Argument. The argument can be formulated using a case that Feldman describes as follows:

> Imagine a person – we can call him 'Porky' – who spends all his time in the pigsty, engaging in the most obscene sexual activities imaginable. I stipulate that Porky derives great pleasure from these

> activities and the feelings they stimulate. Let us imagine that Porky happily carries on like this for many years. Imagine also that Porky has no human friends, has no other sources of pleasure, and has no interesting knowledge. Let us also that Porky somehow avoids pains – he is never injured by the pigs, he does not come down with any barnyard diseases, he does not suffer from loneliness or boredom. (Feldman, 2004, p. 40)

The objection, then, is that standard non-adjusted forms of Hedonism, like SSH and IAH, imply that Porky's life is exceptionally high in welfare value – an intuitively implausible result. To circumvent this objection, Feldman proposes moving to an adjusted form of attitudinal Hedonism. In particular, he proposes that intrinsic values of episodes of attitudinal pleasure be adjusted according to the degree to which their objects deserve to have pleasure taken in them. Modifying the theory in this way would yield the more intuitive result that Porky's life is not very good for him. And so the objection would be avoided.

What is it, more specifically, for something to deserve to have pleasure taken in it, in other words, to be pleasure-worthy? According to Feldman,

> the value of a pleasure is enhanced when it is pleasure taken in a pleasure-worthy object, such as something good or beautiful. The value of a pleasure is mitigated when it is pleasure taken in a pleasure-unworthy object, such as something evil, or ugly. The disvalue of a pain is mitigated (the pain is made less bad) when it is pain taken in an object worthy of pain, such as something evil, or ugly. The value of a pain is enhanced (the pain is made yet worse) when it is pain taken in a object unworthy of this attitude, such as something good or beautiful. (Feldman, 2004, p. 120)

Feldman does not say much more than this about exactly what makes something be pleasure-worthy. However, the basic idea behind the theory that Feldman favors should be clear. According to the passage above Feldman thinks that the value of episodes of attitudinal pleasure should be magnified if their objects are more pleasure-worthy, and the disvalue of episodes of attitudinal pain should be lessened if their objects are pain-worthy. This suggests the following formulation of DAIAH, or as I'll call it, Feldman's Theory:

(FT*) Here is how to determine the amount of well-being in P's life:

1) Find all the episodes of intrinsic attitudinal pleasure in P's life.
2) For each one, find its *raw value* by multiplying its intensity by its duration.
3) For each one, find its *desert-adjusted value* by multiplying its raw value by the degree to which the object of this episode of intrinsic attitudinal pleasure deserves to have pleasure taken in it.
4) Add up the desert-adjusted values of all the episodes of intrinsic attitudinal pleasure in P's life. This is the *total desert-adjusted pleasure* contained in P's life.
5) Find all the episodes of intrinsic attitudinal pain in P's life.

6) For each one, find its *raw disvalue* by multiplying its intensity by its duration.
7) For each one, find its *desert-adjusted disvalue* by multiplying its raw disvalue by the degree to which the object of this episode of intrinsic attitudinal pain deserves to have pain taken in it.
8) Add up the desert-adjusted disvalues of all the episodes of intrinsic attitudinal pain in P's life. This is the *total desert-adjusted displeasure* contained in P's life.
9) Subtract the total desert-adjusted displeasure in P's life from the total desert-adjusted pleasure in P's life.
10) The number you end up with equals the amount of well-being contained in P's life.

This formulation of Feldman's theory is the one suggested by Feldman's own comments. However, it is not the formulation of the theory that he should endorse. (FT*) has a major problem. In particular, the problem arises because of steps 7) and 8). I believe it is a mistake to take it that the disvalue of episodes of intrinsic attitudinal pain should be adjusted for desert. Why is this?

In the passage above, Feldman says that the disvalue of an attitudinal pain should be lessened when the object of that pain is pain-worthy – for instance when the object is 'something evil, or ugly.' Moreover, he says that the disvalue of an attitudinal pain should be made *greater* when the object of that pain is pleasure-worthy – for instance, when the object is 'something good or beautiful.' But now compare two lives. Suppose the first life contains a certain amount of pain taken in objects that are very pain-worthy, while the second life contains that very same amount of pain but taken in objects that are very pain-*un*worthy. To make it concrete, suppose that in Life 1, some soldiers in an invading army rape and kill your family members. This causes you a tremendous amount of attitudinal pain. Moreover, insofar as anything is worthy of having pain taken in it, the objects of the pains in Life 1 are *highly* worthy of having pain taken in them. By contrast, suppose that in Life 2, you are pained to this same extraordinary degree by the fact that your kids make more money than you do, that cancer is cured and that interracial marriage is made legal in your society. The objects of these pains are highly pain-*un*worthy if anything is. Thus since (FT*) implies that pain-worthiness mitigates the disvalue of episodes of pain while pain-unworthiness enhances the disvalue of episodes of pain, (FT*) implies that Life 1 would be better for you than Life 2. That seems to be an extremely counter-intuitive result. Not many people, I suspect, could accept it.

Thus it seems to be a mistake to allow that the disvalue of episodes of attitudinal pain should be modified by the pain-worthiness of the objects of these pains. It seems much more plausible to formulate the theory in such a way that the disvalue of any attitudinal pain, no matter what the pain-worthiness of its objects, should be a function solely of its intensity and duration. If this is the right response to the present objection, it suggests that Feldman's theory is best formulated as follows:

(FT) Here is how to determine the amount of well-being in P's life:

1) Find all the episodes of intrinsic attitudinal pleasure in P's life.
2) For each one, find its *raw value* by multiplying its intensity by its duration.
3) For each one, find its *desert-adjusted value* by multiplying its raw value by the degree to which the object of this episode of intrinsic attitudinal pleasure deserves to have pleasure taken in it.
4) Add up the desert-adjusted values of all the episodes of intrinsic attitudinal pleasure in P's life. This is the *total desert-adjusted pleasure* contained in P's life.
5) Find all the episodes of intrinsic attitudinal pain in P's life.
6) For each one, find its *raw disvalue* by multiplying its intensity by its duration.
7) Add up the raw disvalues of all the episodes of intrinsic attitudinal pain in P's life. This is the *total displeasure* contained in P's life.
8) Subtract the total displeasure in P's life from the total desert-adjusted pleasure in P's life.
9) The number you end up with equals the amount of well-being contained in P's life.

(FT) differs from (FT*) only with respect to the steps that come after step 5). (FT), I take it, is the superior formulation of Feldman's Desert-Adjusted Intrinsic Attitudinal Hedonsim. It should be obvious that (FT) is an Objectively-Adjusted Enjoyment Theory. After all, episodes of enjoyment count towards well-being on this theory, and the degree to which such episodes enhance well-being is to be adjusted by the degree to which their objects are pleasure-worthy. Since pleasure-worthiness seems to be a clear example of a mind-independent variable, (FT) counts as an Objective-Adjustment theory. In the next sections, we will discuss some of the major challenges that (FT) has in common with other Objectively-Adjusted Enjoyment Theories.

4.3 What Grounds Adjustment?

At this point, we have seen reasons for being dissatisfied with Parfit's and Darwall's theories. Of the Objectively-Adjusted Enjoyment Theories, therefore, the theories defended by Adams and Feldman seem to be the best candidates. However, these theories too face a number of challenges. In this section, I will discuss what is perhaps the most obvious question they face. Adams thinks that the intrinsic welfare values of episodes of enjoyment should be modified by the degree to which the objects of these enjoyments are excellent. Feldman thinks that the intrinsic welfare values of episodes of attitudinal pleasure should be modified by the degree to which the objects of these pleasures are pleasure-worthy. But what exactly is it that makes a given object of enjoyment more excellent or more pleasure-worthy? In order for a theory like Adams' or Feldman's to be complete, it needs an account of the objective variable that is supposed to modify the welfare values of the relevant episodes of enjoyment. The theory will be plausible only insofar as such an account can be given.

In this section, I will argue that Adams' theory is doubtful because the notion of excellence he adopts is problematic. Therefore I do not think that Adams' approach represents the best way to defend an Objectively-Adjusted Enjoyment Theory. After this, I will turn to Feldman's theory. Feldman himself does not provide a systematic account of what makes a given state of affairs pleasure-worthy. But I think that such an account can be given and I will propose one. Then I will argue that this account should be modified in a certain way because doing so would answer two prima facie compelling arguments against DAIAH. Once this work is completed, we will be left with the conclusion that DAIAH, supplemented with the account I propose, represents the most plausible way to defend an OAET. Nonetheless, in the final section of this chapter, I will argue that even this most promising of OAETs faces problems.

4.3.1 Adjusting for Excellence

We saw Adams endorse a theory according to which the welfare value of episodes of enjoyment are to be modified by the excellence of the objects of these enjoyments. The more excellent the object is, the more intrinsic welfare value enjoying it would have. But exactly what does this excellence consist in? Adam's view is a theistic one according to which, roughly, 'things are excellent insofar as they resemble or imitate God.' (Adams, 1999, p. 23)[18] Thus, as a first sketch of Adams' view, we may understand excellence as follows:

> E1. An object of enjoyment, O, is excellent to degree d iff O resembles God to degree d.

In order to fully understand E1, however, we need to know how to understand the notion of resemblance. Adams thinks it should be uncontroversial that being beautiful is one way in which something may resemble God. Being sublime seems to be another.[19] In general, the idea behind E1 seems to be that for any (degreed) property, p, that God would have if he existed, the greater the extent to which some object, O, possesses p, the more O resembles God. The operative assumption here is that God has all of his properties (at least, all the important ones) to an infinite degree. So the greater the degree to which an object possesses one of God's properties, the more like God that object would be. Sticking with this assumption, then, we may take it that the level of resemblance between an object, O, and God can be calculated as follows:

1) Find all the properties, p1, p2… pn, that O shares with God.
2) For each shared property, determine the degree to which O possesses that property.
3) Add up the degrees to which O possesses each of the shared properties.
4) This sum equals the degree of resemblance between O and God.

With this method now in place for determining degrees of resemblance, the idea behind E1 should be relatively clear.

Note that this account of excellence does not require the existence of God.[20] After all, it is obvious that X may resemble Y, even if Y does not exist.[21] If I were a talented detective and wore double-brimmed hats and spent a lot of time looking at things through a magnifying glass, then I would strongly resemble Sherlock Holmes even though he

[18] See especially Adams, 1999, ch. 1, §3-4

[19] See Adams, 1999, pp. 38-41

[20] Adams, for some odd reason, seems to deny this. He says that a 'theistic theory of the nature of excellence obviously presupposes or implies the existence of God.' (Adams, 1999, p. 28) But it clearly does no such thing. Adams is, it seems to me, is just mistaken about this.

[21] It might seem that in order for it to be true that X resembles Y, at the very least either *X or Y* must exist, even if both do not have to. I'm not sure about this, though. Anyway, it doesn't matter for our purposes.

never existed. Similarly, it would clearly be possible for an object to resemble God even if God did not exist.

However, Adams points out that E1 is not an adequate account of excellence. In particular, it faces a certain kind of problem, which is apparent in the following passage:

> Consider the phenomenon of parody or caricature. Parodies and caricatures do resemble, but do not in general share the excellences of their original or object. (…) Even something so abstract and free of superfluous properties as a beautiful piece of music can be parodied; and the parody will resemble the original but will not thereby share its virtues. Perhaps one could plausibly maintain that the divine goodness, uniquely, is such that it cannot be parodied or caricatured. But I would not know how to argue for that, and there seem to be counterinstances. It is natural enough to say that Hitler's power is "a caricature of the divine power" – more natural, I suspect than to deny flatly that his power resembles God's in any way. (Adams, 1999, p. 33)

The problem for E1 that this passage suggests is that it seems that certain things may resemble God with respect to certain properties, but would still not thereby be made any more excellent. As examples of this, Adams mentions parodies or caricatures of God. However, to my mind, the most striking example he suggests is that of Hitler's power. The argument against E1 based on this example would go something like this. In virtue of being very powerful, Hitler resembles God. Thus E1 implies that Hitler's power makes him more excellent. But this seems counter-intuitive. Considering the horrible evil to which this power was put, Hitler's power intuitively does not make him more excellent. Thus E1 is false. Similar arguments could be constructed on the basis of Adams' other examples, involving parodies or caricatures of God.

In order to get around this sort of problem, Adams proposes a solution that appeals the idea of *faithful resemblance*. He explains it as follows:

> This suggests a modification of the analysis of excellence in terms of resembling or imaging God. We can suppose that the difference between resemblances to God that do and do not constitute virtues or excellences is analogous to the difference between good portraits (by which I mean faithful portraits) and caricatures. (…) I will not offer here a full account of what the faithfulness of a portrait amounts to, and I am not sure that I could give one. It would surely include the observation that caricatures are *distorted* in a way that faithful portraits are not. The caricature exaggerates one or more features of the original, whereas the faithful portrait represents features in a balanced way and in relation to those other features to which they are most importantly related in the original. (Adams, 1999, p. 33).

Adams does not offer a precise account of faithful resemblance. However, focusing on Adams' talk of exaggerations of some features of the original, it seems we may take the basic idea to be this.[22] An object *faithfully* resembles God when it shares some properties with God and it has these properties in exactly the same proportions as God has them. By contrast, an object resembles God *unfaithfully* when it shares some properties with God but it has these properties in rather different proportions than the proportions in which God has them. In general, the greater the difference between a) the proportions in which an object has certain properties and b) the proportions in which God has those properties, the more *unfaithful* the resemblance between the object and God will be.

This idea can be used to modify Adams' account of excellence. The thought is that the excellence of an object is diminished by unfaithful resemblance. An object that unfaithfully resembles God is not as excellent as an object that faithfully resembles God. The less faithful the resemblance between an object and God, the less excellent it is. Accordingly, we could take the excellence of an object to be the product of two things: a) the degree to which it resembles God (as understood in E1), and b) the faithfulness of this resemblance. To do this, let us suppose that the faithfulness of a resemblance is to be represented by a number between 1 and 0, where 1 represents a perfectly faithful resemblance and 0 represents a perfectly unfaithful resemblance.[23] With this assumption in place, we can take Adams' modified account excellence to be this:

[22] This interpretation of Adams notion of faithful resemblance is similar to the more detailed and precise account that Scott Hill develops in his paper, 'Goodness is Being Like God: Adams' Theistic Axiology.' I am heavily indebted to Hill for helping me understand Adams' view.

[23] It is hard to explain how the faithfulness of a resemblance between an object and God is to be understood. However, I have tried to figure it out. In particular, I will first explain how to determine the level of *dissimilarity* between the proportions in which an object has certain properties and the proportions in which God has them. Then I will use this to define the *faithfulness* of the resemblance between the object and God. However, please feel free to skip this footnote if you're not interested in the details of this.

To explain how to calculate the faithfulness of a resemblance, I first need to define some terms. Suppose O shares three properties with God: A, B and C. Let 'a_g', 'b_g' and 'c_g' stand for the degrees to which God possesses A, B and C, respectively. Let 'a_o', 'b_o' and 'c_o' stand for the degrees to which the object O possesses A, B and C, respectively. Also, I make one simplifying assumption. Although God is typically thought to have many of his properties (viz. love, power, wisdom, etc.) to an infinite degree, to keep things simple, let us take A, B and C to be properties that God possesses to a *finite* degree. (Perhaps the properties in question are a sense of humor, a preference for Aston Martins over BMWs and a desire to see Led Zeppelin reunited.) Now we can characterize the level of unfaithfulness of the resemblance between O and God, with respect to the properties A, B and C, as follows:

1) **First, determine the proportions in which God has A, B and C.** To do this, take the ratios between the degrees to which God possesses these properties. That is, find the values for the following three ratios: a_g/b_g, a_g/c_g, and b_g/c_g. (These are the only proportions we need to take account of. For the other three possible ratios here carry no new information.) Let 'R_g1' refer to the value of the first of these ratios, 'R_g2' to the value of the second ratio, and 'R_g3' to the value of the third.

> E2. An object of enjoyment, O, is excellent to degree d iff d equals [the degree to which O resembles God *times* the faithfulness of the resemblance between O and God].

The faithfulness of the resemblance between an object O and God is a measure of the similarity between the proportions in which God has certain properties and the proportions in which O has these properties. It will always be a number between 0 and 1, where 1 represents a perfect match with respect to these proportions, and 0 represents a perfect mismatch.

2) **Next, determine the proportions in which the object O has A, B and C.** To do this, take the ratios between the degrees to which O possesses these properties. That is, find the values for the following three ratios: a_o/b_o, a_o/c_o, and b_o/c_o. Let 'R_o1' refer to the value of the first of these ratios, 'R_o2' to the value of the second ratio, and 'R_o3' to the value of the third.

3) **Now compare the proportions in which O has A, B and C to the proportions in which God has A, B and C. This will give you the *Raw Unfaithfulness* of the resemblance between O and God.** To do this, take the absolute value of (R_g1 - R_o1), (R_g2 - R_o2) and (R_g3 - R_o3). Then add up these three values. This is the degree to which the proportions in which O has A, B and C **differs** from the proportions in which God has A, B, C. In other words,

$$\textbf{Raw Unfaithfulness}_{\{O,\ G:\ <A,\ B,\ C>\}} = |R_g1 - R_o1| + |R_g2 - R_o2| + |R_g3 - R_o3|$$

After all, if it is the case that $|R_g1 - R_o1| = 0$, and $|R_g2 - R_o2| = 0$, and $|R_g3 - R_o3| = 0$, then the degree of resemblance between O and God is *perfect* with respect to the proportions in which they possess A, B and C. But if, for instance, $|R_g1 - R_o1| > 0$, then the degree of resemblance in the proportions between O and God with respect to the proportions would be *less than perfect*. Thus the larger the difference between R_g1 and R_o1, the less O would resemble God. Ditto for the other two pairs of ratios. Accordingly, the closer to 0 this number is, the greater the faithfulness of the resemblance to God.

4) **Determine the Faithfulness score of the resemblance by assigning an appropriate number between 0 and 1 based on the Raw Unfaithfulness score of the resemblance.** This step is needed in order to get the numbers to work the right way when the notion of faithfulness is plugged into Adams' account of excellence. What is the function that takes you from your Raw Unfaithfulness score to your Faithfulness score? I don't know exactly how to construct this function. However, I can say a few things about it. The Raw Unfaithfulness score will always be a number equal to or greater than 0. As we saw, if the Raw Unfaithfulness score is 0, then the resemblance between O and God is perfect. The larger the Raw Unfaithfulness score is, the more unfaithful the resemblance is. **This means that when the Raw Unfaithfulness score is 0, then the function in question must return a Faithfulness score of 1. Moreover, as the Raw Unfaithfulness score gets higher, the function must return a Faithfulness score that is closer and closer to 0.** There is no upper limit on how unfaithful a resemblance between two objects can be. Thus the function will never *actually* return 0 as the Faithfulness score. **However, as the Raw Unfaithfulness score approaches infinity, the Faithfulness score will approach 0.**

Hopefully, this procedure will give you some sense of how the faithfulness of a resemblance is to be calculated. The procedure is not complete because I don't know how to construct the function from the Raw Unfaithfulness scores (which are always equal to or greater than 0, and have no upper bound) to the Faithfulness scores (which are always between 0 and 1). But I'm hoping that I have at least taken some steps towards clarifying the idea.

Although E2 seems to avoid the Hitler and parody counter-examples to E1, one might still have doubts about E2. Scott Hill has argued (convincingly to my mind) against Adams' faithfulness-based account of excellence, which I think is captured more or less by E2.[24] However, I will raise two different problems here. In particular, I will focus on the problems that arise when one plugs Adams' account of excellence into his theory of welfare. For starters, plugging E2 (or any other theistic account like it) into Adams' theory of welfare seems to have counter-intuitive consequences. After all, it seems that for some of God's attributes, it is not the case that it is better for a person to enjoy things the more they resemble God in respect of these particular attributes. Consider the property of being incorporeal. Or the property of being eternal. Or the property of having all things depend on you for their existence. I don't see why we should suppose that it is better for a person to enjoy things the more they possess these properties. It seems to me that enjoying corporeal things might just as good (and sometimes even better) than enjoying incorporeal things. It also seems that sometimes it is better to enjoy things that last shorter, rather than longer. (For example, it is better to enjoy a meal that lasts the right amount of time than to enjoy and extremely drawn out, seemingly never-ending meal.) And finally, I see no reason to suppose that it is in principle better to enjoy something that lots of other things depend on for their existence than it would be to enjoy something that very few other things depend on. Adams theory of well-being, when supplemented with E2 (or E1 for that matter), would seem to imply that it is indeed better for a person to enjoy things the more incorporeal, eternal or fundamental they are. But I find this implication to be implausible.

There is an additional reason to doubt Adams' theory of well-being when supplemented with a theistic account of excellence of the sort that E2 embodies. I think the problem is most clearly presented as a dilemma. Either God exists or God does not exist, and there are problems in either case. Begin with the first horn of the dilemma. If God exists, then there would be a fact of the matter about what properties God has. And so there would also be a fact of the matter about the degree to which various objects (faithfully) resemble God. However, God is often taken to exist outside the realm of normal experience. If He exists, it is likely to be outside of the empirically observable

[24] See Scott Hill, 'Goodness is Being Like God: Adams' Theistic Axiology.'

universe. Thus there would be significant epistemological problems with figuring out what God's properties are, as well as the degrees to which various objects resemble God. And so for any theory of welfare that accords a central place to the facts about how much various objects of enjoyment resemble God, the implications of this theory are going to be highly unclear.

Perhaps a bigger problem is that many people – myself included – might be inclined to think that God does not exist. Some do, but those of us who don't are not going to be able to accept the first horn of the dilemma. So we are going to have to adopt the second horn of the dilemma instead. In other words, we are going to have to take it that how excellent something is gets determined by the degree to which this thing resembles what God *would* be like if he did exist (even though he doesn't).

Now, there are problems if one adopts this second horn of the dilemma. In particular, how are we to understand what God would be like if he did exist? Well, since we are assuming that God does not exist, we are just going to have to make something up. In other words, we will have to provide some account of what God would be like if he existed – that is, we will have to say what properties God would have, and in what proportions he would have them. Moreover, this account would have to be independently motivated. Perhaps this can be done. Suppose one does manage to provide some independently motivated account of the nature of God. Suppose one says that God has ten basic properties, i.e. A, B, …. J, and that he has all these properties to an infinite degree. In that case, Adams' excellence-adjusted enjoyment theory of welfare would just amount to a theory that says that it is better to enjoy things the more they possess these ten properties A–J. Thus there would be no need to appeal to the concept of God in formulating the theory of welfare. The concept of God would have entirely dropped out of the picture. This is not to say that the theory would as a result be *false*. But it does suggest that, on the assumption that God does not exist, there is no point in formulating an excellence-adjusted theory of welfare that appeals to the concept of God. One could simply formulate the theory so as to say that it is better for a person to enjoy things the more they possess properties A-J.

Thus we may summarize the dilemma for Adams as follows. Either God exists or God does not exist. If God does exist, then although there may be some point in formulating an excellence-adjusted enjoyment theory of welfare that does appeal to the concept of God, it will be exceedingly difficult to get any sense of what the theory implies. What's more, many people (myself included) simply are not going to be able to accept this first horn of the dilemma. However, if one opts for the second horn of the dilemma instead, and assumes that God does not exist, then there will not be any point in formulating an excellence-adjusted enjoyment theory of welfare that appeals to the concept of God. It would be more straightforward and more theoretically elegant to formulate the theory in non-theistic terms. Thus both horns of the dilemma are problematic.

This is not a knock-down argument against Adams' excellence-adjusted enjoyment theory of welfare. However, I do think it raises major doubts about the theory insofar as the notion of excellence it employs is to be understood in terms of resemblance to God. One possible response would be to find a plausible non-theistic account of excellence. But I do not know what that would be.[25] Instead, I propose that we abandon Adams' excellence-adjusted enjoyment theory of welfare, and consider a theory that is similar in structure but that seems to be more plausible because it is cashed out in entirely non-theistic terms. In particular, perhaps we can get a plausible theory of welfare if we jettison the notion of excellence and adopt the notion of pleasure-worthiness instead.

4.3.2 Pleasure-worthiness

Where Adams' theory makes use of the notion of excellence, Feldman's theory, Desert-Adjusted Intrinsic Attitudinal Hedonism (DAIAH), makes use of the notion of

[25] Perhaps one could employ the ancient idea that the excellence of a thing consists in that thing's performing its characteristic function well. This would yield a non-theistic account of excellence. In particular, we would get:

> E3. An object of enjoyment, O, is excellent to degree d iff O performs its characteristic function to degree d.

However, such an account might have problems as well. Not only do many things not have any clear function, but plugging E3 into a theory of well-being like Adams' would have some odd consequences. For instance, consider two episodes of enjoyment. The first episode, which has intensity ten and lasts for ten minutes, is taken in observing a carburetor that works the way it's supposed to. The second episode, which also has intensity ten and lasts for ten minutes, is taken in observing a breathtakingly beautiful landscape. If E3 is adopted for use in Adams' theory, the implication is that the first episode would enhance your welfare significantly more than the second episode. After all, the object of the first episode performs its function very well, whereas this is not the case for the object of the second episode. This, it seems to me, is not a plausible result, however. And so I am inclined to think that E3 does not provide the account of excellence that is needed to save Adams' theory. I do not know what else could be used instead.

pleasure-worthiness. There is more hope, I think, in finding a plausible account of the pleasure-worthiness of states of affairs than there is in finding an account of the excellence of objects of enjoyment. Thus I would be willing to bet more money on DAIAH's being true than Adams' theory.

Feldman does not provide any systematic account of what makes a given state of affairs more worthy of having pleasure taken in it. In *Pleasure and the Good Life* (2004)*,* Feldman mentions beautiful and good states of affairs as examples of ones that would receive high scores with respect to pleasure-worthiness, and states of affairs that are that are ugly, bad, cruel or disgusting as examples of ones that would receive low scores with respect to pleasure-worthiness.[26] In his other major article on this topic, 'The Good Life: A Defense of Attitudinal Hedonism' (2002), Feldman does not say much more by way of a systematic account of pleasure-worthiness. No one else that I know of provides such an account either.

However, in order for DAIAH to be a complete theory, an account is needed of what makes a state of affairs be more pleasure-worthy. One possibility that immediately suggests itself is this:

> PW1) A state of affairs, S, is pleasure-worthy to degree X iff S is 'intrinsically valuable' to degree X.

Intrinsic value here is not supposed to be the same as *welfare* value. That would lead to an obvious circularity. Instead, intrinsic value in this context is the sort of value that makes a *world* better the more of it is present in that world. (This, I believe, is what is known as 'Moorean intrinsic value.')

But on closer insprection, I don't think this is an advisable strategy for understanding pleasure-worthiness. On a certain plausible assumption, which most Utilitarians are going to want to accept, DAIAH in conjunction with PW1 will lead to a worrying circularity. The standard Utilitarian view is roughly that the value of a given possible world or scenario is equal to the sum of the degrees to which the people in that world or scenario are well-off. Thus Utilitarians propose to understand the intrinsic value of worlds or scenarios in terms of welfare value. But this means that it would be circular, for a Utilitarian at least, to turn around and understand welfare value in terms of intrinsic

[26] See Feldman, 2004, p. 120

value, as DAIAH in conjunction with PW1 would have him do. (What's more DAIAH together with PW1 would seem to lead to a similar, but perhaps less obviously vicious circularity even for those who are not Utilitarians, but who nonetheless think that the intrinsic value of a world or scenario is determined *at least in part* by the amount of welfare that people possess in that world or scenario. For even this non-Utilitarian would be understanding intrinsic value at least *in part* in terms of welfare, while DAIAH together with PW1 have it that welfare, again, is to be understood in terms of intrinsic value. Perhaps this circularity is less vicious than the one facing the Utilitarian, but it seems to me to be a worrying kind of circularity nonetheless.)

A more promising strategy might be to seek a list of things that intuitively seem to be especially pleasure-worthy. For instance, G.E. Moore defends the view that two things in particular have intrinsic *world*-value, namely 'the pleasures of human intercourse' and 'enjoyment of beautiful objects.'[27] Thus we might suppose that human intercourse and beautiful objects are the two things that are especially pleasure-worthy. On the other hand, I'm still inclined to think that this two-item list suggested by Moore's views would be a bit too impoverished. Michael Zimmerman (2007) points out that Frankena has provided one of the most comprehensive list of goods in the literature. So perhaps we should appeal to this list in order to account for what states of affairs are especially pleasure-worthy. Zimmerman gives the following summary of Frankena's list:

> life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc. [Presumably a corresponding list of intrinsic evils could be provided.] (Frankena, 1963, pp. 87-88)

This is certainly an impressive list. I think it gives a good indication of what states of affairs we should expect to turn out to be especially pleasure-worthy. With such a list in place, we could proceed to formulate an account of pleasure-worthiness:

> PW2) A state of affairs, S, is pleasure-worthy to degree X iff S instantiates an item on the list to degree X (or if S instantiates more than one item on the list, X is the average of the degrees to which S instantiates the items that it does).

PW2 is clearly superior to PW1 in that it does not make DAIAH turn out to be circular.

[27] See Moore, 1903, ch. VI, § 113

However, I do not think PW2 is a satisfactory account of pleasure-worthiness. In general, providing a list of especially pleasure-worthy states of affairs is not going to be a good strategy for us to pursue. After all, why are these things in particular on the list? In virtue of what does something belong on the list? We do not know yet. Thus the account of pleasure-worthiness that is suggested by PW2 is not *systematic*. To rectify this, I will now propose and defend what I think *is* a systematic account of what makes a state of affairs more pleasure-worthy. In section 4.3.3, I will argue for expanding this notion of pleasure-worthiness in a certain way.

The first step I propose is to adopt what I think should be a pretty uncontroversial *analysis* of pleasure-worthiness:

> PWA) A state of affairs, S, is pleasure-worthy to degree X (or deserves to have pleasure taken in it to degree X) =df. it would be *appropriate* to degree X for one to take pleasure in S.

By itself, PWA is not very illuminating. In particular, we need to know what makes it appropriate to take pleasure in something. The concept of the appropriateness of some response, it seems to me, is the kind of concept that is easily amenable to an ideal observer account. Thus I suggest that the appropriateness of taking pleasure in some state of affairs should be understood in terms of what an ideal observer would approve of taking pleasure in. Of course, this is not the only possible way to understand appropriateness here, but I think it will do. Accordingly, I suggest that PWA be supplemented by the following notion of appropriateness:

> APR) It would be *appropriate* to degree X for one to take pleasure in S iff an ideal observer would *approve* to degree X of one's taking pleasure in S.

To fully understand APR, of course, we need to know what is meant by 'an ideal observer.' I do not have the space to provide a fully worked out account of this here, but in this context, I think it is reasonable to take the ideal observer to be a person who, at a minimum, a) is fully rational, b) knows all the relevant facts about the state of affairs in question, and c) would be deemed to be an open-minded individual, not biased or bigoted.[28]

[28] What if there are many individuals would count as ideal observers and they would differ with respect to what they would approve of taking pleasure in? In that case, we could put it this way: it would be appropriate to degree X for one to take pleasure in a state of affairs S iff the ideal observers, were they to

In order to make APR even more concrete, perhaps we could take it that the facts about what an ideal observer would approve of taking pleasure in are determined by the facts about human nature and about what actual people approve of taking pleasure in. One reasonable way to figure out what items an ideal observer would approve of one's taking pleasure in would be to a) find a representative sample of people from different cultures and historical periods, b) determine what these people would approve of taking pleasure in, and c) provide a theoretical framework that systematizes these judgments. This would, I think, provide a reasonable guide to the facts about what an ideal observer would approve of taking pleasure in.

PWA and APR together provide a systematic account of what it is that makes states of affairs more pleasure-worthy. In a moment I will argue that this account should be expanded in a certain way. However, before that, let me first offer a simple argument for thinking that something like PWA and APR constitutes an acceptable account of pleasure-worthiness. Begin with the following intuitive datum: Frankena has given a good list of some things that are especially worthy of having pleasure taken in them. Thus any plausible account of what makes states of affairs pleasure-worthy will have to explain why the things on Frankena's list are very deserving of having pleasure taken in them. PWA and APR together provide just such an explanation. After all, it seems that an ideal observer would indeed approve highly of one's taking pleasure in the things on Frankena's list. This, I suggest, provides some support for the account of pleasure-worthiness provided by PWA and APR.

4.3.3 A final improvement

I am going to argue in this section that for those who want to defend DAIAH, it would be advisable to expand, in a certain way, the account of pleasure-worthiness given by PWA and APR. The reason is that doing this will allow defenders of DAIAH to avoid two compelling objections to their theory. The first objection has been posed by Michael Depaul.[29] As far as I know, no one has presented the second argument in print. If it weren't for the fact that these objections can be avoided by modifying the account of

discuss the matter among themselves, would reach a consensus to the effect that they collectively approve to degree X of one's taking pleasure in S.

[29] See Depaul, 2002

pleasure-worthiness, I would find them to be compelling arguments. Let me begin by presenting the objections.

DePaul calls his objection the 'Glory Days' objection. I find his own explanation of it to be pretty good:

> Something good happens to Jim Bob – he hits a home run in extra innings to win the high school state championship. Jim Bob has an attitude towards this state of affairs. (…) It turns out that the home run is the high point of Jim Bob's life. After that, poor Jim Bob is more or less a failure at everything he tries. But Jim Bob doesn't let this get him down, he continues to dwell on his one moment of glory, taking great pleasure in this one moment for his entire life. (…)
>
> Now compare Jim Bob to Betty Ann. Something good happened to Betty Ann when she was in high school as well. Betty Ann aced a serve to win the state tennis championship. She is very happy about this state of affairs for awhile. As her life goes on Betty Ann succeeds at many other things. Let's say that in total she has 100 successes and that she enjoys each of these in turn, and that she enjoys each for a moderate period of time. To be more precise, Betty Ann spends 1/100 the time enjoying each of her successes as Jim Bob spends enjoying that one home run. Finally, suppose that Jim Bob and Betty Ann enjoy their successes with equal intensity. (…)
>
> If nothing else happens to please or pain Jim Bob or Betty Ann, it seems that Intrinsic Attitudinal Hedonism is committed to saying that these two lives are equally good. But surely Betty Ann has a much better life than Jim Bob. It does not seem that either of the adjustments for desert can help, since neither Jim Bob nor Betty Ann is taking pleasure in the wrong sorts of things… (Depaul, 2002, p. 630-631)

The argument DePaul is proposing, as applied to DAIAH, is this:

The Glory Days Objection

1) If DAIAH is true, then Jim Bob's life is just as high in welfare value as Betty Ann's life is.
2) But it's not the case that Jim Bob's life is just as high in welfare value as Betty Ann's life is.
3) Therefore, it's not the case that DAIAH is true. [1,2 MT]

This problem for DAIAH arises because a) the quantity of pleasure in Jim Bob's life is the same as the quantity of pleasure in Betty Ann's life, and b) the object of Jim Bob's one episode of pleasure is exactly as pleasure-worthy as all the objects of Betty Ann's 100 episodes of pleasure. If this is the set-up, then DAIAH (supplemented by PWA and APR) would indeed imply that their lives are equal with respect to the amount of welfare they contain.

The second objection has to do with the impact of achievement on the welfare value of a life. To illustrate the objection, compare the lives of Jack and Jill. Neither Jack nor Jill feel any pain in their lives. For both of them, the only pleasure they receive is from observing great works of art. Their lives contain exactly the same amount of pleasure. For every episode of pleasure that Jack experiences, Jill experiences an episode of pleasure of the same intensity and duration. What's more, Jack and Jill are pleased by looking at the very same works of art. So the *objects* of all Jack's pleasures, respectively, are the same as the objects of all Jill's pleasures, respectively.[30] In fact, their lives differ in only one way: *Jill* is the one who has created the works of art that they both take their pleasure in. I stipulate that Jill takes no pleasure in the creation of the art, nor in the fact that it is her creation, nor anything else of that sort. Jack and Jill feel precisely the same amount of attitudinal pleasure, and the objects of their pleasures are exactly alike in their pleasure-worthiness. Thus DAIAH implies that their lives are equal with respect to the amounts of welfare value they contain. Intuitively, however, that is not a very plausible result. Since Jill both experiences pleasure *and* accomplishes something worthwhile in her life – namely creating beautiful works of art – while Jack only experiences pleasure, Jill seems to have a life that is somewhat better for her than Jack's life is for him. And so DAIAH seems to be false. So to put the objection succinctly:

The Achievement Argument:

1) If DAIAH is true, then Jack's life is just as high in welfare value as Jill's life is.
2) But it's not the case that Jack's life is just as high in welfare value as Jill's life is.
3) Therefore, it's not the case that DAIAH is true. [1,2 MT]

I think that prima facie these are two fairly compelling arguments against DAIAH. How are we to diagnose the problem with DAIAH that they reveal? It seems to me that these arguments show that it is problematic to formulate Intrinsic Attitudinal Hedonism in a way that takes the value of episodes of pleasure to be modified only by the pleasure-worthiness of the *objects* of these pleasures. I think we can avoid these two arguments if we move from talk of the pleasure-worthiness of *objects* of enjoyment to simply the *worthiness* of episodes of enjoyment.

The notion of the worthiness of episodes of pleasure involves, but is nonetheless broader than, the notion of the pleasure-worthiness of objects of pleasure (understood along the lines of PWA and ARP). The notion I am suggesting that Intrinsic Attitudinal Hedonists should appeal to is relativized to particular people and the amounts of pleasure they feel. More specifically, the notion is that of a state of affairs, S, deserving to a certain degree to have a certain amount of pleasure taken in it by a particular person. If a

[30] To be more precise, we might say that the proposition Jack takes pleasure in = the proposition Jill takes pleasure in = *this work of art is so beautiful!*

given state of affairs greatly deserves to have a certain amount of pleasure taken in it by a particular person, then that person's taking that amount of pleasure in that state of affairs would have a lot of *worthiness*. By contrast, if a given state of affairs does *not* deserve to have a certain amount of pleasure taken in it by a certain person, then that person's taking that amount of pleasure in it would have very little worthiness. We can capture the concept of worthiness as follows:

> W) The *worthiness* of a person P taking pleasure of intensity I for duration D in a state of affairs S equals X =df. it would be *appropriate* to degree X for P to take pleasure of intensity I for duration D in S.

How are we to understand appropriateness here? As before, it should be understood in terms of the approval of an ideal observer:

> APR*) It would be *appropriate* to degree X for a person P to take pleasure of intensity I and duration D in a state of affairs S iff an ideal observer would *approve* to degree X of P's taking pleasure of intensity I and duration D in S.

I propose that the most plausible version of DAIAH is to be formulated in terms of *worthiness* so understood.[31] On this understanding of DAIAH, the intrinsic welfare value of an episode of attitudinal pleasure would equal its intensity times its duration times its *worthiness*.

Why is this the best version of DAIAH? For one thing, understanding DAIAH in terms of worthiness would incorporate all the *benefits* of the pleasure-worthiness version of the theory (that is, the original version of the theory, which says the value of an episode of pleasure equals its intensity times its duration times the pleasure-worthiness of its object). After all, if some state of affairs, S, has a small degree of *pleasure-worthiness*, then there would not be much *worthiness* in anybody's taking any amount of pleasure in S. On the other hand, if a state of affairs, S', is highly pleasure-worthy, then it is likely to be the case that there would be a great deal of worthiness in people's taking pleasure in S'. Of course, if someone were to take pleasure in a pleasure-worthy state of affairs for an *insanely* long period of time, for instance, then this might decrease the worthiness of this pleasure episode of pleasure ever so slightly.

[31] This messes up the acronym 'DAIAH' a little bit, I admit. But the Worthiness-Adjusted Intrinsic Attitudinal Hedonism that I'm talking about here is so similar to regular DAIAH that I'm just going to keep using 'DAIAH.'

Second, the strategy of appealing to the worthiness of episodes of pleasure differs from the strategy of appealing to the pleasure-worthiness of objects of pleasure because the former avoids certain problems faced by the latter. More specifically, understanding DAIAH in terms of the worthiness of episodes of pleasure avoids the two objections discussed above. First, it avoids the Glory Days objection. Recall that this problem seemed to arise because DAIAH was supposed to imply that the lives of Jim Bob and Betty Ann are equally high in welfare value. But DAIAH understood in terms of worthiness does not have this implication. Granted the sheer amount of pleasure in Jim Bob's life equals that in Betty Ann's life. And granted the objects of Jim Bob's pleasures are just as pleasure-worthy as all the objects of Betty Ann's pleasures. (That's why old DAIAH implied that their lives were equally high in welfare value.) However, the *worthiness* of Jim Bob's insanely long episode of pleasure is not going to be quite as high as the worthiness of Betty Ann's episodes of pleasure. After all, an ideal observer would presumably approve slightly *less* of Jim Bob's being pleased for an insanely long amount of time by one inconsequential thing than such an observer would approve of Betty Ann's taking pleasure in equally good things for normal, sane amounts of time. Thus DAIAH formulated in terms of worthiness would not have the implication that Jim Bob's life is just as good as Betty Ann's life. Instead, DAIAH formulated in terms of worthiness will yield the intuitive result that Betty Ann's life is slightly better for her than Jim Bob's is for him.

Furthermore, DAIAH formulated in terms of worthiness will avoid the Achievement Argument as well. This problem for DAIAH seemed to arise because of the implication that Jill's life, which consists of her being pleased by looking at certain works of art (which she herself happened to have created), is just as good as Jack's life, which consists of his being pleased to the same degree by looking at those very same works of art (even though he has created nothing in his life). However, if W) and APR*) are plugged into DAIAH, as I propose that they be, then the theory will not have this problematic implication. After all, even though the intensities, durations and pleasure-worthiness of the objects of Jack and Jill's pleasures do not differ, the *worthiness* of their pleasures will differ somewhat. This is because an ideal observer would presumably approve slightly more of Jill taking pleasure in the works of art that *she* created than such an observer would approve of Jack's taking pleasure in those very same works of art given that he did *not* create them. Thus a version of DAIAH that is formulated in terms of the worthiness of enjoyments (instead of the pleasure-worthiness of objects of enjoyment) will not yield the counter-intuitive result that Jack's life is just as high in welfare value as Jill's life is. Instead, this version of DAIAH will have the more intuitive implication that Jill's life is somewhat higher in welfare value than Jack's is.

Because formulating DAIAH in terms of the worthiness of episodes of enjoyments instead of the pleasure-worthiness of objects of enjoyment avoids these two prima facie compelling objections, I think this is the version of DAIAH that its proponents should prefer.

4.4 The Problem with the Adjusted Enjoyment Theories

It seems to me that a version of DAIAH that employs the notion of worthiness is a very strong theory. It gets a lot right. I am inclined to think that it is the most plausible Objectively-Adjusted Enjoyment Theory currently on offer. However, it faces one serious problem that seems to be impossible for it to avoid. No kind or amount of adjustment can solve this problem. What's more, it is a problem that is shared by any monistic theory of welfare that qualifies as an Adjusted Enjoyment Theory, no matter whether Subjectively-Adjusted or Objectively-Adjusted. Because of this problem, I think no Adjusted Enjoyment Theory can capture the whole truth about welfare. No theory of this kind be entirely descriptively adequate.

The problem is that DAIAH, even when understood in terms of worthiness, will imply that a life that contains no pleasure is entirely worthless for the one who leads it. Feldman even acknowledges this point himself (at least with respect to IAH):

> No matter how much knowledge, virtue, honor, wealth, health, longevity, loving relationships, etc. he many have, if he takes pleasure in nothing, there is no basis for attributing positive intrinsic value to his life according to IAH. (…) a life without attitudinal pleasure or pain is a life without value. (Feldman, 2004, p. 67)

In fact, any Adjusted Enjoyment theory will have this implication. This is simply a result of the way that these theories do the math. On these theories, episodes of pleasure and pain are the only things that can bear value or disvalue. So even if a life contains a great deal of personal achievement or excellence or remarkable experiences, if it contains no episodes of pleasure or pain there are simply no value-bearing items in that life. No matter what the content of the life, the fact that there is no pleasure or pain in the life leaves no possible way for value to be generated according to the Adjusted Enjoyment Theories. Thus no matter what the content of a life might be, if it happens to contain no episodes of pleasure or pain, then these theories will imply that this life has no welfare value whatsoever. Such a life, in other words, would not be worth living for the one whose life it is.

I cannot bring myself to accept this implication. This seems entirely implausible to me. There are two reasons why. For one thing, certain possible lives seem outright desirable even though they contain no pleasure and no pain. Consider the life of a highly active and successful person who by some genetic fluke is incapable of experiencing enjoyment of any kind. This person has a neuro-physiology that makes it impossible for him to feel intrinsic attitudinal pleasure or displeasure, satisfaction or dissatisfaction, contentment or its opposite, happiness or unhappiness, sensory pleasure or sensory pain, and so on. His life is cold and robotic in many ways. However, the person, as I said, is highly active. He is not motivationally inert. Suppose he has some goals – suppose they are highly worthwhile goals – that guide his action. Moreover, because of his inability to feel pleasure or pain, he has no fear and no distractions, and this allows him to be singularly successful in pursuit of these goals. Such an existence, it seems to me, is by no means a worthless one. (In some moods, I am actually inclined to think that it is a rather good life.) What's more, it never seems to me that this robotic life of activity and success has a welfare value of *zero*. It seems to be one that contains a positive amount of welfare value (of course, not nearly as much as a life filled with a great deal of enjoyment of worthy things). But this intuition is inconsistent with all the Adjusted Enjoyment Theories.[32]

[32] A related problem is that Adjusted Enjoyment Theories like DAIAH have the implausible consequence that all of the lives that contain no pleasure and no pain contain *exactly* the same amount of welfare value. This consequence does not seem plausible either. Of the lives containing no pleasure and no pain, some *clearly* seem to be better for a person than others. For instance, a life of no pleasure and no pain that is spent exclusively watching paint dry would seem to be much worse for the one whose life it is than a singularly successful life that contains a range of remarkable experiences, but that also happens to contain no pleasure and no pain. In general, it is implausible to think that one's welfare level in a life containing no

The second reason it seems to me that a life containing no enjoyment might nonetheless contain a positive amount welfare is this. If a life contains no welfare value, then for the one whose life it is it would not be a life worth living. So we would expect people to be indifferent between leading a life that has zero welfare value and not being alive at all, at least all other things being equal – that is, disregarding other non-selfish reasons one might want to remain alive (like wanting to help one's children survive). However, I suspect that the vast majority of people would *not* be indifferent between a life containing no pleasure and no pain, on the one hand, and not being alive at all, on the other. Even disregarding the non-selfish reasons for which one might want to stay alive (like wanting to help one's children succeed or wanting to save the planet), I suspect that most people would prefer to lead a life that is guaranteed to contain no pleasure or pain over not being alive at all. It seems to me that being conscious and aware of *something*, even if it is not of things that are in any way pleasing or displeasing, seems better than being dead. But this is not consistent with the implications of the Adjusted Enjoyment Theories.

Perhaps one could object to this second line of argument by pointing out that the only reason people would prefer, all else being equal, a life with no pleasure or pain over no life at all is that people have an irrational fear of death. After all, it is quite reasonable to think that there are powerful evolutionary forces that select for being disposed to be afraid of dying. While I agree that a fear of death has probably been selected for, I do not think this undermines my argument. After all, why think that the widespread fear of death is in any way *irrational*? I see no independent, non-theory driven reason to think that it is. I have been suggesting that, all else being equal, people would prefer a life with no pleasure and pain over no life at all, and that this suggests that a life with no pleasure and pain does *not* have zero welfare value. This line of argument would be undermined only by the claim that this preference is the result of an *irrational* fear of death. If it's not the case that the fear of death is irrational, then the ceteris paribus preference for being alive

enjoyment but lots of achievement *is exactly the same* as one's welfare level in a life containing no enjoyment and no achievement either.

without pleasure or pain over being dead would indeed seem to support the idea that there is a greater than zero amount of welfare in a life with no pleasure or pain.[33]

Since all the Adjusted Enjoyment Theories, even very sophisticated ones like DAIAH, imply that the welfare value of a life that contains no pleasure or pain is zero, I do not think that any of these theories capture the whole truth about welfare. Something is missing. There must be some additional component to welfare besides enjoyment. That is the conclusion of this chapter. In later chapters, I aim to defend a theory of welfare that says what this additional component is.

[33] That this is so is supported by the argument of the first three sections of chapter 1.

CHAPTER 5

DESIRE SATISFACTIONISM AND TIME

The notions of desire and time are connected in several ways. For one thing, every desire is had at a particular time, or during a particular interval. For instance, it is today that I have my desire to go skiing, and it is tomorrow that I have my desire not to be sick with a cold. Since desires are had at times, it is clear that our desires can change over time as well. Thus at one point in time I might have a strong desire to be a professional cellist, but later in life I might lose this desire and instead acquire the desire to be a professional philosopher.

A second way in which desire and time are connected is that the *objects* of our desires might have particular times built into them. For instance, some desires are for certain things to happen *now* – e.g. that I go skiing today. Some desires are for certain things to happen in the *future* – e.g. that I go skiing in a week. And some desires are for things to have happened in the *past* – e.g. I went skiing during the winter holiday (which was a month ago). Of course, sometimes we desire that certain things are the case, but not at any particular time. For instance, I might desire that there is an elegant way to prove Fermat's Last Theorem. Thus the basic form of a desire seems to be roughly this:

(D) *S* desires, at *t*, with intensity, *i*, that *p* is the case, where '*p'* here could stand for a state of affairs obtaining at a particular time (or during an interval), or else '*p*' could stand for a state of affairs obtaining but not at any time in particular.[1]

Both of these ways in which desires are connected to time raise difficult philosophical issues. For instance, the way in which the objects of desires may have times built in raises tricky questions for Desire Satisfaction accounts of well-being (or welfare). According to these views, what enhances one's well-being is that one desires that something be the case and that it in fact is the case. Philosophers who are sympathetic to such a view must decide whether to take it that a person's well-being is enhanced also by the satisfaction of desires for some state of affairs to obtain in the future or in the past, and not just by the satisfaction of desires for some state of affairs to obtain now. For instance, suppose I desire, today, that I have children when I am 40 years old and in fact I will have children when I am 40. Is my welfare thereby enhanced? Or suppose I desire, today, to have gone to college between the ages of 18 and 22, and it is in fact the case that I did so. Does this fact, too, enhance my welfare?

The fact that we might hold different desires at different times raises other difficult issues. Much ink has been spilt discussing the question of how to decide what to do at a particular time given that our desires change over time. When deciding what to do now, should we consider only the preferences we have at present? Or should we also consider the preferences we will have in the future or have had in the past, which clearly might conflict with our present preferences? To illustrate the problem, consider this example of Parfit's:

> When I was young what I most wanted was to be a poet. (…) Now that I am older, I have lost this desire. (…) Does my past desire give me a reason to try to write poems now, though I now have no desire to do so? (Parfit, 1984, p. 157)[2]

This is a pressing question for anyone interested in giving an account of how one ought, from e.g. the perspective of rationality or prudence, to act at a given time. The responses that have been offered vary widely. R.M. Hare, in *Moral Thinking*, takes it that the

[1] Note that I am assuming that we may keep our desires even after we get them satisfied. This is because I am understanding the notion of desire in a quite wide way. I will take it that to desire that something be the case is roughly the same as to attach value to this thing's being the case, or to have a preference to some degree or other that this thing is the case.

[2] Many philosophers have offered cases with the very same structure. Cf. Brandt, 1979, p. 249; Heathwood, SDS, ms, p. 9-11; and Bykvist, 2003, p. 17-18.

rational thing to do at any given moment is determined only by the desires one has at that moment.[3] By contrast, Phil Bricker argues that all of one's preferences count: past, present and future.[4] Other philosophers offer different views.[5]

As Chris Heathwood points out,[6] this phenomenon of desire change also poses a problem for simple formulations of Desire Satisfactionism about well-being. Consider again Parfit's example. If the Parfit of today were to write some poems, this would satisfy the desire he had as a young man. Since Desire Satisfactionism (or a simple version of it, at least) takes it that what enhances one's well-being is that one desires that something be the case and that it in fact is the case, the satisfaction of the desire Parfit held as a young man would increase the total amount of well-being contained in Parfit's life. But some might find this to be an odd consequence, given that the Parfit of today – when the poems would be written – has no desire whatsoever to be writing poetry. If you share this intuition, you may have a reason to doubt Desire Satisfactionism.

In this chapter, I will be concerned with the temporal problems that arise specifically for Desire Satisfactionism. In particular, my question will be this: how should the Actual Desire Satisfactionist (i.e. someone who thinks that welfare depends on the desires one *actually* has, as opposed to the ones one *would* have under certain ideal conditions) formulate his view so as to deal with the problems raised by the temporal nature of desires? I will not directly address the question of what one ought, e.g. from the perspective of rationality or prudence, to do at a particular time (which, it should be noted, has been the focus of most of the literature on desires and time). However, there is clearly some overlap here. For instance, several of the versions of Desire Satisfactionism I discuss here are analogous to views about how to decide what to do at a time.

The order of business will be as follows. To start, I will introduce some concepts to clarify our discussion. In particular, I will offer a conceptual scheme that organizes the different ways in which the objects of desires may (or may not) have times built in, and

[3] Hare, 1981, pp. 104-105. See Rabinowicz, 1989 for a criticism of Hare's view.

[4] Cf. Bricker, 1980. In particular, he thinks that the prudentially right thing to do is to bring about – out of all the possible worlds that one can bring about at the time in question – the possible world that maximizes the degree to which all of the preferences one holds at one time or another in that possible world get satisfied. Also see Bykvist, 2006.

[5] See for example, McKerlie 2007b. What's more, Parfit gives a lengthy and insightful discussion of a number of different responses to this problem in *Reasons and Persons*(Cf. Parfit, 1984).

[6] Heathwood, SDS, ms, p. 9-11

the different ways in which desires may (or may not) change over time. Then I will consider the different ways in which a Desire Satisfactionist may formulate her view in order to deal with the temporal nature of desire. In particular, I will consider half a dozen different versions of Desire Satisfactionism that have either been explicitly defended by various philosophers or suggested by comments on related topics. I will argue that the best way out of the problems raised by desire and time is provided by a version of Desire Satisfactionism that I will call *Weak Concurrentism*. My defense of this view will consist of two components: I will first attempt to show that the other formulations of Desire Satisfactionism on offer face insurmountable problems, and then I will go on to argue that the prima facie implausible consequences of Weak Concurrentism are, on closer inspection, in fact not implausible at all. Thus my conclusion is going to be a conditional one: insofar as one is committed to Actual Desire Satisfactionism, Weak Concurrentism is the version of the view one should accept.

5.1 The Temporality of Desire

As noted, there are two main ways in which desire is connected to time: we have desires at times, and the objects of our desires may have times built in. It will help clarify the discussion to follow if I introduce some concepts to keep all this straight.

5.1.1 Time and the objects of desire

I propose to think of desires in such a way that their *objects* typically have two components: i) a desired state of affairs, *p*, and ii) an interval of time, *t*, at (or during) which it is desired that *p* obtain.[7] For simplicity, I am just going to assume that the state of affairs referred to by '*p*' here is not itself relativized to any time. Thus I assume that 'p' refers to some state of affairs like 'I go skiing' or 'I am typing', not a time-relativized

[7] Another approach might be to take it that the objects of desires have times *built in*. That is, one might think that the object of a desire consists of just one component: viz. a time-relativized state of affairs. However, I opted for the other approach just because I find it simpler (and because I have some questions about the nature of time-relativized states of affairs in general). But I am not sure that much of substance hangs on my picking this approach rather than the other one.

state of affairs like ‘I go skiing today’ or ‘I am typing at noon on Feb. 3rd 2009’. This assumption will simplify matters greatly.[8]

But it is clear that not all desires are temporal desires of this sort. There also seem to be desires whose objects contain a state of affairs, but no time at which it is desired that this state of affairs obtains. For instance, I might desire that Fermat’s Last Theorem have an elegant proof, but also not have any time, *t*, in mind such that *t* is the specific time I want this state of affairs to obtain at. The object of a desire like this seems to consist of a desired state of affairs and a time-slot that is left empty. I am going to call desires of this sort *atemporal desires*.

Nonetheless, many – perhaps even most – of our desires are not atemporal in this way. Often we want something to happen at a particular time, or during a particular interval of time. For example, this is the case with desires like my desire to go skiing tomorrow, my desire to own a beach house when I’m 40 and my desire to have graduated from college. The objects of these desires contain a state of affairs *and* a time at which it is desired that the state of affairs in question obtain. I am going to call desires of this sort *temporalized desires*.

What, then, are the main kinds of temporalized desire? Hare proposes a helpful way of categorizing desires like these:

> Suppose that I now prefer that at some later time (then, for short), *x* should happen, but that I shall then prefer that *x* not happen. (…) It will simplify the example if we suppose that the second preference (the preference then), is what I shall call a then-for-then preference. By this I mean that it is a preference for what should happen then. (…) The first preference [by contrast], is a now-for-then preference… (Hare, 1981, p. 101-102)

In addition to the then-for-then desires and the now-for-then desires that Hare mentions here, he also goes on to mention now-for-now desires, which are desires one has now for some state of affairs to obtain now. Thus there are three main types of desires, on Hare’s proposal: now-for-now, then-for-then and now-for-then. To make the picture complete, we can assume that the now-for-then desires include not only desires one has now for some state of affairs to obtain in the *future*, but also desires one has now for some state of affairs to obtain in the *past*. (We might also want to add then-for-now desires.)

[8] It means we don’t have to deal with, say, desires like this one: I desire, now, that during all of 2009 I have a desire to get a lot of work done in 2009. Presumably more work could be done to accommodate desires like this that are relativized to times in three (or more) places. But I’m not going to do it here.

Hare's terminology seems to pick out some of the main ways in which the objects of desires can be associated with times. But his terminology is unstable because it allows that the same desire can belong to different types depending on when 'now' is taken to be. For instance, suppose I desire at noon on Feb. 3rd, 2009 that I go skiing on Feb. 16th 2009. If 'now' is Feb. 1st, 2009, this will be a then-for-then desire. But if 'now' is Feb. 3rd 2009, this very same desire will be a now-for-then desire.

However, I think the awkwardness of Hare's terminology can be avoided. Hare is pointing out three important kinds of temporalized desires, which I will call *present directed desires*, *future directed desires* and *past directed desires*. The present directed desires are the ones where the object of your desire is *present* relative to the time at which you hold the desire. Thus we get the following definition:

> For any two times (intervals), t_1 and t_2, if S desires, at t_1, that p obtains at t_2, and t_1 and t_2 are simultaneous, then S's desire is **present directed**.

A few clarifications. First, while it's clear what it is meant by desiring something at a moment, what is meant by desiring something at an *interval*? If t is an interval of time, then 'S desires *at t* that p obtains' simply means that for every moment in t, S has a desire that p obtain. Second, what does it mean to say that two intervals are simultaneous? I will take it that that two intervals of time, t_1 and t_2, are *simultaneous* when it's the case that for any moment of time, it occurs during t_1 iff it also occurs during t_2. Third and finally, in this definition (as with the next two), I am going to assume for simplicity that the state of affairs referred to by 'p' here is a non-temporalized state of affairs, like 'I go skiing' or 'I am typing' – not 'I go skiing today' or 'I am typing at noon on Feb. 3rd 2009'. So for example, what would it be for me to have a present directed desire to go skiing during a certain interval, for example? Well, there would have to be some interval of time (e.g. 9am Feb. 3rd 2009 to noon Feb 3rd 2009) such that I have a certain desire during that whole interval, and the object of this desire is for me to be skiing during that whole interval.

Next, the future directed desires may be defined as follows:

> For any two times (intervals), t_1 and t_2, if S desires, at t_1, that p obtains at t_2, and t_2 is later than t_1, then S's desire is **future directed**.

The future directed desires are so-named because with them, the object of your desire is in the *future* relative to the time at which you hold the desire. Again, I want to be clear that '*p*' here refers to some non-temporalized state of affairs like 'I am typing'.

Finally, the past directed desires may be defined like this:

> For any two times (or intervals), t_1 and t_2, if S desires, at t_1, that *p* obtains at t_2, and t_2 is prior to t_1, then S's desire is **past directed**.

The past directed desires are so-named because with them, the object of your desire is in the *past* relative to the time at which you hold the desire. As before, '*p*' here refers to some non-temporalized state of affairs.

This terminology is more stable than Hare's. After all, a past directed desire, say, will always be past directed, no matter where in time you or I happen to find ourselves and no matter when we take 'now' to be. Similarly for the present directed and the future directed desires. Once a present-directed desire, always a present-directed desire; once a future-directed desire, always a future-directed desire.

Still, one might think that there is something strange about my proposed definitions. Consider what the definition of present-directed desires implies about the following case. Suppose that at noon on Feb 3, 2009 I briefly wake up from a heavy sleep, form a short-lived desire that I be skiing all day on Feb 3, and then fall back asleep. Would I then be having a desire that is *not* present-directed, since I don't possess this desire all day? The definition of present directed desires seems to imply this, but it might be an implausible result.

However, I am not worried about cases like this, because my definitions provide the resources to plausibly account for such cases. After all, if I wake up at noon, then presumably what really is happening is that I have two desires: one past-directed desire for me to have gone skiing during the morning hours (instead of being asleep the whole time), and one future-directed desire for me to go skiing in the afternoon. This seems to be a plausible way to understand what I would mean when, in the scenario described, I wake up and think 'I want to go skiing all day today'. On the other hand, you might think this is not what's going on in this case. Suppose the case really is such that what I desire when I wake up is *literally* to go skiing all day. In other words, I desire that every moment during the day today be such that I am skiing at that moment. It seems unlikely that a real person would ever have such a desire, but if this really is what is going on in the case at hand, then my desire would indeed not be a present-directed one. The definition admittedly does have this implication. But if what I desire is *literally* to be skiing all day long, then in virtue of this fact I will also have a derivative desire to be skiing *right now*, i.e. at the moment when I wake up. (After all, the moment when I wake up is part of the interval that is picked out by 'all day'.) And this derivative desire of mine really would be a present-directed desire. So if you have the intuition that I've got some kind of present-directed desire in this case, then my terminology can account for this intuition too. This example will illustrate, I hope, how my proposed definitions can capture a wide variety of temporalized desires. One may have to be careful about exactly how one individuates the desires in any given case. But the definitions can account for most of the standard cases, I think.

Nonetheless, it seems clear that the tri-partite distinction between present-directed, past-directed and future directed desires cannot capture *all* the temporalized desires that a person might have. For people often seem to have desires with temporalized objects, but where it's not entirely clear what the relevant time is. For instance consider the following desires:

> D1) I desire, now, to own a beach house by the time I am 40 years old.
> D2) I desire, now, to live to be at least 90 years old.
> D3) I desire, now, that I got a good college education at some appropriate time in the past.

These all seem to be desires whose objects have times built in, but where it's not clear what the relevant times are. We might call these *diffusely temporalized desires*. I think there is a natural way to unpack specifically *these* diffusely temporalized desires so that it is easy to see what the conditions are under which these desires would be satisfied. D1), it seems, can be re-interpreted as something like this:

> D1*) I desire, now (noon, Feb. 3rd 2009), that there is an interval of time occurring between now and my 40th birthday (Nov. 30th 2022) during (all of) which I own a beach house.

Similarly, D2) seems to amount to something like this:

> D2*) I desire, now (noon, Feb. 3rd 2009), that there is some interval of time occurring on or after my 90th birthday (Nov. 30th, 2072) during which I am alive.

And D3) can be re-interpreted as follows:

> D3*) I desire, now (noon, Feb. 3rd 2009), that there is some interval of time, occurring before now but not too recently (e.g. not after I turned 23), during which I received a good college education.

Given this way of re-interpreting D1) – D3), it should be easy to see when their objects would obtain. In particular, since D1) – D3) amount to desires for there to be an interval of time that possesses a certain feature, these desires would be satisfied as long as there exists some interval of time with the specified feature.

However, it should be noted that re-interpreting D1) – D3) in this way does not make them conform to the canonical form of a past or future directed desire. After all, the objects of past or future directed desires were said to consist of two things: i) some non-temporal state of affairs and ii) a time at which it is desired that this state of affairs obtains. But D1*) – D3*) are not like this. They are desires for there to be some interval of time (whether past or future) that possesses a certain feature. So these desires do not fit the canonical form of a past or future directed desire. (Nonetheless, it might be natural to think that D1*) and D2*) represent another species of future directed desire than the canonical one, and that D3*) represents another species of past directed desire than the canonical one.)

In what follows, however, I will be focusing primarily on the canonical temporalized desires, viz. the present directed, past directed and future directed desires. These are sufficient for illustrating the sort of problems that the temporal nature of desires raise for desire satisfactionist theories.

5.1.2 The evolution of desire over time

When it comes to past, present and future directed desires, there seem to be four particularly important ways in which such desires develop over time. We might have future directed desires that either match our later desires or that don't, just as we might have past directed desires that match our previous desires or that don't. Begin with the future-directed desires. First, we have cases in which a future directed desire is *matched* by a later present-directed desire. For instance, consider the following pair of desires:

> D4) I desire, on Feb. 3rd 2009, that I go skiing on Feb. 16th 2009.
> D5) I desire, on Feb. 16th 2009, that I go skiing on Feb. 16th 2009.

If I hold both of these desires, then we have a case in which my future directed desire is matched by a subsequent present directed desire.[9] Next, we have cases of *unmatched* future directed desires. Suppose, for instance, that I hold the desire in D4), but on Feb. 16th 2009, I have no desires pertaining to skiing at all – not because, say, I am asleep or in a coma, but because my desire to ski has entirely evaporated. In such a case, my future directed desire D4) would not be matched. Similarly, suppose I hold the desire in D4), but then I come to hold:

D6) I desire, on Feb. 16th 2009, that I do not go skiing on Feb. 16th 2009.

In this case my future directed desire, D4), is not matched either. These seem to be the two main types of cases concerning future directed desires: those where the future directed desire is *matched* and those where it is *unmatched*.[10]

We find a similar pair of cases when it comes to the past directed desires. First, we have cases in which a past directed desire is *matched* by an earlier present-directed desire. We would have such a case if I held, for instance:

D7) I desire, on Feb. 3rd 2009, that I go to college during the interval 2001-2005.
D8) I desire, during the interval 2001-2005, that I go to college during the interval 2001-2005.

By contrast, we would have a case of an *unmatched* past-directed desire if I held the past directed desire in D7) together with this present directed desire:

D9) I desire, during the interval 2001-2005, that I *not* be in college during the interval 2001-2005.

Similarly, my past directed desire D7) would be unmatched if during the interval of 2001-2005, I had *no* desires pertaining to going to college at all – not because I am asleep or in a coma, but rather because, say, my upbringing prevented me from having any knowledge of college at all at that time. These seem to be the two main types of cases

[9] To be more precise, what I mean by a matched future desire is this: i) during the whole interval, t_1, I have a future directed desire that p obtain during the whole interval t_2 (which is later than t_1), and ii) during the whole interval t_2, I have a present directed desire that p obtain during the whole interval t_2.

[10] For a more precise account of what makes a future directed desire matched or unmatched, see Bykvist, 2003. The matched desires correspond to Bykvist's notion of a desire *with* full inside support, and the unmatched desires correspond to Bykvist's notion of a desire *without* full inside support.

concerning past directed desires: those where the past directed desire is *matched* and those where it is *unmatched*.[11]

5.2 Simple Desire Satisfactionism

The question I will be concerned with in the remainder of this paper is how the Actual Desire Satisfactionist should deal with the temporal nature of desires. In general, the Actual Desire Satisfactionist about well-being takes it that one's well-being depends on the desires one *actually* has, as opposed to the ones one *would* have under certain ideal conditions.[12] The most basic version of Actual Desire Satisfactionism is what Heathwood calls *Simple Desire Satisfactionism*, which is roughly the view that what is good for one is getting what one wants. This theory is almost certainly false, and one of its deepest flaws is its failure to account for the temporal nature of desires. It will be instructive to begin by explaining these flaws.

Heathwood formulates Simple Desire Satisfactionism, or SDS, as follows:

> (i) Every desire satisfaction is intrinisically good for its subject; every desire frustration is intrinisically bad for its subject.
> (ii) The intrinisic value for [the] subject of a desire satisfaction or frustration is a function of the intensity and the duration of the desire satisfied or frustrated.
> (iii) The intrinsic value of a life for the one who lives it = the sum of the intrinsic values of all the desire satisfactions and frustrations in the life.[13]

So as not to confuse SDS with other versions of Actual Desire Satisfactionism to be discussed below, it should be noted that SDS employs a particularly simple notion of desire satisfactions and desire frustrations. What has intrinsic value for a person on SDS are *simple desire satisfactions*, which are states consisting of one's actually desiring that some state of affairs, *p*, obtains and *p*'s in fact obtaining. By contrast, what has intrinsic disvalue for a person on SDS are *simple desire frustrations*, which are states consisting of one's actually desiring that *p* obtains and *p*'s in fact failing to obtain. (Notice that the

[11] Again, for a more precise account of this distinction, see Bykvist, 2003.

[12] The contrasting view, Ideal Desire Satisfactionism, has many defenders. (For example, see Brandt,1979; Griffin, 1986; Railton, 2003.) However, this view faces the same problems because of the temporal nature of desires as Actual Desire Satisfactionism does. To keep the scope of this chapter manageable, I focus only on Actual Desire Satisfactionism. But presumably, the same sorts of solutions I discuss here are available to the Ideal Desire Satisfactionist as well.

[13] Heathwood, SDS, ms, pp. 4-5

definitions of simple desire satisfaction and simple desire frustration do not include any mention of times. This is the source of the problems for SDS.)

Desire Satisfactionists should not accept SDS as the way to formulate their view. Some think that this is because SDS has strange implications about certain cases of unmatched future directed desires (though in the last section I'm going to argue that they might not be that strange after all). Moreover, there is another, simpler reason to reject SDS. But begin with the alleged problem concerning unmatched future directed desires.

Heathwood raises the following case in order to show a problem specifically with SDS:

> *Ellie's 50th Birthday Party* Teenage Ellie, a rock 'n' roll fan, is imagining her 50th birthday party. She wants live rock 'n' roll at the party. She continues to desire for years and years that there be rock 'n' roll at her 50th birthday party. But a month before the party, Ellie ceases enjoying rock 'n' roll. She now prefers easy listening, and finds rock 'n' roll loud, childish, and annoying. She will continue to feel this way on her 50th birthday. She would have the time of her life at her party if she got easy listening, but would be miserable if rock 'n' roll were played. (Heathwood, SDS, ms, p. 10)[14]

Brandt offers another case of this sort:

> a convinced sceptic who has rebelled against a religious background wants, most of his life, no priest to be called when he is about to die. But he weakens on his deathbed, and asks for a priest. Do we maximize his welfare by summoning a priest? Some would say not in light of his past desires. (Brandt, 1979, p. 250)

Hare offers one, too:

> I wanted, when a small boy, to be an engine-driver when I grew up; when I have graduated as a classical scholar at the age of 18, and am going to take the Ph.D. in Greek literature, somebody unexpectedly offers me a job as an engine-driver. In deciding whether to accept it, ought I to give any weight to my long-abandoned boyhood ambition? (Hare, 1981, p. 159)

And Parfit's poet case, mentioned in the introduction of this paper, is basically the same as Hare's.[15] [16] What all these cases have in common is that the person in the story begins with a future directed desire for some state of affairs (a 50th birthday party with rock 'n' roll, no priest present at one's deathbed, being an engine-driver or a poet as an adult), but when the time in question comes around, this desire is not matched by a corresponding present directed desire. What does SDS imply about these cases?

[14] This case is a version of a case offered by Brandt: 'Suppose my six-year-old son has decided he would like to celebrate his fiftieth birthday by taking a roller-coaster ride. This desire now is hardly one we think we need attend too in planning to maximize his lifetime well-being. Notice that we pay no attention to our own past desires. Are we then to take into account only the desires we think my son will have at the time his desire would be 'satisfied', here at the age of fifty?' (Cf. Brandt, 1979, p. 249)

[15] Cf. Parfit 1984, p. 157

[16] Griffin also mentions a case of this sort. Cf. Griffin 1986, p. 16.

For simplicity, let's focus just on the first two cases. (The point carries over, albeit in a more complicated way, to the cases of Hare and Parfit). Let's suppose that Ellie gets a 50th birthday party with rock 'n' roll, and that no priest visits the dying atheist on his deathbed. Thus Ellie and the atheist get their respective future directed desires satisfied, but their present directed desires frustrated. What's more, let's suppose that both in the case of Ellie and in the case of the atheist, the future directed desire in question (i.e. Ellie's for rock 'n' roll, and the atheist's for no priest) has the same average intensity as the present-directed desire in question (i.e. Ellie's for easy listening, and the atheist's for a priest). On these suppositions, then, what SDS implies is that Ellie and the atheist are made better off, on balance, by what actually happens to them. For the satisfaction of the future directed desire in these cases outweighs the frustration of the present directed desire. Thus SDS implies that in these cases, assuming we are interested in maximizing the well-being of the protagonist of the stories, we should try to satisfy the original future directed desire, not the new present directed desire. This is what Bykvist calls a 'present-for-past sacrifice.' (Bykvist, 2007, pp. 74-75)

But some find these implications to be counter-intuitive. Bykvist, for instance, says he 'want[s] to avoid the present-for-past sacrifices illustrated by Parfit's poet example.' (Bykvist, 2007, p. 74) What's more, Heathwood argues explicitly that SDS is refuted by the case of Ellie.[17] After all, it might seem that, intuitively, Ellie would not be made better off on balance by getting rock 'n' roll played at her party. Since she no longer desires rock 'n' roll, she would in no way be benefited by having it played at her party. Instead, what Ellie would be benefited most by is a birthday party with music that conforms to her current preferences. Similarly, for the atheist case. If you were a life-long friend of this person and wanted what is best for him, wouldn't you call a priest to his bedside? Isn't this the thing a caring fried would do here? If you think so, then you'll find the implications of SDS about these cases to be counter-intuitive.[18]

However, these are complex cases and I will discuss them at length in the last section. I will argue that present-for-past sacrifices are in fact not as problematic as they might at

[17] Cf. Heathwood, SDS, ms, pp. 10-11

[18] Heathwood thinks the theory goes wrong in allowing the satisfaction of desires held in the past to count towards one's welfare. As he puts it, 'That Simple Desire Satisfactionism gives equal weight to merely past desires makes it, in some cases, as paternalistic as any objective list theory.' (Heathwood, SDS, ms, p. 11) Bykvist thinks we should 'respect the autonomy of person-stages.' (Cf. Bykvist, 2007)

first seem. Thus in defending my own view about how the Desire Satisfactionist should deal with the temporal nature of desire, I will argue against the judgments that Heathwood, Bykvist and others are inclined to make about these cases. While some might think that these cases provide sufficient reason to reject SDS, I am not convinced.

Nonetheless, I think SDS is defective for another, simpler reason. In particular, the notions of simple desire satisfaction and simple desire frustration are flawed because they are entirely insensitive to time. Start with desire frustration. Because SDS appeals to the notion of simple desire frustration, it allows that if one has a desire, at some time, for a given state of affairs, *p*, then one will be harmed as long as there is *some time* at which *p* fails to obtain. Suppose that right now I have a desire (an atemporal one) for my dissertation to be finished sometime, but not at any particular time. The state of affairs I now desire did not obtain when I was 15 years old. So one might think that SDS implies that I am thereby harmed. For here we seem to have an episode of simple desire frustration. There is a state consisting of i) my desiring a certain state of affairs – viz. my being done with my dissertation – and ii) this state of affairs not obtaining while I am 15. However, it is absurd to say that I am harmed by this state. It is entirely natural that one's dissertation not be finished almost a decade before one starts working on it. There is no harm in that.

The same problem comes up concerning simple desire satisfactions. SDS implies that one will receive some benefit as long as one has a desire, at a given time, for some state of affairs and there is any time at all that this state of affairs obtains. This is not supposed to be the same problem as the cases of past-for-present sacrifice discussed a moment ago. Suppose I have a desire right now to write a very influential book on moral philosophy – not at any time in particular, but just whenever. Suppose that in 20 years, it just so happens that I write such a book. SDS might seem to imply that my welfare is thereby enhanced right now. For here we seem to have an episode of simple desire satisfaction. There is, right now, a state consisting of i) my desiring a certain state of affairs, and ii) this state of affairs obtaining 20 years in the future. However, it would be absurd to say that I am benefited right now by this state.

Thus no matter what you think of the cases mentioned above – offered by Heathwood, Brandt, Hare and Parfit – SDS clearly has got to go. What we need is a formulation of desire satisfactionism that takes times into account. But SDS does not. In the next few sections, then, I will consider a range of different ways of taking time into consideration.

5.3 Concurrent Desire Satisfactionism

One plausible way for the Desire Satisfactionist to take times into account is to say that what is good for one is *to get what one wants want while one wants it*, and what is bad for one is to *fail* to get what one wants while one wants it. This is the basic idea underlying what I'll call *Concurrent Desire Satisfactionism*. However, there are several ways in which this basic idea might be developed. For instance, according to what I'll call *Weak Concurrentism*, it is *literally* the case that what is good for you is to get something you want while you want it – even if you want that this thing to happen at some time other than the present. By contrast, what I'll call *Strong Concurrentism* is the more restrictive view that the only thing that is good for you is to have a desire now for something to happen now, while this thing does in fact happen now. In the last section of this chapter, I will defend Weak Concurrentism – despite the fact that it does not have the consequences about cases like Ellie's birthday and the dying atheist case that Heathwood and others seem to think are the intuitive ones. Strong Concurrentism, by contrast, does have the supposedly 'intuitive' consequences about these cases. But, as we'll see, it is has other fatal flaws.

5.3.1 Weak Concurrentism

The motto of Concurrent Desire Satisfactionism is that what is good for you is to get something you want while you want it. Weak Concurrentism takes this motto literally. On this view, one's welfare is enhanced if one desires, at a time, that some state of affairs obtain, and this state of affairs does obtain at that time. Similarly, one's welfare is diminished if one desires, at a time, that some sate of affairs obtain, and this state of affairs does obtain at that time.

To state the theory in full, let me introduce the notions of concurrent desire satisfaction and concurrent desire frustration:

> S gets an episode of **concurrent desire satisfaction** during the whole interval <t_1, t_2> iff for every time, *t'*, that occurs during <t_1, t_2>, it is true both that i) S has, at *t'*, a desire that some state of affairs, *p*, obtain, and ii) *p* in fact obtains at *t*'.
>
> S gets an episode of **concurrent desire frustration** during the whole interval <t_1, t_2> iff for every time, *t'*, that occurs during <t_1, t_2>, it is true both that i) S has, at *t'*, a desire that some state of affairs, *p*, obtain, and ii) *p* does *not* obtain at *t*'.

Notice that this lets us get around the simple problem that refuted SDS. If, for instance, I now have a desire for the non-temporal state of affairs [my dissertation is finished], we cannot say that my desire is concurrently frustrated just because my dissertation was not finished back when I was 15. In order for this desire to be concurrently frustrated, it must be the case that the state of affairs I desire now fails to obtain now.

Given the notions of concurrent desire satisfaction and frustration, we can state the view as follows:

Weak Concurrentism

(i) Every concurrent desire satisfaction is intrinisically good for its subject; every concurrent desire frustration is intrinisically bad for its subject.

(ii) The intrinsic value for S of an episode of concurrent desire satisfaction equals the duration of the episode times the average intensity of the relevant satisfied desire; the intrinsic disvalue for S of an episode of concurrent desire frustration equals the duration of the episode times the average intensity of the relevant frustrated desire.

(iii) The total amount of welfare contained in a person's life equals the sum of the intrinsic values of all the episodes of concurrent desire satisfaction in the life minus the sum of the intrinsic disvalues of all the episodes of concurrent desire frustration contained in the life.[19] [20]

I think Weak Concurrentism really might be the best way for the Desire Satisfactionist to deal with the temporality of desire.[21] This is what I will argue in the last

[19] It seems to me that this view is roughly equivalent to the view that is defended in Bricker, 1980. Bricker formulates his view as answer to the question of how one prudentially ought to act at a given time, as opposed to the question of how the total welfare contained in a life is to be determined. Still, I take the views to amount to the same roughly thing. For on Bricker's view, the satisfaction or frustration of any desire one holds – whether present directed, past directed, future directed, or entirely atemporal – will have an impact one's welfare. (This is also similar to a view that Dennis McKerlie mentions, but rejects. Cf. McKerlie, 2007b, pp. 53-57)

[20] Note that this view, as formulated here, faces serious double-counting problems. However, in chapter 6, I argue that these problems can be solved by formulating Desire Satisfactionism in terms of the notion of a *basic desire*. See chapter 6 for a detailed discussion of how this solution is supposed to work.

[21] At the beginning of this chapter, I said that, in its canonical form, desire is a 4-place relation: 'S desires, at *t1*, that *p* obtains at *t2*.' However, Weak Concurrentism is stated in terms of the notions of concurrent

section. But for now, I want to point out that the view has some consequences that some might find troublesome. These consequences, for some people, motivate the search for an alternative approach to the temporality of desire.

In particular, Weak Concurrentism has just the sort of implication about cases of unmatched future directed desires that Heathwood, Bykvist and others take to be counter-intuitive. To take just one example, consider the case of Ellie (though similar points could be made using the other cases mentioned above). For many years of her life, Ellie has a future directed desire that rock ‘n’ roll be played at her 50th birthday party. On the day itself, though, she has lost this desire entirely. Now she has a present directed desire that easy listening – not rock ‘n’ roll – be played at her 50th birthday party. Weak Concurrentism implies that the total amount of welfare contained in Ellie’s life would be greater if rock ‘n’ roll is played at her party than if easy listening is played. Thus Weak Concurrentism implies that, if we’re interested maximizing the amount of welfare contained in Ellie’s life, we should perform a ‘present-for-past sacrifice’ and play rock ‘n’ roll at the party.[22] Those who find this implication counter-intuitive (e.g. Heathwood, Bykvist) thus might want to reject Weak Concurrentism.

desire satisfaction and frustration, which make use of a 3-place notion of desire. One might wonder why the 4-place notion of desire was not used instead.

The reason is this. On Weak Concurrentism, it doesn’t seem to matter what times the objects of desires are relativized to. According to this theory, what is good for you is a) for you to have a desire at a time, *t*, for some state of affairs (whether relativized to a time or not), and b) for this state of affairs to obtain at *t*. So on Weak Concurrentism it doesn't matter what times the objects of your desires are relativized to. Of course, many of your actual desires may in fact have objects that are relativized to times. But to determine your welfare according to Weak Concurrentism, all you need to know is whether your desires (temporalized or not) count as satisfied during the stretches of time that you have these desires. (By contrast, we will see in a moment that things are different for a theory like Strong Concurrentism, where what benefits you is getting your *present directed desires* concurrently satisfied. In order to state this theory it *is* necessary to employ the 4-place notion of desire in order to formulate this view.)

Of course, it would be easy to formulate a version of Weak Concurrentism that employs the 4-place notion of desire. To do that would require modifying the above definitions of concurrent desire satisfaction and frustration. For example, one would have to define concurrent desire satisfaction to be this: ‘S gets an episode of concurrent desire satisfaction during the whole interval <*t1, t2*> iff for every time, *t’*, that occurs during <*t1, t2*>, it is true both that i) S has, at *t’*, a desire that some state of affairs, *p*, obtains **at some time *t3***, and ii) *p* in fact obtains at *t’*.’ However, this seems needlessly confusing because this definition appeals a third time, *t3*, which isn't related to anything else in the definition.

[22] Note that Weak Concurrentism implies that Ellie’s *welfare level during the party* would be much higher if easy listening is played rather than rock ‘n’ roll. And it is not unreasonable to think that when we care about a person, we tend to try to make their *current welfare level* as high as we can, not maximize the total amount of welfare contained in that person’s life. After all, the current person is right there before us, while the person’s past and previous selves are far removed view. However, more on this later.

The sort of problem that Weak Concurrentism faces in these cases can be brought out even more clearly by considering the following pair of lives[23]:

Life 1:
At t_0, Jenny desires with intensity +10 that some event, E_1, happens at t_1.
At t_1, Jenny has no desire that E_1 happens. (That is, she has no desires at all concerning this event.)
E_1 happens at t_1.
At t_1, Jenny desires with intensity +10 that E_2 happens at t_2.
At t_2, Jenny has no desire that E_2 happens.
E2 happens at t_2.

Life 2:
At t_0, Kelly has no desire that E_1 happens.
At t_1, Kelly desires with intensity +10 that E_1 happens at t_1.
E_1 happens at t_1.
At t_1, Kelly has no desire that E_2 happens.
At t_2, Kelly desires with intensity +10 that E_2 happens at t_2.
E_2 happens at t_2.

If this is a complete description of these two lives, then Weak Concurrentism implies that they contain the same amount of welfare. Jenny has two future directed desires of intensity +10 that get concurrently satisfied – for the one desire, this happens at t_0, and for the other, this happens at t_1. Jenny has no other desires that are either concurrently satisfied or frustrated, so the total amount of welfare in her life is +20. By contrast, Kelly has two present directed desires of intensity +10 that get concurrently satisfied. For the one desire, the satisfaction happens at t_1, and for the other, it happens at t_2. Kelly has no other desires that get satisfied or frustrated, so the total amount of welfare in her life is also +20. But intuitively, it might seem that life goes much better for Kelly than for Jenny. After all, while Jenny and Kelly both get all their desires concurrently satisfied, Kelly gets 'the things she wants' while she wants them, but Jenny gets them when she no longer wants them.

Thus Weak Concurrentism might seem problematic. Instead, one might think, the Desire Satisfactionist should formulate her theory in such a way that concurrent desire satisfactions of the sort that Jenny gets are not as good for a person as concurrent desire satisfactions of the sort that Kelly gets.

[23] This case was suggested to me by Fred Feldman, though I know of no discussion of it in print.

5.3.2 Strong Concurrentism

Perhaps the most natural way to avoid the sort of implication that Weak Concurrentism has about the lives of Jenny and Kelly is to say that desire satisfactions of the sort that Jenny gets simply do not enhance one's welfare. Thus the Desire Satisfactionist might formulate a theory according to which what is good for a person is to desire, at a given time t, that something happens at t and that this thing does happen at t. This is the idea behind Strong Concurrentism. It amounts to saying that it is only concurrent satisfaction or frustration of one's *present directed desires* that count towards one's welfare. Hare, insofar as I understand him correctly, seems to endorse such a view of welfare.[24] (Heathwood and Bykvist also briefly mention views of this sort, but find them implausible for various reasons.[25])

Given the notions of concurrent desire satisfaction and frustration that were employed in Weak Concurrentism, Strong Concurrentism can now be stated very simply. The view takes it that the bearers of welfare value are episodes of what I'll call strong concurrent desire satisfaction, while the bearers of welfare disvalue are episodes of strong concurrent desire frustration:

> S gets an episode of **strong concurrent desire satisfaction** during the whole interval $<t_1, t_2>$ iff S has a present directed desire, D_{pr}, during all of $<t_1, t_2>$, and D_{pr} is concurrently satisfied during all of $<t_1, t_2>$.
>
> S gets an episode of **strong concurrent desire frustration** during the whole interval $<t_1, t_2>$ iff S has a present directed desire, D_{pr}, during all of $<t_1, t_2>$, and D_{pr} is concurrently frustrated during all of $<t_1, t_2>$.

Thus if I desire, during 1pm and 2pm this afternoon, to be skiing during this whole time, and in fact I am skiing during this whole time, then I receive an episode of strong concurrent desire satisfaction during the interval of 1pm to 2pm. So my welfare is

[24] Cf. Hare, 1981, pp. 101-104. Hare discusses whether it is 'more rational' to act so as to always maximize the satisfaction one's now-for-now and now-for-then desires, or whether it is better to maximize only one's now-for-now and then-for-then desires. (The former option corresponds roughly to what Weak Concurrentism recommends, whereas the latter option corresponds roughly to what Strong Concurrentism recommends.) Then Hare writes: 'It is possible to define "greatest happiness" (…) as the maximal satisfaction of now-for-now and then-for-then preferences. The happiest man is then, in this sense, the man who most has, at all times, what he prefers to have at those times. (…) The simplifying assumption which I shall shortly be making will turn my theory, in effect, into a happiness theory of this kind…' If Hare is using 'happiness' as interchangeable with 'well-being', then he seems to be committing himself to what I'm calling Strong Concurrentism.

[25] Cf. Heathwood, SDS, ms, p. 29. Bykvist, 2003, pp. 23-24

enhanced. By contrast, if I have this same desire, but don't get to do any skiing between 1pm and 2pm, then I receive an episode of strong concurrent desire frustration during this whole interval. What happens if I don't get to go skiing for this *whole* time, but, say, for just 45 minutes during 1pm and 2pm? In that case, we can still say that I receive an episode of strong concurrent desire satisfaction – it is just a shorter episode. In particular, if I desire, during 1pm and 2pm this afternoon, to be skiing during this whole time, but in fact I get to go skiing only between 1pm and 1:45pm, then it follows that i) I receive an episode of strong concurrent desire satisfaction that lasts for the interval between 1pm and 1:45pm, and ii) I receive an episode of strong concurrent desire frustration that lasts between 1:45pm and 2pm.

Using these notions of strong concurrent desire satisfaction and frustration, Strong Concurrentism can be stated as follows:

Strong Concurrentism

(i) Every strong concurrent desire satisfaction is intrinisically good for its subject; every strong concurrent desire frustration is intrinisically bad for its subject.

(ii) The intrinsic value for S of an episode of strong concurrent desire satisfaction equals the duration of the episode times the average intensity of the relevant satisfied desire; the intrinsic disvalue for S of an episode of strong concurrent desire frustration equals the duration of the episode times the average intensity of the relevant frustrated desire.

(iii) The total amount of welfare contained in a person's life equals the sum of the intrinsic values of all the episodes of strong concurrent desire satisfaction in the life minus the sum of the intrinsic disvalues of all the episodes of strong concurrent desire frustration contained in the life.[26]

In the cases of Ellie, the dying atheist and so on, Strong Concurrentism does not recommend present-for-past sacrifices. If one thinks that this is the right result (as Heathwood and Bykvist do), then one might think that Strong Concurrentism has an advantage over Weak Concurrentism, which did recommend past-for-present sacrifices in these cases. According to Strong Concurrentism, the future directed desires that Ellie had during most of her life to have rock 'n' roll played at her 50th birthday party have no relevance to her welfare. What does matter to her welfare, however, is the present directed desire that Ellie has during her 50th birthday party to have easy listening played

[26] Again, this view is going to face double-counting worries. As noted previously, my fix for this problem will be presented in chapter 6.

at her birthday party. Thus if rock ‘n’ roll is played at her party, she will receive an episode (a rather intense one) of strong concurrent desire frustration, and her welfare would be decreased. So Strong Concurrentism implies that the present directed desires that Ellie holds during her birthday party cannot be overridden by the future directed desires that Ellie had for all those years prior to the party. For similar reasons, Strong Concurrentism implies that it would be best for the dying atheist to have a priest called to his bedside, and that Hare’s welfare would not be enhanced to any degree if he were to take the job as an engine-driver. What’s more, the implications of Strong Concurrentism differ from those of Weak Concurrentism when it comes to the case of Jenny and Kelly. As these lives were described above, Jenny does not receive a single episode of strong concurrent desire satisfaction, while Kelly receives two. (And since Jenny has no present directed desires, she receives no episodes of strong concurrent desire frustration either.) Thus according to Strong Concurrentism, there is nothing that happens to Jenny that enhances her welfare, while there are things that happen to Kelly that *her* welfare is enhanced by. So Kelly has the better life.

Even if one finds these consequences of Strong Concurrentism to be more plausible than the consequences that Weak Concurrentism has about these cases, one should resist the temptation to endorse Strong Concurrentism. For it has other unacceptable consequences, which give sufficient reason to reject the view.

For one thing, the view fails to account for the fact that quite often our desires are just not directed at any time – past, present or future. For instance, I might desire my life would make a compelling story. Or I might desire that my life display a certain narrative structure: say, adversity followed by success (as opposed to the other way around).[27] Or I might desire that I lead a highly autonomous life, and that I not be manipulated or controlled or coerced by others. These global desires are not such that I want their objects to obtain at any particular time. Thus they are not present directed desires, and their satisfaction or frustration cannot affect my welfare according to Strong Concurrentism. However, surely a desire satisfactionist would want to say that I would be made better off if I got such these desires as these satisfied during my life. Thus Strong Concurrentism is an implausible way for the desire satisfactionist to formulate her theory.

[27] For a number of examples of this sort, see Velleman’s ’Well-being and Time’ (Velleman, 1991)

Even more worryingly, though, is that Strong Concurrentism mistakenly implies that past directed desires are entirely irrelevant to one's welfare. First, consider this case:

> *The college student who stuck it out* – Jeremy hated every minute of college. At no time during his college tenure did he desire to be in college. The whole time he desired to drop out and do something else. But he didn't. His parents threatened to punish him in various ways, so he grudgingly stuck it out and completed his degree. Suppose that in Jeremy's 30s, he becomes extremely thankful for his college education. He realizes that in his society, the people who don't go to college are at a huge disadvantage professionally. As a result, he comes to attach a lot of value to his college education. Jeremy forms a past directed desire to have gotten a college education in his youth, and he holds this desire for the rest of his life.

Since the satisfaction of past directed desires cannot impact one's welfare according to Strong Concurrentism, the view implies that Jeremy receives no benefit whatsoever from his coming to value the fact that he went to college as a youngster. However, this is highly implausible. Anyone who is sympathetic to a desire satisfactionist approach to welfare should think that Jeremy's life goes better for him because of the fact that he comes to attach value to his college education. After all, the degree to which Jeremy's desires actually are satisfied in life is clearly greater than the degree to which they would have been satisfied if he had never come to appreciate his college education. If Jeremy, for the rest of his life, had continued to wish that he had never gone to college, then the total amount of desire satisfaction contained in his life would be much lower than the total amount of desire satisfaction that his life actually contains, given that he actually comes to value his college education. But this is something that Strong Concurrentism cannot capture.

A case offered by Phil Bricker illustrates a similar flaw with this view:

> A man in his youth sets out various goals for himself, and, in the course of his life, succeeds in attaining them all. But as he enters old age, he looks back upon all his earlier activity with disgust and regret; he now believes that he has wasted his youth upon vain pursuits. How shall we evaluate this man with respect to prudence? (Bricker, 1980, p. 383)

Bricker goes on to argue that the intuitive thing to say about this case is that, when it comes to the total amount of desire satisfaction his life contains, things go worse for this man because he comes to regret the goals he worked so hard to attain as a young man. As Bricker puts it,

> his life would have been more prudent if he could have acted so as to satisfy the preferences of both his earlier and later selves (perhaps by changing himself into a later self who could accept the goals of his earlier self)… (Bricker, 1980, p. 383)

So there would have been a greater degree of match between this man's actual life and the desires, aims and goals he holds during life if he had not come to regret, as an old man, the pursuits of his youth. Thus insofar as one favors a desire satisfactionist approach to welfare, one should think that this man is harmed in terms of welfare because he forms desires not to have done the things he did as a young man.

But Strong Concurrentism cannot yield this result. After all, the desires that this man forms in his old age are all past directed. He wishes not to have pursued and accomplished the things he did when younger. But the frustration of past directed desires can have no impact on a person's welfare according to Strong Concurrentism; only present directed desires matter on this view. Thus Strong Concurrentism is unable to give the result, which desire satisfactionists should all accept, that this man is harmed by his coming to desire as an old man to have spent his youth otherwise than he actually did. Thus I take it that Strong Concurrentism is an untenable formulation of the desire satisfactionist position.[28]

[28] Could a defender of Strong Concurrentism take it that in each of the above cases, there is some present-directed desire that is indeed satisfied? The college student who stuck it out, Jeremy, might be thought to have a present directed desire to *have* completed college, and similarly the old man might be thought to have a present directed desire not to *have* wasted his time on all these frivolous activities. (Thanks to Fred Feldman for pointing out to me this way of responding to the objection.)

However, I have doubts that this response will work. After all, the canonical form of a present directed desire is such that its object consists of two things – viz. a non-temporalized state of affairs, and a time at which it is desired that this state of affairs obtains – and for the desire to be present directed, the time at which the agent has the desire must be simultaneous with the time that constitutes part of the object of the desire. But now consider the alleged 'present directed' desire that the defender of Strong Concurrentism wants to attribute to Jeremy as a 30 year old, in the case of the college student who stuck it out. The object of this desire consists of i) the state of affairs of Jeremy having gone to college in his youth, and ii) the time *now*, as the time at which it is desired that this state of affairs obtain. However, notice that the state of affairs in i) is a *temporalized* state of affairs. It is relativized to a time, namely the time when Jeremy was young. Thus the desire that the Strong Concurrentist needs to attribute to Jeremy does not fit the canonical form of a present-directed desire. After all, its object has *two* times built in. (I have been assuming throughout that allowing for multiply-relativized desires is undesirable because of the many complications it raises.) Now, Strong Concurrentism is stated in such a way that it permits only present directed desires of the canonical form to impact one's welfare. Thus Strong Concurrentism, at least as stated, cannot be defended in the way suggested above.

Of course, perhaps this defense of Strong Concurrentism can be made to work if the view is modified so that also *non-canonical* present directed desires can impact welfare. That is, perhaps the Strong Concurrentism can be modified so that also desires whose objects have *more than one* time built in can impact one's welfare. Then the suggested defense of the view might work. Nonetheless, this modification of Strong Concurrentism seems to me to opens up a pandora's box of problems. For it is not clear how to deal in a plausible way with desires whose objects have more than one time built in.

5.3.3 Discount Concurrentism

To avoid the problems with Strong Concurrentism, we might try to formulate a version of Concurrentism that allows that past directed desires, future directed desires and atemporal desires do count towards one's welfare, but at a discounted rate. Thus we might have a theory on which, all other things being equal, the satisfaction or frustration of a present directed desire would affect one's welfare *n* times as much as the satisfaction or frustration of a past directed, future directed or atemporal desire would. However, I will not bother formulating this view in detail here. It should be obvious how this would go.

Even though Discount Concurrentism does not suffer from the same defects as Strong Concurrentism, Discount Concurrentism still faces two problems that make it not be an ideal way to accommodate the temporal nature of desire. First of all, the theory does not solve the problems it was intended to solve. What was the motivation for abandoning Weak Concurrentism and seeking some other theory? It was because one might want to avoid the implications that Weak Concurrentism has in cases of unmatched future directed desires, i.e. cases like those of Ellie, the dying atheist, and Jenny & Kelly. Weak Concurrentism implies that Ellie's life would, all in all, contain more welfare if rock 'n' roll were played at her 50th birthday party than if easy listening were played. It implies that the dying atheist's life, all in all, would contain more welfare if no priest were summoned to his bedside than if a priest were summoned. It implies that Jenny's life and Kelly's life contain exactly the same amount of welfare. One might not like these implications.

At first sight, it might seem that Discount Concurrentism can avoid these consequences. On Discount Concurrentism, Ellie's present directed desire to have easy listening at her party counts for more than her the future directed desire for rock 'n'roll she held for all the years prior to her party. So perhaps this view implies that it would be better, all in all, for Ellie to get easy listening than rock 'n' roll. For similar reasons, the view might also be thought to imply that it would be best for the dying atheist to be paid a visit by the priest, and that Jenny's life is not as high in welfare value as Kelly's life is.

But on closer inspection, the Discount Concurrentist's solution will not work in all cases. In other words, if you think that Weak Concurrentism should be abandoned because of its consequences about cases like those of Ellie, the dying atheist and Jenny & Kelly, then you should not be satisfied with Discount Concurrentism either. After all, we can just construct a version of each of these cases in which Discount Concurrentism gives the supposedly ‘counter-intuitive’ result.[29] For simplicity, let’s just suppose that the discount rate for future and past directed desires is 0.5. That is, present directed desires will count for *twice* as much as future directed desires and past directed desires.[30] Suppose that for the whole time period between the ages of 15 and 49, Ellie has a future directed desire for rock ‘n’ roll at her 50th birthday party. On the day of her 50th birthday, Ellie loses the desire for rock ‘n’ roll and gets a present directed desire for easy listening instead. Thus Discount Concurrentism, too, implies that the total amount of welfare contained in Ellie’s life would be enhanced much more by having rock ‘n’ roll at the party than easy listening. So if one doesn’t like the fact that Weak Concurrentism has this result, one shouldn’t like Discount Concurrentism either. A similar point applies in the case of the dying atheist.

Discount Concurrentism doesn’t do any better than Weak Concurrentism when it comes to cases like that of Jenny & Kelly either. Granted Discount Concurrentism implies that, as the two lives were described above, Kelly’s life will be twice as high in welfare as Jenny’s life, which you might think is the right result. But this does not mean the problem is completely solved. For we can just change the numbers in the case so that Discount Concurrentism, too, has the supposedly implausible consequence that some Jenny-like life contains exactly the same amount of welfare as some Kelly-like life. For instance, let’s keep all the facts about the two lives the same, except let’s make Kelly’s desires be half as intense as Jenny’s desires. If the case is modified in this way, then Discount Concurrentism (assuming that the discount rate is still 0.5) will imply that Jenny’s life contains the same amount of welfare as Kelly’s life – even though Jenny in some sense ‘never gets what she wants while she wants it,’ while Kelly does. Thus if you

[29] Recall, that I will eventually argue that the Weak Concurrentism actually gives the *correct* result about all these cases. My point now is just that people like Heathwood and Bykvist who wouldn’t like Weak Concurrentism’s implications about Ellie and the dying atheist should not be satisfied with Discount Concurrentism either.

[30] I don’t claim to know *what* discount rate it really is best to use, but it doesn’t matter because a similar problem case can be constructed no matter what discount rate is chosen.

wanted to abandon Weak Concurrentism because of its implications about cases like Jenny & Kelly, then you should want to abandon Discount Concurrentism, too.

There is another, perhaps even more important reason to reject Discount Concurrentism. In particular, it seems totally arbitrary to discount the contributions that past directed and future directed desires, like atemporal desires, make to a person's welfare. Why make desire satisfactionsm be biased in favor desires that are directed at the same moment the desire is held, as opposed to some time that comes later or earlier than the time at which it is held (or that is not directed at any time at all)? There seems to be nothing particularly special about present directed desires, which could provide reason to think that their satisfaction counts for more than the satisfaction of other desires. I can see no independent rationale for this kind of temporal bias.[31]

Thus since Discount Concurrentism not only is unmotivated, but also temporally biased, we should not consider it to be a good way to formulate desire satisfactionism.

5.4 Bykvist's Theory

Krister Bykvist has proposed a different answer to the question of how one's well-being is impacted by the temporalized desires one might hold during life.[32] On Bykvist's view, only *some* of one's desires for things to happen at other times than the present count. This view is a very interesting proposal, but, I will argue, ultimately misguided.

5.4.1 Stating Bykvist's view

Stated in my terminology from above, the view is roughly this. For starters, the satisfaction or frustration of all one's present directed desires impact one's welfare. When it comes to future directed desires, the satisfaction or frustration of one of them impact one's welfare if and only if it is completely *matched* (in the sense I explained in section 5.1) by a present directed desire. So consider, for example, Ellie's future directed desire, which she held for all those years early in life, that rock 'n' roll music be played at her

[31] I'm not alone in disliking temporally biased formulations of desire satisfactionism. Bykvist, for instance, argues in favor of temporal neutrality. (Cf. Bykvist, 2003, p. 19-20). What's more, Bricker seems to favor temporal neutrality. (Cf. Bricker, 1980, p. 383-384). (He has also told me as much in personal correspondence.)

[32] Cf. Bykvist, 2003

50th birthday party. This desire would be matched in the relevant sense iff at the time specified in the object of the desire – the time of her 50th birthday party – Ellie also has a present directed desire to for rock 'n' roll to be played at the party. Since Ellie's future directed desire is in fact *not* matched in this way, its satisfaction or frustration will not impact Ellie's welfare, on Bykvist's view. When it comes to past directed desires, although Bykvist doesn't explicitly mention them, presumably something similar holds for them: viz. that a past directed desire counts towards one's welfare iff it is matched by a present directed desire.

This gives a rough picture of Bykvist's view. But to state it precisely, let's use Bykvist's own notion of a desire with *full inside support*. On Bykvist's theory, it is only the desires with full inside support whose satisfaction or frustration can impact one's welfare. Bykvist explains the notion of full inside support as follows:

> Roughly put, my preference for something to happen at a certain time or during a certain time period t has full inside support just in case it is *perfectly matched* by those preferences that occur at or within t. More formally, my wanting, at t_1, with intensity i, that p at t_2, has full inside support iff for every time t within t_2, either
>
> a) I do not have any preferences at t_1, or
> b) I want, at t, with intensity I, that p at t_2.[33]

This definition implies that present directed desires (i.e. desires that you have at t for something to happen at t) will always count as having full inside support. Since such a desire is about the very time at which the desire itself is held by the agent, these desires will count as 'supporting themselves.' Next, future directed desires typically need to be matched by an appropriate present directed desire in order to have full inside support. Suppose I have a desire at some time t_1 – call the desire 'D1' – that p (a non-temporal state of affairs) obtains at some later time t_2. If I still desire at this later time, t_2, that p obtains at t_2, then it follows that my future directed desire D1 has full inside support. There is another way in which D1 could have full inside support, too, namely if I am completely out of commission (e.g. unconscious, dead) at t_2. Bykvist includes this way for a desire to get full inside support in order to ensure that the satisfaction of your desires for things that happen after your death can enhance your welfare.[34] Things work

[33] Bykvist, 2003, p. 29

[34] This, of course, is a controversial view. Mental state theorists about welfare – e.g. Hedonists, or defenders of Heathwood's Subjective Desire Satisfactionism – will not be able to accept that things that happen after your death can have any impact on your welfare.

similarly when it comes to past directed desires. Suppose I have a desire at some time t_1 – call the desire 'D2' – that *p* (a non-temporal state of affairs) obtains at some earlier time t_0. D2 would have full inside support if it was the case, at the earlier time t_0, that I also desired that p occur at t_0. The other way in which D2 would count as having full inside support is if I were completely out of commission (e.g. unconscious, not born yet) at t_0.

Bykvist endorses the thesis that when it comes to welfare, 'a (…) preference counts iff it has full inside support.' (Bykvist, 2003, p. 29) We can state a version of desire satisfactionism that conforms with this thesis. To state the theory in a simple way, I'm going to coin two terms:

> S gets an episode of **Bykvist-ian desire satisfaction** during the whole interval $<t_1, t_2>$ iff S has a desire, D, with full inside support during all of $<t_1, t_2>$, and D is concurrently satisfied during all of $<t_1, t_2>$.
>
> S gets an episode of **Bykvist-ian desire frustration** during the whole interval $<t_1, t_2>$ iff S has a desire, D, with full inside support during all of $<t_1, t_2>$, and D is concurrently frustrated during all of $<t_1, t_2>$.

Using these terms, we can state a version of desire satisfactionism that conforms to Bykvist's views as follows:

> <u>Bykvist-ian Desire Ssatisfactionism</u>
>
> (i) Every Bykvist-ian desire satisfaction is intrinisically good for its subject; every Bykvist-ian desire frustration is intrinisically bad for its subject.
>
> (ii) The intrinsic value for S of an episode of Bykvist-ian desire satisfaction equals the duration of the episode times the average intensity of the relevant satisfied desire; the intrinsic disvalue for S of an episode of Bykvist-ian desire frustration equals the duration of the episode times the average intensity of the relevant frustrated desire.
>
> (iii) The total amount of welfare contained in a person's life equals the sum of the intrinsic values of all the episodes of Bykvist-ian desire satisfaction in the life minus the sum of the intrinsic disvalues of all the episodes of Bykvist-ian desire frustration contained in the life.

Bykvist's view is designed to avoid the implications that a view like Weak Concurrentism has for cases of unmatched future directed desires, like the cases of Ellie, the dying atheist, Hare the engine-driver, Parfit the poet, and so on. In other words, the theory is supposed to avoid

> the *present-for-past* sacrifice illustrated by the cases about Hare and Parfit. Other things being equal, it is never right to sacrifice a person's present preferences about his present life for the sake of her past preferences about her present life. (Bykvist, 2003, p. 24)

The reason that this is a desirable result, Bykvist thinks, has to do with considerations of autonomy:

> I argued that the problem of past preferences arises when these preferences of past selves 'poke their nose' into present selves' lives and private concerns. This way of putting things suggests that what is at issue is some kind of respect for the autonomy of person-stages, and I chose to spell out this respect in the following way: each person-stage has a veto over what they should do with their lifestage so any conflicting preferences of other persons or other temporal stages of the same person are to be completely disregarded. I stressed, however, that in the intrapersonal case the autonomy of person-stages is not the only thing that matters. Since we do not just lead our lives from the perspectives of individual moments but also from a more comprehensive diachronic point of view, it seems plausible to count those past preferences of a person that agree with her present preferences. (Krister Bykvist, 2007, p. 74)

Bykvist's idea here is that since each person-stage is autonomous, it would be misguided to allow the preferences of a particular person-stage to be overridden by the conflicting preferences of earlier or later person-stages.[35] Bykvist-ian Desire Satisfactionism is a theory that seems to give these results.

5.4.2 Against Bykvist's view

Nonetheless, I do not think that Bykvist-ian Desire Satisfactionism represents a good way for those who are sympathetic to a desire satisfactionist approach to welfare to formulate their view. There are four reasons for this. First of all, I think the intuitions that Bykvist has designed his theory to capture are misguided. I will argue in the last section of this chapter that the desire satisfactionist can and should allow that your total welfare would be enhanced by the satisfaction of desires that you held earlier in your life, even if you no longer hold those desires at the current time. Thus I will argue, for example, that there is nothing wrong with thinking that the total amount of welfare contained in Ellie's life would be enhanced more by playing rock 'n' roll at her party than by playing easy listening – even though easy listening is what she wants during the party itself. Accordingly, since Bykvist-ian Desire Satisfactionism is specifically designed so as not

[35] Bykvist emphasizes that he does not mean his talk of personstages to commit him to any particular view about personal identity, e.g. the denial of view that a person is 'wholly present' at any given moment. He says that 'my use of 'person-stage' was only meant to be a convenient way of referring to a person as he is at a particular time…' (Bykvist, 2007, p. 74)

to give these results about cases like that of Ellie, I think Bykvist-ian Desire Satisfactionism is unmotivated. However, more on this in the last section of this chapter.

What's more important right now is that Bykvist's view doesn't fully capture the intuitions that it is designed to capture in the first place. Suppose for the moment that we *do* want our theory of welfare not to recommend what Bykvist calls present-for-past sacrifices (i.e. to not imply that your welfare can be maximized by giving you something you wanted in the past but which you don't want at present). Bykvist's theory fails to rule out such sacrifices because of a problem with his definition of 'full inside support.' Consider the following modified version of the dying atheist case:

> *The case of the comatose dying athiest*: Between the ages of 20 and 70, the atheist has a strong future directed desire, D1, that no priest come to visit him when he is on his deathbed. When he is 70, the atheist's resolve weakens and he now forms a new desire, D2, that a priest visit him when he's on his deathbed. He holds D2 for a total of one week before he gets into a car crash and falls into a coma. The atheist is put on life support. He will not wake up. Thus the hospital bed is his deathbed. However, if he were to wake up, he would desire what he did right before the accident, viz. to have a priest come visit him on his deathbed.

In fact, Bykvisti-ian Desire Satisfactionism implies that the total amount of welfare contained in the atheist's life would be maximized by having no priest come to give him his last rites. After all, both of the atheist's future directed desires, D1 and D2, count as having full inside support. This is because the atheist has no desires whatsoever during the time towards which D1 and D2 are directed – i.e. the time during which he is on his deathbed. Thus the satisfaction or frustration of D1 and D2 would indeed have an impact on the atheist's welfare, according to Bykvist-ian Desire Satisfactionism. Because the duration of D1 is so much longer than the duration of D2, the theory implies that getting D1 satisfied would enhance the atheist's welfare much more than getting D2 satisfied.

However, I do not think Bykvist would want to accept this result. What Bykvist-ian Desire Satisfactionism recommends in this case seems to be very close to the sort of present-for-past sacrifice that Bykvist is so concerned to avoid. The theory implies that we could enhance the atheist's welfare significantly more by ignoring his most recent desires – which are also the desires he would have if he were to wake up from his coma – and instead satisfy his prior longstanding desire for no priest. If Bykvist is troubled by theories that recommend present-for-past sacrifices in the original dying atheist case (involving no coma), then he should be troubled by the implications that his own theory has about the present case, too. Thus it seems Bykvist's theory doesn't do the job it was designed to do.

A third problem for Bykvist's view is that it seems arbitrary.[36] Why should past directed and future directed desires count only if they are matched by an appropriate present directed desire? If we are willing to admit that *some* future or past directed desires can count towards welfare, why not say that they all count? Suppose that right now I hold two equally intense future directed desires: one to write an influential book on moral philosophy, and another to make a million dollars. However, I justifiably come to believe that I can't do both, so I force myself to give up my desire for the million dollars. I hang on to my desire to write the book, though, and finally as an old man I succeed in writing a book on moral philosophy that is influential. Moreover, suppose that (as unlikely as this is) my book becomes a best-seller and I make a million dollars from the royalties. But at this late stage in my life, I remain entirely uninterested in the money. I have no desire for it. Since my future directed desire to write the book on moral philosophy is matched by a present directed desire, Bykvist-ian Desire Satisfactionism implies that my welfare is enhanced by the satisfaction of *this future directed desire itself*. But since my other future directed desire (to make a million dollars) is not matched by a present directed desire, Bykvist's view implies that *its* satisfaction does not enhance my welfare. This seems arbitrary. After all, considered in and of themselves, the first future directed desire is no different in kind from the second.[37] Why, then, should the satisfaction of the first one enhance my welfare but not the satisfaction of the second one?

Even more damagingly, Bykvist-ian Desire Satisfactionism has a range of unacceptable consequences. Recall the two cases of unmatched past directed desires, which Strong Concurrentism could not accommodate. Bykvist's view cannot accommodate them either. First, there was *the case of the college student who stuck it out*. The whole time Jeremy was in college, he desired not to be. Ten year's later, he

[36] A similar criticism is made in McKerlie, 2007b, pp. 62-63

[37] Notice that Bykvist's suggestion makes the intrinsic value of a state consisting of the satisfaction of a future directed desire depend on features that are *extrinisic* to this state. This might conflict with the axiological principle that the intrinsic value of something must depend on its intrinsic features. (Cf. Feldman, 2004, p. 73)

comes to value his college education. Thus Jeremy forms a past directed desire to have gone to college, but it is unmatched by any present directed desire in the relevant time period. I claimed that desire satisfactionists should think that Jeremy is benefited by forming a desire after the fact to have gone to college. After all, this makes the total degree of fit between his life as a whole and his desires higher. His life seems to contain a greater amount of desire satisfaction than it would if he hadn't formed any desire after the fact to have gone to college. But Bykvist-ian Desire Satisfactionism cannot yield this result. After all, Jeremy's past directed desire to have gone to college does not have full inside support, since it is not matched by any present directed desire. Thus its satisfaction does not count towards his welfare.

For similar reasons, Bykvist-ian Desire Satisfactionism cannot account for Bricker's case of the old man who comes to regret the activities of his youth. We rejected Strong Concurrentism on the grounds that it failed to imply that the man's welfare is decreased by the fact that he comes to develop, in his old age, a past directed desire not to have spent his youth as he did. Bykvist-ian Desire Satisfactionism cannot yield this result either. After all, the old man's past directed desire does not have full inside support. It is not matched by any present directed desires that he held in his youth. So the frustration of the old man's past directed desire does not reduce his welfare. This, too, is an unacceptable result. Thus the implications that Bykvist-ian Desire Satisfactionism has about these two cases give sufficient reason to reject the theory.[38]

[38] Not only does Bykvist's theory have trouble with cases of unmatched past directed desires, it might also have problems with certain cases of unmatched *future directed desires*. Suppose that as a youngster I desire that I will be in a loving relationship. But suppose that I grow bitter as I get older, and as a result never again desire to be in a loving relationship. Finally, suppose as an older man, I have a chance to be in a relationship that would be very tender and loving. If I were to be plunged into the relationship, would I thereby be benefited? Bykvist-ian Desire Satisfactionism implies that the answer is 'no.' My bitterness as an old man prevent me from having any present directed desires for this relationship. Thus the future directed desire that I had as a young man does not have full inside support. So its satisfaction would not benefit me in any way. But this seems strange. Intuitively, I would be benefited by being plunged into the loving relationship and getting the future directed desire I held as a young man satisfied. However, this case is admittedly complex. So the present argument might not convince everyone.

5.5 Some Other Ways to Accommodate the Temporality of Desire

For completeness, I want to briefly mention three other ways in which a desire satisfactionist might try to accommodate the connection between desires and time. None of these solutions are successful, either, in my view.

5.5.1 Desires that are conditional on their own persistence

Bricker,[39] Bykvist,[40] and Heathwood[41] all point out that a desire satisfactionist might try to appeal to the idea of desires that are conditional on their own persistence in order to accommodate cases in which desires change over time (like the cases of Ellie, the dying atheist, etc.). When are desires conditional on their own persistence? Suppose you desire to eat peach ice cream after dinner. We can say that this desire is conditional on its own persistence if it is *really* the case that the object of your desire here is this: to eat peach ice cream provided that this is still what you want at the time. Heathwood gives a clear explanation of how this idea can be employed in the service of desire satisfactionism:

> when a person has a desire about the future that is conditional upon its own persistence, its satisfaction is intrinsically good for the person when and only when the condition is satisfied – i.e., when the desire persists. Perhaps Ellie's original desire for rock 'n' roll was implicitly conditional upon its own persistence: as a teenager she wanted rock 'n' roll at her 50th birthday party only if, on that day, she would still want rock 'n' roll. Since the condition isn't satisfied, satisfying the conditional desire doesn't benefit her.[42]

However, even if the desire satisfactionist can use this strategy in some cases, it will not do the trick in all cases – as Bricker, Bykvist and Heathwood all point out. After all, clearly not all our desires are conditional on their own persistence. Bykvist points out that 'all preferences that express *ideals* seem not to be conditional on their own persistence.'[43] If the desires involved in the case of Ellie, or the dying atheist, or Hare the engine-driver or Parift the poet are unconditional ones (which might be especially plausible for the dying atheist and Parfit cases), then the problem would remain.

[39] Cf. Bricker, 1980, p. 389

[40] Bykvist, 2003, p. 21

[41] Heathwood, SDS, ms, p. 11. (Also see Parfit, 1984, p. 151)

[42] Heathwood, SDS, ms, p. 11

[43] Bykvist, 2003, p. 21

5.5.2 Heathwood's Subjective Desire Satisfactionism

Another way for the desire satisfactionist to deal with the problems caused by the temporality of desire is to introduce an Experience Requirement into the theory. This amounts to demanding that desire satisfactions and desire frustrations must be *experienced* in order for them to affect your welfare. Chris Heathwood defends a desire satisfaction theory of welfare that is explicitly based on this idea. He sketches his view, which he calls *Subjective Desire Satisfactionism*, as follows:

> Let's call a state of affairs in which a subject simultaneously desires and believes a single proposition a *subjective desire satisfaction*. A *subjective desire frustration* is a state of affairs in which a subject negatively desires and simultaneously believes a single proposition. The theory I wish to defend – *Subjective Desire Satisfactionism* – begins, to a first approximation, with the thesis that every subjective desire satisfaction is intrinsically good for its subject, and every subjective desire frustration is intrinsically bad for its subject. (Heathwood, SDS, ms, p. 17)

Heathwood goes on to refine his view in several ways. Here, I will not state the theory in full, since Heathwood does this well enough himself.[44]

One of the advantages Heathwood thinks his view possesses is that it gets cases like that of Ellie right:

> SDS is able to answer the Argument from Changing Desires because it requires simultaneity: desire and belief must be simultaneous for a subjective desire satisfaction to occur. So if a person's longstanding desire changes, we do not benefit her by giving her the thing that is no longer desired, for then she'll have the belief without the desire. More traditional desire satisfaction theories are different in this respect. Since merely past desires about the (then) future can be satisfied, we can benefit people by giving them what they no longer want. This is what happened at Ellie's 50th Birthday Party. (Heathwood, SDS, ms, 28)

The idea here is that a theory of welfare should not imply that Ellie would be benefited by getting 'what she no longer wants', and Heathwood thinks Subjective Desire Satisfactionism is a plausible theory, in part, because it indeed does not imply this.

In my view, Subjective Desire Satisfactionism does not represent a good way for the desire satisfactionist to deal with the temporality of desire. For one thing, I think Heathwood is mistaken in claiming that the implications of Subjective Desire Satisfactionism about the Ellie case constitute a reason to prefer that theory over other versions of desire satisfactionism. I will argue in the last section that, from the

[44] One thing Heathwood should do to improve his theory (which he mentions only briefly – see p. 23) is to incorporate degrees of belief. After all, belief is not an all-or-nothing affair. Sometimes we attach more credence to a proposition, sometimes less. Heathwood's theory can easily be made to accommodate this. He should just say that a higher degree of belief in a proposition that one desires true would enhance welfare more.

perspective of desire satisfactionism at least, there is nothing wrong with the idea that 'we can benefit people by giving them what they no longer want.' Thus my view is that it is no advantage of Subjective Desire Satisfactionism that people actually cannot be benefited in this way.

More importantly, however, Subjective Desire Satisfactionism is not an attractive route for the desire satisfactionist to take because Heathwood's theory is vulnerable to Experience Machine arguments.[45] According to Subjective Desire Satisfactionism, what intrinsically enhances one's welfare is a certain kind of mental state: viz. the state of desiring that something is the case and simultaneously believing that this thing is the case. Thus Subjective Desire Satisfactionism is a *mental state theory*. These are the theories that imply that there can be no difference in the welfare value of two people's lives without there also being some difference in these people's mental states.

However, Experience Machine scenarios pose a well-known problem for the mental state theories. Suppose A leads a life in the real world, in which he overcomes adversity to lead a successful life that is filled with meaningful and loving relationships. Suppose A's twin, B, leads an experientially indistinguishable life, but his takes place entirely inside the experience machine. Who has the better life? Most would agree that A's life is the better one, even though A and B are mentally indistinguishable. If one agrees with this judgment (and I, for one, do agree), then one must reject the mental state theories of welfare. So it seems that adopting Subjective Desire Satisfactionism in order to deal with the temporality of desire would be to trade one problem for a much bigger one.

Moreover, it is often taken to be an advantage of desire satisfaction theories of welfare, in general, that they have the resources to avoid the mental state theories' problematic implications about Experience Machine scenarios.[46] But adopting Subjective Desire Satisfactionism requires giving up this advantage. Desire satisfactionists would presumably not be happy to do this just in order to deal with the temporality of desire. Accordingly, I don't think that Subjective Desire Satisfactionism represents the best route for the desire satisfactionist to pursue.

[45] Though he doesn't mention Heathwood specifically, Bykvist makes a similar criticism of the 'Experience Requirement' solution. (Cf. Bykvist, 2003, p. 22)

[46] See, for example, Kagan, 1998, pp. 35-37

5.5.3 Ideal Desire Satisfactionism

A final strategy that the desire satisfactionist might appeal to is the move to some kind of ideal desire satisfactionism. That is, one might adopt a version of desire satisfactionism on which it is not the case that you are benefited by the satisfaction of the desires you *actually* have, but rather by the satisfaction of the desires you *would* have if you were made more ideal in certain respects – e.g. you are given full information, enhanced intelligence, complete rationality, etc.[47] The idea behind this solution, as Bykvist explains it, is that one could try to argue that 'we will not have any intertemporal conflicts between rational preferences, since rational preferences cannot change over time: If I rationally want *p* at one time I will always rationally want *p*.'[48] None of the cases of desire change we have been struggling with here – viz. Ellie, the dying atheist, Hare the engine-driver and Parfit the poet – involve people's ideal desires. And so these cases can pose no problem for ideal desire satisfactionism.

However, this strategy will not provide a complete solution either. After all, why think that one's ideal preferences must remain unchanged? Why think it is impossible that, if an ideal version of myself, at t_1, would desire something, then at a later time, t_2, an ideal version of myself might not desire it anymore? Sometimes our desires change not because we gain new information, or become more rational or what have you, but simply because our feelings change. Sometimes our desires change for no special reason. Thus I think there can be cases of *ideal* desire change, too. So we could construct versions of the cases of Ellie, the dying atheist, and so on, in which all the relevant desires are ideal ones. Accordingly, ideal desire satisfactionism must deal with the issue of desire change, just like actual desire satisfactionism.

5.6 In Defense of Weak Concurrentism

We have seen a number of unsuccessful ways in which the desire satisfactionist might try to deal with the temporality of desire. I will argue that, in light of this sad state of

[47] Brandt (1979), Griffin (1986) and Railton (2003) defend different versions of such a view.
[48] Bykvist, 2003, p. 21

affairs, Weak Concurrentism is the best route open to the desire satisfactionist. It does not have the unacceptable consequences about unmatched past directed desires that led us to reject Strong Concurrentism and Bykvist-ian Desire Satisfactionism. Nor is it arbitrary in the way that cast doubt on both Discount Concurrentism and Bykvist-ian Desire Satisfactionism. Nor does it suffer from the disadvantage of being a mental state theory, which undermines Heathwood's Subjective Desire Satisfactionism.

However, Weak Concurrentism has some challenges to overcome, too. In particular, Weak Concurrentism might be thought to have trouble with cases of unmatched future directed desires – i.e. with the cases of Ellie, the dying atheist, Hare the engine-driver, Pafit the poet, and the Jenny & Kelly case. The problem is that Weak Concurrentism implies, as Heathwood puts it, that 'we can benefit people by giving them what they no longer want.' It sometimes recommends, in other words, what Bykvist calls 'present-for-past sacrifices.'

Nonetheless, my view is that these implications of Weak Concurrentism, which Bykvist and Heathwood find so troubling, are in fact not a problem for the view. I will argue that desire satisfactionists should not think there is anything wrong with the idea that 'we can benefit people by giving them what they no longer want.' In particular, my strategy will be to explain away the judgments of people like Heathwood and Bykvist about the offending cases.

First, briefly recall the argument against Weak Concurrentism. Really it is a string of arguments: for we have several cases of unmatched future directed desires that Weak Concurrentism supposedly has unacceptable consequences about.

a) In the case of Ellie, Weak Concurrentism implies that the total amount of welfare contained in Ellie's life would be greater if rock 'n' roll is played at her party than if easy listening is played. But, intuitively, it would be better for Ellie to get what she wants during the party itself, namely easy listening.

b) In the case of the dying atheist, Weak Concurrentism implies that the total amount of welfare contained in the athiest's life would be greater if no priest gets called to his deathbed than if a priest were called. But, intuitively, it would be better for the atheist if he gets what he wants when on his deathbed itself, namely a priest.

c) In Hare's case, Weak Concurrentism implies that Hare's welfare would indeed be enhanced somewhat if he takes the engine-driver job (even though at the time in question he has no desire to do so). But, intuitively, this is not the case.

d) In Parfit’s case, Weak Concurrentism implies that Parfit’s welfare would indeed be enhanced somewhat if he writes some poetry (even though at the time in question he h as no desire to do so). But, intuitively, this is wrong.

e) When it comes to Jenny and Kelly, Weak Concurrentism implies that their lives contain the same amount of welfare. But, intuitively, this is not the case. Kelly has the better life.

I do not think desire satisfactionists should be troubled by these results. It seems to me that the only reason why one might find the implications of Weak Concurrentism about these cases to be implausible is that certain facts about happiness and unhappiness (or pleasure and pain, if you like) are contaminating one’s intuitions. For example, consider Heathwood’s description of the Ellie case:

> She continues to desire for years and years that there be rock ’n’ roll at her 50th birthday party. But a month before the party, Ellie ceases enjoying rock ’n’ roll. She now prefers easy listening, and finds rock ’n’ roll **loud, childish, and annoying**. She will continue to feel this way on her 50th birthday. **She would have the time of her life** at her party if she got easy listening, but **would be miserable** if rock ’n’ roll were played. (Heathwood, SDS, ms, p. 10)

Notice the phrases I’ve put in bold. They seem designed to give the impression that if rock ‘n’ roll were played at Ellie’s party, she would experience a lot of unhappiness and discomfort, but if easy listening were played instead, she would experience a great deal of happiness and enjoyment. Thus it seems that the happiness and unhappiness facts about the case are what, in large part, is responsible for the judgment that it would be better for Ellie to get rock ‘n’ roll at the party than easy listening. Presumably something similar can be said about the other cases. Of course, the appeal to pleasure and pain facts is not as explicit in those other cases as it is in Heathwood’s description of the Ellie case. But it is likely that our judgments about those cases are driven by a tacit appeal to considerations about happiness and unhappiness, as well.

However, if this is right, then the current objection to Weak Concurrentism would be based on general Hedonistic intuitions, not intuitions that are specifically concerned with the relation between time and the satisfaction of desire. Now, perhaps it is a good argument against desire satisfaction theories in general that they can capture neither the intrinsic value that happiness or pleasure seem to have for a person, nor the intrinsic disvalue that unhappiness or pain seem to have for a person.[49] But this would be an altogether different problem from the one we are concerned with in this chapter. When it comes specifically to the temporal nature of desires – and the resulting phenomena of desire change, future or past directed desires, etc. – Weak Concurrentism gives perfectly acceptable results.

This is apparent when the cases are described in such a way as to neutralize any tacit appeal to facts about happiness and unhappiness. The story of Ellie must be told so as to make explicit that the amount of happiness minus unhappiness that she receives will be the same no matter whether it is rock 'n' roll or easy listening that ultimately is played at her party. The story of the dying atheist must be told in such a way that he receives the same amount of happiness minus unhappiness no matter whether a priest is called to his death bed or not. Similarly for the other cases. The alternatives in each one must be made hedonically equivalent. For it is only in this way that we can isolate our intuitions specifically about the relation between time and the satisfaction of desire.

Once the stories are told in this way, the implications of Weak Concurrentism do not seem troubling any more. For instance, if Ellie will feel the same amount of happiness minus unhappiness no matter whether she gets rock 'n' roll or easy listening at her party, then it does not seem implausible to say that her total welfare would be greater if she were to get rock 'n' roll. In fact, it seems that this is what the clear-headed desire satisfactionist *must* say. After all, in light of the longstanding future directed desire for rock 'n' roll that she held for all those years, the total amount of desire satisfaction in Ellie's life would seem to be much greater in the case where she gets rock 'n' roll than it would be in the case where she gets easy listening. Viewed as a whole, Ellie's life would seem to *fit her various desires to a much higher degree* in the possible world where she gets a rock 'n' roll party than in the possible world where she gets an easy listening party.

Similar points apply to the other cases. To take just one more, recall the case of Jenny and Kelly. To isolate our intuitions specifically about the relation between time and desire (and to avoid a tacit appeal to hedonistic intuitions), suppose that the total amount of happiness minus unhappiness is exactly the same in Jenny's life as it is in Kelly's life.

[49] See Heathwood's discussion of the 'dead sea apples' objection. Also see the last section of chapter 7 of this dissertation.

Now who has the better life? Granted, Kelly in a certain strong sense 'gets what she wants while she still wants it', while Jenny does not. But it still seems as though the degree to which Jenny's life, viewed as a whole, fits with her various desires is exactly the same as the degree to which Kelly's life, viewed as a whole, fits with *her* various desires. The two lives seem to be equal with respect to the total amount of desire satisfaction that they contain. Thus the clear-headed desire satisfactionist should not have any trouble accepting the implication that Weak Concurrentism has about this case, namely that Jenny's life contains the same amount of welfare value as Kelly's life.

Perhaps one might object to what I have said here. If one feels strongly that these implications of Weak Concurrentism are problematic, one might object that even if the cases are described in the anhedonic way that I want, the results of Weak Concurrentism are still wrong. For instance, suppose we grant that a rock 'n' roll party would produce the same amount of happiness minus unhappiness for Ellie as an easy listening party would. Even so, one might argue, it is counter-intuitive to say that her welfare would be enhanced most by the rock 'n' roll party. After all, the objection goes, if *you* were the friend of Ellie's who was put in charge of arranging her party, and then you found out that she had completely lost her longstanding desire for rock 'n' roll at her party, would *you* disregard that fact and throw a rock 'n' roll party nonetheless? No. Even if you knew Ellie wouldn't be particularly pained or pleased either way, any good friend would still try to comply with her current wishes and get her some easy listening for her party. That's what someone who cared about Ellie would do.

Two points in response to this objection. For one thing, it should be emphasized that even Weak Concurrentism implies that what would be best for Ellie *as she is now*, on the day of her party, is to get easy listening. The person-stage (if I can be permitted to use this phrase without committing myself to any views about the nature of personal identity) that consists of Ellie on her 50th birthday would clearly be benefited more by getting an easy listening party than a rock 'n' roll party. After all, that's what Ellie as she is now wants now. And Weak Concurrentism does indeed imply this. (Though it also implies that what would maximize the total amount of welfare contained in her life as whole is to get the rock 'n' roll party.)

Second, if one still has the intuition that what a good friend would do for Ellie (even in the anhedonic version of the case) is to arrange an easy listening party, then I would respond by claiming that this shows nothing about what would in fact maximize the total amount of welfare in Ellie's life. For it does not seem that the *only* thing involved in being a good friend to a person, or in caring about that person, is seeking to maximize the total amount of welfare contained in this person's life. I think that, for purely psychological reasons, being a good friend to some person, P, or caring about P, is likely to involve a certain amount of *temporal bias* in favor P as P is at the current moment. Suppose I care deeply about my new wife, who I met just two years ago. I know her as she is now, but not as she was many years ago. I see her as she is now. I interact with her as she is now. I have a very vivid image of who she is now, while I have but a dim inkling of who she was as a child. Thus it makes sense, psychologically, that I would be much more concerned with doing what is good for her as she is now (who I am intimately acquainted with), as opposed to what would be good for her as she was in earlier times (who I don't know well at all). For that reason, I might – precisely *because* I care about her – feel a strong (but temporally biased) inclination to do what maximizes her welfare as she is now, rather than what maximizes the total amount of welfare contained in her life as a whole.[50]

I think this explains why some people might think that what a good friend would do for Ellie (even in the anhedonic version of the case) is to arrange an easy listening party. It is not because an easy listening party would maximize the total amount of welfare contained in Ellie's life, but rather because an easy listening party would maximize the welfare of Ellie *as she currently is* and her current friends are, for natural psychological reasons, quite likely to be biased in favor of Ellie's current self.

So my view is that the implications of Weak Concurrentism about the cases we have been considering here are in fact quite plausible ones, as far as the desire satisfactionist is concerned. If one is committed to being a genuine desire satisfactionist about welfare, then one must resist the Hedonistic intuitions that seem to have been behind the original

[50] I think that something exactly analogous to this can be said also when it comes to *self*-concern, or *self*-love. While I remember to some extent who I was as a kid, I experience who I am right now with a much greater level of vividness. So of course I'm going to care more about doing what's good for my current self than what would be good for my self as a child. But this doesn't show that it's always the case that the total amount of welfare contained in my life would be maximized by doing what is good for my current self.

judgments about the cases of Ellie and the rest. A genuine desire satisfactionist cannot reject Weak Concurrentism because it conflicts with these Hedonistic intuitions (i.e. because a genuine desire satisfactionist cannot accept these Hedonistic intuitions in the first place). And once this source of confusion is eliminated, the implications of Weak Concurrentism should seem perfectly acceptable to the desire satisfactionist. Thus, in light of the failure of all the other versions of desire satisfactionism to plausibly account for the temporality of desire, I think Weak Concurrentism represents the best way for desire satisfactionists to handle the ways in which desire and time are connected.

CHAPTER 6

CLOUD DESIRE SATISFACTIONISM

My main aim in this chapter is to formulate an adequate version of the well-known theory, Actual Desire Satisfactionism. The view, although widely discussed, is much more problematic than is often acknowledged. It faces a number of difficult technical problems that too often simply get glossed over. Some of these problems – namely the ones concerning desire and time – have already been discussed at length in the previous chapter. In this chapter, then, I discuss three other serious difficulties that the desire satisfactionist must address before she can state her view adequately. I develop a theory called 'Cloud Desire Satisfactionism' in order to solve these problems.

6.1 Classical ADS

Desire Satisfactionism is roughly the view that your life goes well for you to the extent that you get your desires satisfied. The simplest version of this theory is Actual Desire Satisfactionism (ADS), according to which it is the satisfaction of your *actual* desires that benefits your welfare. I begin by presenting a simple version of ADS, which I will call ‘Classical ADS’ because it corresponds – more or less[1] – to standard ways of formulating the view. To state Classical ADS, we need a few basic concepts.

Desire. Desires are pro-attitudes that people bear towards states of affairs.[2] If you want it to be the case that you own a weekender yacht, then you bear the attitude of desire towards the states of affairs in which you own a weekender yacht. Note that it will not be necessary to introduce a corresponding negative attitude, e.g. of aversion, that you have when you want some state of affairs *not* to obtain. After all, any time you have an aversion to some state of affairs, S, it would trivially be the case that you have a desire for the more complex state of affairs of [not-S]. As a result, I think we can forgo the notion of aversion and just make do with the notion of desire. [3]

Duration. Sometimes one desire will last longer than another. Thus I am assuming that desires have durations. They persist for certain amounts of time. I have had my desire to own the *Deadwood* series on DVD for a couple weeks now, and when I eventually lose interest in the series (as often happens with such things) my desire will be gone. The stretch of time during which I have the desire is the desire’s duration.

Strength. Sometimes we desire some things more strongly than other things. For instance, while I currently desire that I own the television series *Deadwood* on DVD and I currently desire that I get a new pair of gloves, my desire to own the television series is stronger than my desire for the gloves. If I can’t have both, I would prefer to forgo the gloves and just get the DVDs. Thus it makes sense to talk about some desires being *stronger* than others.

The strength of a desire is a matter of how much you want its object to obtain. As a rough way to ascertain the strength of a desire, we might ask you, for instance, “How much pain would you endure to see this desire realized?”. Alternatively, we could ask

[1] I think the only non-standard element of Classical ADS is its appeal to the notions of concurrent desire satisfaction and frustration. But I have already defended the use of concurrency in formulating desire satisfaction. So I lump it in with Classical ADS.

[2] I take it we could just as well say that desires are attitudes that people bear towards *propositions*. Because I talk so much in other chapters about how various theories say the welfare values of states of affairs are to be determined, it will be convenient to understand Desire Satisfactionism in terms of states of affairs rather than propositions. I don’t think anything significant turns on this.

[3] What’s more, although this might contrast with everyday speech, I am assuming that we often keep our desires even after we get them satisfied. (Thanks to Kelly Trogdon for pressing me on this.) This is because I am understanding the notion of desire in a rather wide way. I will take it that to desire that something be the case is roughly the same as to attach value to this thing’s being the case, or to have a preference to some degree or other that this thing is the case.

you, “For these two particular desires, which of them would you prefer to get satisfied if you can’t satisfy them both?”. Thus the stronger your desire for X, the more inclined you will be to prefer getting X over getting other things instead. What’s more, the stronger your desire for X, the more inclined you will be to take the apparent means to obtaining X. To formulate ADS, it will be helpful to assume that the strength of a desire can be represented on a numerical scale. That is, for any desire, there is going to be some number that corresponds to its strength.[4]

Note that some philosophers have used the phrase ‘intensity of desire’ to talk about the strength of a desire.[5] I think this invites a confusion, however, and so I prefer not to use this terminology. The problem is that saying that desires differ in intensity makes it sound as though what matters is the strength of the *feelings* associated with the desire.[6] Talk of the intensity of desire makes it sound as though having a desire is like experiencing a powerful craving for, say, Indian food. It sounds as though the intensity of your desire for X is a function of how hot and bothered you get when you think about X, how infatuated you are with the thought of getting X, and so on.

But this is a mistake. A desire need not be accompanied by any particular feeling. I might strongly want something to be the case without having any feelings of infatuation or craving or what have you. For instance, right now, I strongly desire to have a dry and warm place to sleep at night. I would decisively prefer having a dry and warm place to sleep at night to having, say, the *Deadwood* DVDs or a new pair of gloves. I would be much more inclined to take the apparent means to getting myself shelter than to getting either the DVDs or the gloves. Thus right now my desire for shelter is stronger than my

[4] I am not sure there is any need to take it that there is a maximum degree to which a person can desire that some state of affairs be the case. So for simplicity, I’m just going to assume that there is in principle no upper bound on the strength of a desire.

[5] Parfit seems to accept this appeal to intensity of desire: ‘In deciding which alternative would produce the greatest total net sum of desire-fulfillment, we assign some positive number to each desire that is fulfilled, and some negative number to each desire that is not fulfilled. How great these numbers are depends on the intensity of the desires in question.’ (Parfit, 1984, p. 496) Also see Heathwood, SDS, ms. Heathwood uses the term ‘intensity’ but notice that he is careful to avoid the sort of confusion I describe here.

[6] Thinking that the intensity of a desire is a matter of the strength of the feelings associated with it is, I suspect, what led James Griffin to argue as follows:

> One does not most satisfy someone’s desires simply by satisfying as many as possible, or as large a proportion. One must assess their strength, not in the sense of felt intensity, but in a sense supplied by the natural structure of desire. (…) felt intensity is too often a mark of such relatively superficial matters as convention or training to be a reliable sign of anything as deep as well-being. (Griffin, 1986, p. 15)

desire for either the DVDs or for the gloves. However, since I know I have a decent apartment to go home to, it's not the case that there are any intense feelings associated with my desire for shelter. I do not crave shelter. I am not infatuated with the thought of having shelter. Nonetheless, I have a strong desire for it.

Accordingly, it is a mistake to think that the strength of a desire is a the strength of the feelings associated with the desire. Instead, it would be better to say that the strength of a desire of yours corresponds to the strength of your disposition to prefer getting that desire satisfied over getting other things (in situations where you can't get both). The more this desire would tend to outweigh other conflicting desires, the more strength it has.

Instrumental vs. Non-instrumental desire. To make the preliminary version of ADS more plausible, we should assume that one's welfare would not be enhanced by the satisfaction of one's instrumental desires. What is an instrumental desire? A (perhaps not very illuminating) account of instrumental desire is this:

> For any state of affairs, *p*, and person, S, S has an *instrumental* desire that *p* obtain iff S desires that *p* obtain only because i) there is some other state of affairs, *q*, that S also desires and ii) S believes that if *p* were to obtain, then this would cause *q* to obtain.

A non-instrumental desire, then, is a just a desire that is not instrumental. That is, one has a non-instrumental desire for some state of affairs, *p*, when one desires *p* for its own sake, and not as the means to something else one desires. So if I desire happiness for its own sake, and not as a means to something else, then this desire is non-instrumental. The desire satisfactionist would do well to take it that it is only the satisfaction of non-instrumental desires that can enhance a person's welfare.[7]

[7] To see why, consider Jamie. She is raised to believe that people ought to be good citizens and contribute to their communities. So early on, Jamie develops a deep desire to lead a life in which she has a positive impact on her community. Now suppose that she comes to believe that drinking milk is good for children, and so she campaigns to get all the school cafeterias in her community to serve milk with lunch. It's not that she cares about drinking milk for its own sake. Rather, she wants to get the schools to serve milk at lunch because of the good effects she believes this to have. Thus Jamie's desire to get the schools to serve milk at lunch is an instrumental one. Now, compare two possible scenarios for Jamie. In the first scenario, Jamie succeeds in her campaign, but it turns out that getting kids to drink milk with lunch doesn't really have any good consequences. The kids all have just as healthy bones and teeth (or what have you) as they had before. Thus Jamie's non-instrumental desire to make a difference to her community remains unsatisfied. In the second scenario, Jamie does not get the schools to serve milk at lunch (but even if she had, it wouldn't have mattered because, like the first scenario, we're supposing that getting kids to drink milk doesn't make them better off).

Concurrent Desire Satisfaction and Frustration. In the last chapter we saw that desire satisfactionist theories encounter vexing problems because of the connections between desire and time. There were issues concerning future directed and past directed desires, the problem of desire change, and several others besides. I argued that the best way for the desire satisfactionist to deal with these problems was to adopt the view I called *Weak Concurrentism*. To move things along, let's assume I was right.

This means formulating our preliminary version of ADS in such a way that it appeals to the notion of concurrent desire satisfaction and frustration. You have a concurrently satisfied desire if you desire, at some time, *t*, that some state of affairs, *p*, obtain, and *p* does obtain at *t*. By contrast, you have a concurrently frustrated desire if you desire, at *t*, that *p* obtain, and *p* does not obtain at *t*. In stating the view, it will be useful to have the notion of an *episode* of concurrent desire satisfaction or frustration.

> S gets an episode of *concurrent desire satisfaction* during the whole interval <t_1, t_2> iff for every time, *t'*, that occurs during <t_1, t_2>, it is true both that i) S has, at *t'*, a desire that some state of affairs, *p*, obtain, and ii) *p* in fact obtains at *t*'.
>
> S gets an episode of *concurrent desire frustration* during the whole interval <t_1, t_2> iff for every time, *t'*, that occurs during <t_1, t_2>, it is true both that i) S has, at *t'*, a desire that some state of affairs, *p*, obtain, and ii) *p* does *not* obtain at *t*'.

Notice that this definition applies even in cases when the object of the desire is that some state of affairs obtain in the future, in the past or at no time in particular.

Now, at last, we are in a position to state our simple version of ADS:

<u>Classical ADS</u>

(i) Every concurrent non-instrumental desire satisfaction is intrinsically good for its subject; every concurrent non-instrumental desire frustration is intrinsically bad for its subject.

(ii) The intrinsic value for S of an episode of concurrent non-instrumental desire satisfaction equals the duration of the episode times the average strength of the relevant satisfied desire; the intrinsic disvalue for S of an episode of

In which of these two scenarios does Jamie have higher welfare? If the desire satisfactionist allows that the satisfaction of both instrumental and non-instrumental desires can enhance welfare, then Jamie would be better off in the first scenario. After all, according to Simple ADS, the satisfaction of *any* desire has a positive impact on a person's welfare. But this result is counter-intuitive, I claim. Jamie seems to be equally well off (or poorly, as the case maybe) in both scenarios. After all, Jamie wants to get the schools to serve milk at lunch only because she (mistakenly) thinks this is the means to satisfying her non-instrumental desire to do something good for her community. And in both scenarios, Jamie's non-instrumental desire remains unsatisfied. In both cases, she fails to obtain what she *really* cared about. Accordingly, the lesson to draw from this case seems to be that the satisfaction of instrumental desires does not in and of itself enhance a person's welfare.

concurrent non-instrumental desire frustration equals the duration of the episode times the average strength of the relevant frustrated desire.

(iii) The total amount of welfare contained in a person's life equals the sum of the intrinsic values of all the episodes of concurrent non-instrumental desire satisfaction in the life minus the sum of the intrinsic disvalues of all the episodes of concurrent non-instrumental desire frustration contained in the life.

Thus, according to Classical ADS, the total amount of welfare contained in a person's life equals the total amount of net concurrent non-instrumental desire satisfaction in that person's life.

6.2 Three Problems for Classical ADS

The search for a better version of ADS is motivated by three technical problems. (The view also faces some serious substantive problems, but I will discuss them in chapter 7. They threaten even the best formulation of ADS.)

6.2.1 The Problem of Irrelevant Desires

One difficult problem for Classical ADS has to do with desires whose satisfaction or frustration seem to be irrelevant to the welfare of the person who holds these desires. Parfit's well-known 'stranger on the train case' provides a good example of the problem:

> Suppose that I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I strongly want this stranger to be cured. We never meet again. Later, unknown to me, this stranger is cured. On the Unrestricted Desire-Fulfilment Theory, this event is good for me, and makes my life go better. This is not plausible. We should reject this theory. (Parfit, 1984, p. 494)

Shelly Kagan offers another good example:

> Suppose, then, that I am a large fan of prime numbers, and so I hope and desire that the total number of atoms in the universe is prime. Imagine, furthermore, that the total number of atoms in the universe is, in point of fact, prime. Since this desire is satisfied, the preference theory must say that I am better off for it (…). But this is absurd! The number of atoms in the universe has nothing at all to do with the quality of my life. (Kagan, 1998, p. 37)

So the argument here is that intuitively, one's welfare would not be enhanced by the satisfaction of desires for a state of affairs like that the stranger on the train gets cured or that the number of atoms in the universe is prime. But this intuition conflicts with the implications that Classical ADS has about these cases. So Classical ADS has to go.[8]

One way to deal with this problem would be to build an experience requirement into the desire satisfaction theory. As Sumner explains the move, 'Such a condition would stipulate that a state of affairs can make me better off only if, in one way or another, it enters or affects my experience.' (Sumner, 1996, p. 127) Thus the idea is to formulate the theory in such a way that what enhances one's welfare is the complex state consisting of a) one desiring a given state of affairs, b) this state of affairs obtaining, and c) one's knowing that this state of affairs obtains.[9]

However, I do not think that such an experience requirement would be a good way for the desire satisfactionist to deal with the problem of irrelevant desires. For this strategy conflicts with certain intuitions that it seems the desire satisfactionist should want to accept. Perhaps most importantly, it makes posthumous benefits or harms impossible. Suppose a person strongly desires that her fortune should be used to create a charitable foundation upon her death. She inserts instructions to this end in her will, but when she dies her instructions are not carried out; the charitable foundation is not created. In such a case, it seems that the total degree of conformity between what she wants to happen and what does in fact happen is *lower* than it would have been if the instructions in her will had been carried out. Thus it seems that the clear-headed desire satisfactionist should accept that this person's life goes less well for her than it would have if her will had been carried out after her death. But the desire satisfaction theory cannot yield this result if an experience requirement is built into it. After all, after one is dead there is nothing that can 'enter into one's experience.' Since it is impossible for this person, once dead, to know that her desires do not get satisfied, it is impossible for her to be harmed by their frustration, according to the desire satisfaction theory with an experience requirement built in.

[8] Many others have discussed this problem as well. Heathwood has a good discussion of it (cf. SDS, ms, pp. 13-14). Also see Griffin, 1986, p. 21, Sumner, 1996, pp. 125-127, and Carson, 2000, pp. 74-76.

[9] Heathwood's theory, Subjective Desire Satisfactionism, builds in the experience requirement in a different way. On his view, what enhances welfare is the complex state of desiring some state of affairs and believing that it obtains – irrespective of whether it in fact obtains or not. However, his theory is problematic for other reasons. Since his view is a mental state theory, it is undermined by Experience Machine objections.

A more promising route for the desire satisfactionist to pursue, in my view, is to state the theory in such a way that it's not the satisfaction or frustration of just any desire that affects welfare, but rather so that it is only success or failure with respect to one's goals, aims or projects that affects welfare. Though this idea has been suggested in other places,[10] it seems to be pursued in most detail by Simon Keller in a recent article, where he distinguishes between goals and mere desires. Goals, he explains as follows:

> Successfully achieving a goal involves having the goal be attained partly as a result of your own efforts. That is why you cannot take as goals things you know you cannot influence. I cannot make it my goal that Geelong [a sports team] wins the Premiership, no matter how much I hope that that happens, because I know that nothing I do will make a difference either way.' (Keller, forthcoming, ms, p. 27)

By contrast, mere desires are

> desires that do not count as goals. I might desire a Geelong victory, but that does not mean that I take it as a goal. (…) Desiring that the world be a certain way (…) does not necessarily involve any commitment to making the world accord with your desire. It can make perfect sense to say, "I desire that this happens, but I am not setting out to make it happen." (Keller, forthcoming, ms p. 29)

Having sketched a distinction between goals and mere desires, Keller proceeds to claim that what affects your welfare is succeeding or failing with respect to your goals, not just getting your mere desires satisfied or frustrated.

One of the advantages that Keller says his account of welfare has is that it provides a solution to the problem of irrelevant desires. As he puts it,

> the account has the ability to avoid of the main problems for the desire theory: the problem of desires for things that do not, intuitively, have anything to do with our welfare. The reason why it is not in your best interests that the stranger [in Parfit's case] recover is that you merely desire his recovery. Whether or not he recovers has nothing to do with how things go for you, because you do not dedicate anything of yourself to his recovery. It would be different if you took the stranger's recovery as your *goal*. Then, you would put some effort into his recovery, and whether or not he recovers would reflect something about whether that effort of yours pays off, hence whether you yourself are successful, hence whether things go well for you. (Keller, forthcoming, ms, p. 31)

In my view, the strategy that Keller pursues for dealing with the problem of irrelevant desires is promising. There is indeed some plausibility to the idea that what matters to our welfare, in particular, are the goals that we put effort into attaining. This idea of Keller's gives us a neat way to distinguish between the desires that matter to welfare and those that don't.

However, his idea also faces problems. It seems that the desire satisfactionist should allow that a person can be benefited by other things in addition to just the achievement of

[10] For instance, see Griffin, 1986, p. 21

one's goals *through one's own efforts*. Sometimes we attain our goals but not because of our efforts. Sometimes we just get lucky. Griffin puts the point as follows:

> It is not that (…) desires count only if they become the sort of aims or goals or aspirations on which the success of a life turns. Good things can just happen; manna from heaven counts too. So we should try saying (…) that what count are what we aim at and what we would not avoid or be indifferent to getting. (Griffin, 1986, p. 22)

Keller's strategy needs to be expanded in such a way that it is not just achievement of one's goals through one's own efforts that enhances welfare, but also what I take Griffin to mean by 'manna from heaven' – viz. attaining one's goals due to events outside of one's control.

In fact, Keller, in an earlier article, mentions a good illustration of this very point.[11] Suppose that one of my goals in life is to become wealthy. I work hard to become wealthy, but have only moderate success. Then a rich uncle of mine, whose existence I had no prior knowledge of, dies. I am his sole living relative and I inherit his fortune. My goal is attained, even though it is not through my own efforts. Is my welfare thereby enhanced? Keller thinks not, on the grounds that 'it's an individual's achieving her own goals – meaning her attaining them *through, in part, her own efforts* – that contributes to her welfare.'(Keller, 2004, p. 33) However, this seems counter-intuitive to me. In the inheritance case, the degree to which my life goes the way I want it to go is higher because I inherited the money than it would have been if I had not gotten the money. Thus any desire satisfactionist, it seems to me, should allow that my inheriting this money enhances my welfare somewhat.[12]

Thus if Keller's idea for how to avoid the problem of irrelevant desires – viz. by formulating desire satisfactionism in terms of goals, not mere desires – is to succeed, it must be developed in a way that accommodates this point. That is, it must be developed in such a way as to allow that also 'manna from heaven' can enhance one's welfare. In section 6.3, I will formulate a version of ADS that does just this.

[11] Cf. Keller, 2004, p. 33

[12] Perhaps a desire satisfactionist sympathetic to Keller's view could say that my life would have gone *even* better for me if I had gotten rich through my own efforts. But that does not change the overall fact that I am at least somewhat benefited by getting rich through inheritance as opposed to through my own efforts.

6.2.2 Redundant desires and double-counting

A second problem for Classical ADS is that it sanctions some dangerous forms of double-counting. There are several different sorts of desires that cause problems of this sort. First, one might desire a certain state of affairs that is *part* of a bigger state of affairs that one also desires. Second, one might have certain desires that seem to follow trivially from other desires one has.

To see the problem posed by the first kind of desire, suppose I have a non-instrumental desire to drink a whole cup of coffee. Call this 'D1.' Moreover, suppose that this desire leads me to have another desire, namely the desire to take a sip from this cup next to me that has coffee in it. This latter desire – call it 'D2' – is not an instrumental desire. For it is not the case that I desire to take a sip from this cup because I believe it is the *means* to satisfying any other desire of mine. In other words, the reason I have D2 is not that I believe its satisfaction would *cause* the satisfaction of D1. Rather, I have D2 because I believe that the satisfaction of D2 would partly *constitute* the satisfaction of D1. After all, taking a sip from this cup is what my drinking a whole cup of coffee would in part *consist in*. Thus since both D1 and D2 are non-instrumental desires, Classical ADS implies that the satisfaction of each would enhance my welfare. However, this seems wrong. D2 is a derivative desire. It is one that I have purely in virtue of some other, more fundamental desire of mine. Thus to allow the satisfaction of D2 as well as D1 to count towards my welfare would seem to lead to a dangerous form of double-counting. So Classical ADS is in trouble.

When it comes to the second kind of desire that leads to double-counting worries, consider two people, A and B. Each has as his only project to write a dissertation in philosophy. However, A is a less excitable and effusive type than B is. A has just one single desire: he desires to complete a dissertation in philosophy by August 2009. What's more, the object of this desire obtains: he will in fact complete the dissertation by August 2009. B's life, on the other hand, is identical in every respect to A's life, except that B has many more non-instrumental desires than A does. For one thing, B desires – just like A – to complete a dissertation in philosophy by August 2009. But in addition, B also has certain desires that follow trivially from this first desire. Thus B desires to complete a dissertation in philosophy *some day*, to complete it *before he dies*, to complete it *before* *he turns 30*, to complete it *in a reasonable amount of time*, and so on. Just like A, B too will complete his dissertation in philosophy by August 2009. Thus all of B's desires count as satisfied. Since B has so many more non-instrumental desires that are satisfied than A does, Classical ADS implies that B's welfare is immensely higher than A's welfare. However, this seems odd. Intuitively, things seem to be going just as well for A as they are for B. The only difference between them is that B has a whole bunch of redundant desires that A happens to lack. This difference, however, does not seem to be the sort of difference that can ground such a dramatic difference in welfare. Thus Classical ADS seems to sanction a dangerous form of double-counting in this sort of case, as well.

To put the worry somewhat differently, note that these cases illustrate that on Classical ADS, there are some extremely easy ways to improve the quality of your life. In particular, for any satisfied desire, D, that you have, either a) get yourself to desire the various states of affairs that partially constitute the object of D, or b) get yourself to desire various states of affairs that obtain trivially in virtue of the fact that D is satisfied. However, it seems pretty clear that no plausible theory of welfare should allow one to enhance one's welfare simply by forming desires of these kinds. So we need to formulate ADS in a way that does not have this consequence. The version of ADS that I formulate in the next section has various features that help avoid this problem.

6.2.3 Partially Satisfied Desires

A third problem for Classical ADS is that sometimes our desires seem to be only partially satisfied. For instance, consider the following state of affairs:

(S_{diss}) Alex completes his dissertation in philosophy.

Right now, I have a strong desire that S_{diss} obtain. However, it doesn't obtain right now. For I have not yet completed my dissertation. Nonetheless, I have completed a good portion of it: let's say two thirds. Classical ADS is formulated in such a way that it allows of only two possibilities: that a desire is either concurrently satisfied or concurrently frustrated. If these are the only two options, the desire for S_{diss} that I currently hold must be regarded as frustrated. Thus according to Classical ADS, if I were to die tomorrow, all the work that I have done on my dissertation – all the steps I took towards the completion of my project – would do nothing to enhance my welfare. In fact, the frustration of this desire *diminishes* my welfare. But this seems implausible. The intuitive thing to say is that if I were to die tomorrow, I would have been benefited at least *somewhat* by the fact that I have completed two thirds of my dissertation, while desiring to complete my dissertation. Perhaps my welfare is not as high as it would have been if I had fully completed my dissertation before I died. But my welfare is surely higher than it would have been in the scenario where I die tomorrow having completed absolutely nothing of my dissertation. Thus the implications of Classical ADS about cases of partially satisfied desires (or partially completed projects) seem implausible.

A natural solution to this sort of problem might be to modify Classical ADS so that it allows desires to be partially satisfied or partially frustrated. Brad Skow has advocated such a modification to ADS for other reasons (viz. that it might allow for a solution to Feldman's paradox of desire).[13] His suggestion is that we modify ADS so as to state that:

> the intrinsic value of an episode of desire satisfaction is equal to the intensity of the desire times the duration of the episode times the degree to which the desire is satisfied. (Since I use negative levels of satisfaction to represent desire frustration, the intrinsic value of an episode of desire frustration is equal to intensity x duration x (-1) x degree of satisfaction. …) (Skow, ms, p. 8)

But how are we to understand the notion of degrees of desire satisfaction in the first place? Skow suggests a helpful simplifying assumption: namely that for every desire, there is a maximum degree to which it can be satisfied and a maximum degree to which it can be frustrated.[14] Thus we might assume that for every desire, there is some degree to which it is either satisfied or frustrated, and this degree can be represented by some number in the interval <1, –1>. The number 1 would correspond to the maximal degree of satisfaction, while –1 would correspond to the maximum degree of frustration.

There are some problems with Skow's suggestion, however. First, one might question the assumption that there is a maximum degree to which a desire can be satisfied. Suppose I desire that the following state of affairs obtain:

S_{money}: Alex is wealthy.

For such a desire as this, it might not make sense to say that there is a maximum degree to which it can be satisfied. For intuitively, the more wealthy I become, the more my desire for S_{money} gets satisfied. This makes my desire for S_{money} different from my desire

[13] Cf. Skow, ms,

[14] Cf. Skow, ms, p. 7

for S_{diss}, since it would indeed make sense to say that my desire for S_{diss} can be maximally satisfied.

The second problem is even more worrying. Consider my desire for S_{diss}. Suppose I die before completing a single page of my dissertation. In this case, is my desire satisfied to degree 0 or is it frustrated to the maximal degree –1? It is not clear what score on the satisfaction-frustration scale my desire for S_{diss} would get in this case. After all, what could it be for this desire to be frustrated except for it to *not* be satisfied?

Of course, for some desires, it really does make sense to distinguish between this desire's failing to be satisfied (i.e. its being satisfied to degree zero) and its being positively frustrated (i.e. its getting a negative satisfaction score). For instance, suppose I desire:

($S_{good\text{-}boy}$) Alex's son is a morally good person as an adult.

Here it seems to make sense to distinguish between this desire's failing to be satisfied and its being frustrated. If my son ends up being an evil person, then the desire is frustrated. But if he just turns out to be morally neutral – i.e. neither a good person nor an evil person – then the desire simply fails to be satisfied. Another way in which this desire might simply fail to be satisfied (as opposed to positively frustrated) would be if my son were to die before he reaches adulthood. So for desires of this sort, it seems to make sense to assume that failure to be satisfied can be represented by the number 0, while maximum frustration can be represented by the number –1.

Nonetheless, we just saw that this might not hold when it comes to desires for states of affairs like S_{diss}. It is unclear how to understand what it would be for my desire for S_{diss} to be maximally frustrated. If I have not written a single page on my dissertation, is my desire for S_{diss} satisfied to degree zero or maximally frustrated (i.e. satisfied to degree –1)? It is unclear. Thus it is unclear how Skow's proposal concerning degrees of desire satisfaction is to be implemented.

I will deal with these problem in the following way. I will not adopt Skow's idea of degrees of desire satisfaction. Instead, I will offer an account of partial success with respect to a project at a time, and this account proceeds in terms of regular, all-or-nothing satisfaction or frustration of desires. In particular, the general idea is that, if one of your central projects in life is to write a dissertation in philosophy, then we can say that you are successful with respect to this project right now to the extent that your dissertation-related desires are satisfied right now. Roughly put, if more of your dissertation-related desires are satisfied than frustrated right now, then right now you are successful to a positive degree with respect to your project of writing a dissertation. And if more of your dissertation-related desires are frustrated than satisfied right now, then right now you are successful to a negative degree with respect to this project. This proposal will be developed in detail below.

6.3 Cloud Desire Satisfactionism

In this section, I present a refined version of ADS that solves the three technical problems discussed in the previous section. I will call this version of the theory 'Cloud Desire Satisfactionism.'[15] To solve the problem of irrelevant desires, this theory will draw on Keller's insight about the tight connection between one's goals and one's welfare. Moreover, I will provide an account of degrees of success with respect to projects in order to accommodate the problems concerning desires that seem to be partially satisfied. What's more, a certain element in this theory (viz. the notion of a basic desire) counters the threat of double-counting. The theory is still a work in progress. I present it in as much detail as I'm able to at present. However, I acknowledge that more work may need to be done to develop certain details of the theory. Nonetheless, I think the theory represents a promising avenue for the desire satisfactionist to pursue.

6.3.1 Stating Cloud Desire Satisfactionism

This new version of the theory begins from the recognition that our desires are much more numerous and fine-grained than traditional desire satisfaction theories seem to recognize. Suppose the project that is most important to me right now is to finish my dissertation. A traditional desire satisfactionist might be inclined to say that in this case I have just one desire, namely a desire to finish my dissertation. However, this would not be a realistic description of the case. A real person who possesses the goal of writing a

[15] The theory developed in this section is based on some ideas that Fred Feldman suggested to me in conversation. I am very grateful for his help in designing this theory.

dissertation is most likely going to have a whole range of desires related to the project of writing a dissertation. For instance, the person might desire to finish his dissertation some day, to be making good progress on his dissertation right now, to manage to write ten pages today, to be the sort of person who is able to finish his dissertation on time, and so on. Thus instead of taking it that what matters to welfare are very course-grained desires like the desire to write a dissertation, it is more realistic to think that what matters to welfare are entire clouds of very fine-grained desires whose objects are somehow related to one another. This is the basic thought underlying Cloud Desire Satisfactionism. To present the theory precisely, I will need to introduce some new concepts.

Desire Clouds. For starters, we need the concept of a cloud of desires. Very roughly, a desire cloud is a set of desires that are 'about the same thing,' or whose objects are related in some important way. But what makes a collection of desires be 'about the same thing' or be related in the relevant way? It is hard to say, precisely. So as a preliminary account, I suggest the following:

> A given set, C, of S's desires constitutes a **desire cloud** iff a) every desire in C is a non-instrumental desire, and b) every desire in C is such that there is some salient (non-trivial) way in which its object is related to the objects of all the other desires in C.[16]

Admittedly, this is not precise. When are the objects of a given set of his desires are related in a 'salient' way? I am not sure. I suspect that David Lewis' notion of relevance (which proceeds in terms of the 'subject matters' of various propositions) would be of some help here.[17] However, it is beyond the scope of this chapter to try to give a fully worked out account of two desires being 'about the same thing'. For now, I will simply help myself to this notion.

Nonetheless, I think I can offer at least the following *sufficient* condition for a given set of desires to be related in the intended way: a given set of S's desires are related in a salient way if there is some unified goal, end or ideal scenario such that the satisfaction of any of these desires would in part constitute the realization of that goal, end or ideal scenario. Thus suppose I desire to complete my dissertation on time, to be making good

[16] This account allows that one and the same desire may be part of several desire clouds. In other words, desire clouds may overlap. However, given the way the rest of the theory is formulated, I do not think this implication is problematic. (Thanks to Fred Feldman for pointing this out to me.)

[17] Cf. Lewis, 1998, ch. 8.

progress on my dissertation right now, to write ten pages of my dissertation today, to be the sort of person who is able to finish his dissertation on time, and so on. These desires will be part of the same desire cloud. For there is a unified ideal – my completing a dissertation – that the satisfaction of these desires would in part constitute the realization of.

Notice that the notion of a desire cloud is intended to be quite inclusive. It includes more than just aims, goals, projects and the like.[18] Consider, for instance, Jimmy Fallon's character in the movie 'Fever Pitch' (2005). He is an ardent Boston Red Sox fan and he deeply desires that they win the World Series each season. His being a Boston Red Sox fan is an important component of who he is and his concern for the success of the Red Sox guides his behavior in important ways. Thus there is a desire cloud comprising those of his desires that pertain to the success of the Red Sox. We might suppose that his Red Sox desire cloud contains, among other things, a desire for the following: that the Red Sox to win the World Series every season, that the Red Sox win the World Series this season, that the Red Sox win this game, that he be celebrating the Red Sox' victory in this season's World Series, that he be a fan of the best team in baseball, and so on. These all seem to be non-instrumental desires. Their objects are all related to the success of the Red Sox. Most people would agree that these desires are naturally part of the same group. I think this is good evidence for thinking that we are dealing with a desire cloud.

Projects. There is one species of desire clouds that will be particularly important for our purposes, namely projects. A project of yours is identical to a cloud of desires that you seem to be in large part capable of satisfying through your own efforts. To be more clear, we can put it like this:

> A given desire cloud, C, of S's is a **project** of S's iff according to the evidence available to S, a preponderance of the desires in C are such that there are things S can do that would be likely to result in the satisfaction of these desires.

Thus a project is a desire cloud of yours where it's reasonable to think that most of the desires in this cloud can be satisfied through your own efforts.[19] So while Jimmy Fallon's

[18] Thus a desire cloud can include 'mere desires' in Keller's sense.

[19] This definition admits of vagueness because of its use of terms like 'most', but I see no plausible way to avoid this. It would be implausible to require that *every* desire in a desire cloud, C, must seem to be satisfiable through one's own efforts in order for C to count as a project. But it is not clear to me precisely where the cut-off should be drawn.

character has a desire cloud pertaining to the success of the Boston Red Sox, this desire cloud does not count as a *project* of his. After all, his evidence suggests that there is nothing he can do that would lead to the satisfaction of a large portion of the desires in this cloud. For no matter how loyally he attends the Red Sox' games, how loudly he cheers, or how hard he prays, it won't affect the Red Sox' performance one bit. Thus the desire cloud in question here does not count as a project. Of course, in order for some desire cloud to be a project, it doesn't have to be the case that the desires in the cloud can be satisfied *only* through your own efforts. Suppose you have a desire cloud pertaining to your being happy. You desire to be as happy as you can be, to be happier than you were last week, to be feeling happy right now, to be happier than most people, etc. There are all sorts of ways in which these desires could become satisfied due to events outside your control. But since many of the desires in this cloud are at least *capable* of being satisfied by things you do, this desire cloud still counts as a project.[20]

This distinction between desire clouds that count as projects and those that don't is going to provide a plausible solution to the problem of irrelevant desires. Cloud Desire Satisfactionism will take it that one's welfare is enhanced *only* by success (in a sense soon to be defined) that one has with respect to one's projects – i.e. success with respect those desire clouds over which one seems to have a reasonable degree of control. While you may have a desire cloud that pertains to just about anything – e.g. the success of the Boston Red Sox, the number of atoms in the universe, the recovery of the stranger you met on the train, etc. – your welfare can be impacted only by the satisfaction or frustration of desires in a cloud that counts as a project. So if you desire, for example, that the number of atoms in the universe is prime and this turns out to be the case, then your welfare will not thereby be enhanced. After all, this desire does not belong to a desire cloud over which you have a significant degree of control, i.e. a cloud that counts as a project. This provides a solution to the problem of irrelevant desires. For we can now

[20] As Scott Hill has pointed out to me, this account of a project implies that if you are locked up in a room, so that there are no actions you are able to perform any more, then desire clouds which before your incarceration counted as projects might all of a sudden cease to count as projects. Upon your release, however, these desire clouds would regain their status as projects provided you regain your ability to do things that would advance the satisfaction of the desires in this cloud. I am willing to accept these implications of my definition of 'a project'.

distinguish between the desires that are relevant to welfare and those that are not: a desire is relevant to one's welfare iff it belongs to a cloud that counts as a project.

Moreover, the present suggestion allows us to avoid the problem that undermined Keller's suggestion. Recall that on Keller's view, what enhances welfare is achieving your goals *through your own efforts*. But this fails to allow that you can be benefited by unforeseen windfalls, i.e. by what Griffin called 'manna from heaven.' By contrast, my suggestion is not undermined by this problem. For on my view, what enhances your welfare is (roughly) to succeed in satisfying desires that are part of a desire cloud that counts as a project – no matter whether the desires get satisfied through your own efforts or not. If you have some desire cloud that counts as a project, and the desires in this project end up getting satisfied through some lucky event that you had no hand in causing, then this may still enhance your welfare.[21] Thus it is possible for you to be benefited by 'manna from heaven.'

Projects' contributions to welfare. According to Cloud Desire Satisfactionism, your level of welfare at a given time, *t*, is going to be the sum of the welfare-contributions made by the various projects you have at *t*. The contribution that a given project of yours, P, makes to your welfare at *t* is the product of two things: a) how successful you are with respect to P at *t*, and b) how important P is to you at *t*. To be more precise:

> The **contribution** to S's welfare made by project P at *t* = [S's level of success with respect to P at *t*] x [the overall importance of P for S at *t*].

To be able to determine how much a given project of yours contributes to your welfare at a given time, we obviously are going to need accounts of both *success with respect to a project* and the *importance of a project*. I will take them in this order. My account of success enables Cloud Desire Satisfactionism to avoid the problem of partial desire satisfaction from section 6.2.3, while my account of importance helps the theory to avoid problems concerning double-counting.

Degrees of success with respect to a project. For every project, P, that a person, S, has at a time *t*, there is some degree to which S is successful or unsuccessful with respect

[21] Perhaps it would be a good idea to modify the theory so that, all other things being equal, it is *better* for you to accomplish your projects through your own efforts than in some other way that is not through your own efforts. However, this introduces an extra level of complexity into the theory (which is already quite complex). So I will pass over this modification for the time being.

to P at *t*. I am going to assume – following Skow's proposal discussed above – that S's level of success is representable by a number between +1 and -1. The number –1 would represent complete failure with respect to the project in question and +1 would represent complete success with respect to that project. As a first stab at understanding levels of success, we might say that your level of success with respect to P at *t* equals the proportion of the desires in the P-cloud that are satisfied at *t* minus the proportion of the desires in the P-cloud that are frustrated at *t*. Thus if three-quarters of the desires in the P-cloud are satisfied at *t* and one quarter of them are frustrated, your level of success with respect to P at *t* = 0.75 – 0.25 = 0.5.

However, this first-pass account of success needs to be refined. After all, different desires in a given desire cloud will have different strengths. Consider a project, P, that consists of ten desires of strength +1000 and a thousand desires of strength +0.5. Suppose that all ten desires of strength +1000 are satisfied at *t*, while all of the thousand desires of strength +0.5 are frustrated at *t*. The first-pass account of success would imply that your level of success with respect to this project is extremely low. After all, the P-cloud contains many, many more frustrated desires than satisfied desires. But, intuitively, this result is incorrect. Since the satisfied desires are so much more weighty than the frustrated desires are, it seems that you are in fact quite successful with respect to this project at *t*. After all, the most significant desires in cloud the are all satisfied, while it's only the negligible ones that are frustrated. Thus a plausible account of success with respect to a project must be sensitive to the fact that the project may consist of desires of different strengths.

Here, then, is my official account of success with respect to a project. It is indeed sensitive to the fact that the desires in a given cloud may have different strengths. Letting 'the P-cloud' refer to the cloud of desires that constitute a given project, P, here is the procedure for how to determine S's level of success with respect to P at *t*:

1) Add up the strengths of all S's desires that are in the P-cloud at *t*.[22] Let '*T*' stand for this number.

[22] Really, to fully avoid double-counting, step 1) should not take into consideration the *duplicate* desires in the desire cloud in question. Thus it would be better if step 1) said this: 'Find all the pairs of duplicate desires in the P-cloud at *t*, and then for each such pair, remove one of them. Then add up the strengths of all of the remaining desires in the P-cloud at *t*. Let "T" stand for this number.' I introduce the concept of a duplicate desire below. For more on this, see the second to last footnote of the paper.

2) Add up the strengths of all S's desires in the P-cloud at *t* that are satisfied at *t* (i.e. whose objects are true at *t*). Let '*S*' stand for this number.
3) Add up the strengths of all S's desires in the P-cloud at *t* that are frustrated at *t* (i.e. whose objects are not true at *t*). Let '*F*' stand for this number.
4) S's level of success with respect to P at *t* equals the proportion of *T* that comes from satisfied desires minus the proportion of *T* that comes from frustrated desires. More precisely, **S's success with respect to P at *t* = (*S/T*) – (*F/T*).**

On this account, S's success with respect to P at *t* will always be a number between –1 and +1. After all, both terms in this equation, (*S/T*) and (*F/T*), are always going to be numbers between 0 and 1. Thus we have an account of degrees of success with respect to a project that is consistent with the assumption of Skow's discussed in section 6.2.3.

What's more, this account is not threatened by the problems that beset Skow's idea that desires themselves can be partially satisfied or frustrated. For instance, we saw above that it was unclear how to determine the degree to which my desire to write a dissertation would be frustrated in a scenario where I had written zero pages. For we were unable to decide whether my desire should be taken to be satisfied to degree zero or to degree –1. This confusion is eliminated once we move to talking in terms of desire clouds, however. For at any given time, my cloud of dissertation-related desires will contain a determinate number of desires (all non-instrumental), and the ratio of satisfied desires to frustrated desires (together with their strengths) will determine my level of success with respect to this project at the time in question. So if one of my projects is to write a dissertation, but I have written zero pages because, say, I am still in the early stages of planning the dissertation, then presumably my dissertation-related desire-cloud would not contain, e.g., the desire to have written lots of pages or the desire to be almost done, etc. Thus I would not count as a failure with respect to the dissertation-project. But if I have tried very hard for years to write a dissertation and still have written zero pages, then my dissertation-related desire cloud would presumably contain all sorts of desires that end up being frustrated (like the desire to have written lots of pages and the desire to be almost done). Accordingly, in this case I would indeed be a failure with respect to the dissertation-project. So it is fairly clear, on Cloud Desire Satisfactionism, how to understand degrees of success or failure with respect to a project.

Importance of a project for you. Now that we have a plausible account of success with respect to a project, the final ingredient we need is a plausible account of the importance of a project. In general, the importance of project P for you is supposed to reflect the strength of your preference for P *as a whole*. Thus one natural way to understand the importance of a project for you might be this: it just equals the sum of the strengths of all the desires in the cloud associated with that project. The problem with this suggestion, however, is that it would make Cloud Desire Satisfactionism vulnerable to double-counting problems. After all, a given desire cloud might include both a) desires for states of affairs that partially constitute some bigger states of affairs that one also desires, and b) desires that seem to follow trivially from other desires one has. If the strengths of these desires are allowed to count towards the total importance of a project for you, then double-counting will ensue. (More on this in section 6.3.2, and the second to last footnote of the paper.)

Thus the importance of a project for a person should not be taken to be the sum of the strengths of simply *all* the desires in the cloud associated with that project. A more refined account is needed to avoid problems of double-counting. The account of importance that I favor, then, is this:

> The **overall importance** of project P for S at *t* = the sum of the strengths of all the *basic desires* contained in the P-cloud at *t*.

What is meant by a 'basic desire'? Recall that desire clouds contain only non-instrumental desires. The basic desires are a particular sub-class of the non-instrumental desires. To precisely explain what I mean by a basic desire, I need to introduce two other concepts. First of all:

> One of S's desires, D, in a given cloud C is **derivative** =df. there is some other desire (or desires) also in C such that S has D purely *in virtue of* having that other desire (or those other desires).

There are several ways for a desire to be derivative. For one thing, instrumental desires will be derivative. If you desire something because you think it is the means to getting something else you want, then you have the former desire purely in virtue of having the latter. Thus the former desire will be derivative. Second, consider a desire whose object is *a part* of the object of some other desire you have. For instance, suppose I desire to drink a whole cup of coffee, and I also have a desire to take a sip from this cup in front of me with coffee in it. This latter desire is not an instrumental one, since taking a sip is not *the* *means* to drinking a whole cup of coffee; rather, it is part of what drinking a whole cup of coffee consists in. Still, my desire to take a sip is a derivative desire because I have it *purely in virtue* of my desire to drink a cup of coffee. Third, there are certain desires of mine that might trivially follow from other desires of mine. For example, suppose I desire to finish my dissertation by August 2009. Because of this, it may be the case that I also desire to finish my dissertation by August 2010. This latter desire would also count as derivative, because I have it purely in virtue of another desire of mine. Finally, suppose I have a given desire, D, but there is no other *single* desire in virtue of which I have D. Still, suppose I have D in virtue of having several other of desires: D', D'' and D'''. In such a case, D would count as derivative as well.[23]

These were some examples of desires that were derivative, for one reason or another. By contrast, here's an example of a pair of desires neither of which are derivative. Suppose that one of my projects is to write a dissertation, and the desire cloud pertaining to dissertation-writing contains just two items:

D1- I desire that I will complete my dissertation by August 2009.
D2- I desire that my dissertation will be original and influential.

It's not the case that I have D1 purely in virtue of having D2 and it's not the case that I have D2 purely in virtue of having D1. Nor is there any other desire in this cloud in virtue of which I have D1 or D2. Thus neither D1 nor D2 is derivative.

Because of a certain complication, however, we can't simply say that a basic desire is a non-derivative desire. Here's why. Suppose I have the following pair of desires:

D3- I desire to complete my dissertation someday.
D4- I desire to complete my dissertation eventually.

Moreover, suppose these are the *only* two desires in the relevant desire cloud. In this case, neither D3 nor D4 are derivative. After all, there is no other desire in the cloud in virtue of which I have D3, nor is there any other desire in the cloud in virtue of which I have D4. I don't have D3 in virtue of D4 and I don't have D4 in virtue of D3 (and D3 and D4

[23] I haven't tried to give a precise account of what it is for one to have a given desire 'in virtue of' having some other desire. However, there is a sizable body of literature on what it is for one material object to exist in virtue of another, as well as what it is for an object to possess one property in virtue of possessing some other property. Thus I take it that the accounts of the 'in virtue of' relation that turn out to be most plausible with respect to these topics will show us the best way to understand the notion of having one desire in virtue of having some other desire. (See for instance Schaffer, *forthcoming*, Rosen, *ms* and Trogdon 2009)

are the only desires in the cloud). Nonetheless, it would seem to be wrong to let both of these desires count towards my welfare. For they clearly are, in some sense, the 'same desire.' To handle this complication, I'm going to introduce the notion of *duplicate desires*:

> Two desires of mine, D and D', are **duplicates** iff the conditions under which D would be satisfied are *identical* to the conditions under which D' would be satisfied.[24]

Accordingly, D3 and D4 here would be duplicates. For D3 would be satisfied iff D4 is also satisfied.

Given the notions of derivative desires and duplicate desires, I can now say what I mean by the 'basic desires' in a given cloud:

> For any project, P, the **basic desires** in the P-cloud at *t* are to be found by carrying out the following steps (in the order given):
> 1) List all the desires in the P-cloud at *t*.
> 2) Find all the pairs of duplicate desires.
> 3) For each such pair, delete just one of the desires in that pair.[25]
> 4) From the desires that remain in the P-cloud, delete all the derivative desires.
> 5) The remaining desires are the basic desires in the P-cloud at *t*.

Given this account of the basic desires in a project, it should be clear what I mean when I say that the overall importance of project P for S at *t* equals the sum of the strengths of the basic desires that are contained in the P-cloud at *t*.

Stating Cloud Desire Satisfactionism. Now we are in a position to state Cloud Desire Satisfactionism in full. It would be cumbersome to state the theory in the tri-partite way that I stated Classical ADS. So instead, I formulate the view as a procedure for determining the total amount of welfare contained in a person's life. (This corresponds to item (iii) in the tri-partite schema.)

> <u>Cloud Desire Satisfactionism (CDS)</u>:
>
> The total amount of welfare contained in S's life is given by the following procedure:
> 1) For every time, *t*, during S's life, identify the desire clouds that S has at *t*.
> 2) Determine which of S's desire clouds at *t* count as projects.

[24] Lewis introduces the notion of a 'least subject matter,' which is roughly the minimal set of possible worlds on which the truth of a proposition supervenes entirely. Using this terminology, I think we could say that two desires, D and D', are duplicates iff the least subject matter of the object of D is identical to the least subject matter of the object of D'. (Cf. Lewis, 1998, ch. 8)

[25] In the strange event that two duplicate desires differ in strength, delete the weaker of the two.

3) For each project, P, that S has at *t*, determine two things: a) S's level of success with respect to P at *t*, and b) the overall importance of P for S at *t*.
4) The contribution to S's welfare made by P at *t* equals [S's level of success with respect to P at *t*] x [the overall importance of P for S at *t*].
5) S's welfare level at *t* equals the sum of the contributions to S's welfare made by all the projects that S has at *t*.
6) The total amount of welfare contained in S's life equals the integral of S's welfare levels for all times in S's life.

What has intrinsic value for a person, according to CDS, is success with respect to those of one's desire clouds that count as projects, i.e. the ones that seem to be largely within one's control. What has intrinsic disvalue for a person, according to CDS, is failure with respect to one's projects. The sum of the strengths of the basic desires in the cloud of a project is what determines *how much* value or disvalue one's success or failure at this project has.

CDS represents my best attempt to solve the technical problems discussed in section 6.2. For one thing, CDS offers a solution to the problem of irrelevant desires (a proposal which is builds on some of Keller's ideas). What's more, CDS builds on Skow's proposal by incorporating degrees of success and failure with respect to one's projects. Finally, CDS avoids the double-counting worries raised by redundant desires because CDS is formulated in terms of the notion of basic desires. (Though see second to last footnote in this chapter.)

6.3.2 A question about CDS

I realize that my formulation of CDS might raise some questions. In particular, one might think that a strange feature of CDS is this: on CDS, it's only the basic desires that count when calculating the overall importance of a project for you, but all non-instrumental desires count when calculating your level of success with respect to a project. Despite the *prima facie* oddness of this feature of CDS, however, there are good reasons for the theory to have this feature.

For starters, why can't the importance of a project just be equal to the sum of the strengths of *all* the desires in that project? As I briefly mentioned above, the reason is that this would lead to double-counting. To see why, consider Jack. He has two projects: to write a dissertation and to get married. To keep it simple, suppose Jack's dissertation-related project is composed of only two non-instrumental desires:

D5- Jack desires to write *some* dissertation (irrespective of its quality)
D6- Jack desires to write a dissertation that is original and influential.

D5 has a strength of +20, while D6 has a strength of +30. (Notice that Jack has D5 here purely in virtue of D6. After all, the satisfaction of D5 would be guaranteed by the satisfaction of D6. Thus D5 can't be a basic desire.) Next, suppose that Jack's marriage-related project contains just one non-instrumental desire:

D7- Jack desires to be married.

Suppose D7 has a strength +50.

Which of these two projects has more importance for Jack? If we were to take it that the importance of a project just is the sum of the strengths of *all* the desires it contains, then these two projects would be equally important to Jack. After all, (D5+D6) = D7 = 50. But this result seems wrong. Not only would Jack prefer the satisfaction of D7 to the satisfaction of either D5 or D6 alone, it also seems that Jack would prefer the satisfaction of D7 to the satisfaction of D5 and D6 *together*. This is because D5 here is derivative. It is a desire that Jack has purely in virtue of having D6. Given that D6 has a strength of just +30 and that the satisfaction of D6 would guarantee the satisfaction of D5, how could it be the case that Jack desires the satisfaction of both D5 and D6 to a degree *more* than +30? It can't. Thus it would be wrong to take the importance of a project for a person just to be the sum of the strengths of all the desires contained in that project.

Instead, my suggestion was that we should take the importance of a project for a person to be the sum of the strengths of all the *basic* desires in that project. Since D5 is not basic, but D6 is basic, it would follow that the importance of the dissertation-writing project for Jack equals the strength of D6, viz. +30. This, it seems to me, is a plausible result.

The next question, then, is this: when calculating your success with respect to a project, why should some *non-basic* desires get to count? After all, recall that your level of success with respect to project P at time *t* is determined by 1) adding up the strengths of all the desires in the P-cloud at *t*, and then 2) subtracting i) the proportion of this total that comes from frustrated desires from ii) the proportion of this total that comes from satisfied desires. This will give you a number between -1 and +1.

Here's the reason why some non-basic desires get to count when determining your success with respect to a project. Consider Jack's dissertation-writing project again, which consisted of D5 and D6. Suppose Jack actually writes a dissertation, but not an original or influential one. Thus D5 would be satisfied, but D6 would be frustrated. It seems clear that Jack should receive *some* benefit from the satisfaction of D5, even though D5 is not a basic desire (i.e. it is derivative). This is why I think we should let *all* the desires in a project[26] (which will all be non-instrumental) count when calculating one's degree of success with respect to that project.

Thus it should be clear why I have formulated CDS in such a way that it is only the basic desires that count when calculating the *importance* of a project for you, while all the desires in the project will count when calculating your level of *success* with respect to that project.

For purposes of illustration, let me conclude my presentation of CDS by considering the implications that the theory would have about the example of Jack. Suppose Jack has just finished his dissertation, but not it is not a very original or influential one. Moreover, suppose Jack has just been married. Jack's welfare at the present moment, then, should be calculated as follows, according to CDS. First, take Jack's dissertation-writing project. The current importance of this project for Jack = the sum of the strengths of all the basic desires presently in the dissertation-cloud = the strength of D6 = +30. Jack's current level of success with respect to this project = (20/50 – 30/50) = -0.2. Thus the contribution of the dissertation-writing project to Jack's welfare level = -6. Next consider Jack's marriage project. The current importance of this project for Jack = the strength of D7 = +50. Jack's level of success with respect to this project = (50/50 – 0/50) = 1. Thus the contribution of this project to Jack's welfare level = +50. The upshot of all this, then, is that Jack's current welfare level = (50 – 6) = +44. I hope this example helps to illustrate the workings of CDS.

[26] Perhaps with the exception of every other non-derivative duplicate desire.

6.4 Conclusions

In this paper, I have presented a version of Actual Desire Satisfactionism, which I called Cloud Desire Satisfactionism (CDS). The theory has many features that makes it attractive. For one thing, its appeal to the idea that people have entire clouds of fine-grained desires lends the theory a degree of realism, psychologically speaking, that might not seem to be present in the work of more traditional desire satisfactionists. But more importantly, CDS has the resources to avoid three serious problems that threaten classical formulations of the actual desire satisfaction theory.

For one thing, CDS offers a plausible solution to the problem of irrelevant desires. This is because CDS implies that one's welfare may be impacted only by desires that are part of a project. A project is a desire cloud of yours where it's reasonable to think that most of the desires in it can be satisfied through your own efforts. Thus CDS implies that your welfare cannot be impacted by desires like the desire for the recovery of the stranger on the train, or the desire for the number of stars in the sky to be prime. This is not a new solution to the problem of irrelevant desires. Keller, for instance, has already suggested it. But the solution to the problem offered by CDS is *better* than Keller's. For as we saw, CDS implies that one can also be benefited by 'manna from heaven', while Keller's solution does not.

Second, CDS largely avoids double counting problems that undermine classical versions of Actual Desire Satisfactionism. This is because of the central place that the notion of *basic desires* occupies in CDS. We saw in section 6.2.2 that double-counting problems are generated by both desires for states of affairs that are part of some bigger state of affairs that one also desires, and desires that follow trivially from other desires one has. However, for reasons explained above, neither of these desires will count as basic. So because CDS takes it that the importance of a project for you equals the sum of the strengths of just the *basic* desires in that project, CDS would seem to avoid the sort of double-counting that cast doubt on Classical ADS.[27]

[27] Perhaps double-counting problems remain. One might worry that since non-basic desires count when calculating one's level of success with respect to a project, double-counting will still take place on CDS. In particular, perhaps one could beef up one's level of success just by getting oneself to form desires that follow trivially from one's already satisfied desires.

Finally, CDS avoids the problem of partial desire satisfaction. Classical ADS seemed to be threatened by the apparent fact that we sometimes manage to satisfy a given desire only partially. But once we recognize that real people possess entire clouds of very fine-grained desires, it becomes possible to offer an account of degrees of success with respect to a project, and this in turn yields an explanation of what is going on when we say that a given desire is satisfied only partially. In particular, when you seem to have a partially satisfied desire, what is really going on is that there is some project (i.e. cloud of desires) you have that you are more successful than not with respect to, but that you still fall short of *complete* success with respect to. This is what people mean by their loose talk of a desire's being partially satisfied.[28]

Thus, in light of the benefits that Cloud Desire Satisfactionism offers, I think the theory represents a promising avenue for the desire satisfactionist. Nonetheless, I must acknowledge that more work needs to be done on the theory. In particular, I have not offered a satisfactory account of what makes desires belong to the same cloud. Because of this, the implications of CDS are still not entirely clear. However, I am confident that such an account can eventually be given. (Perhaps Lewis' work on relevance will prove helpful here.) And when this account is in place, I think CDS will then be the best version of Actual Desire Satisfactionism.

I don't see how to fully avoid this. However, I argued in section 6.3.2 that allowing non-basic desires to count when determining one's level success does not lead to the dangerous kind of double-counting. After all, it seems *correct* that some desires that you have in virtue of other more basic desires of yours should count towards your level of success. Nonetheless, I admit that there might still be a problem here. So more work might need to be done in order to fully avoid double-counting objections to Cloud Desire Satisfactionism.

But I suspect that a good first step would be this. While desires that you have in virtue of some other desire in the P-cloud should count when determining your level of success with respect to P, *duplicate desires should not get to count when calculating one's level of success.* Thus before running the official procedure that calculates one's level of success with respect to P at *t*, we would have to identify the pairs of duplicate desires in the P-cloud at *t*, and then for each such pair, delete one of the desires in that pair from the cloud. This would further mitigate the amount of double-counting that would occur on CDS.

[28] Couldn't the defender of Classical ADS also avail herself of this insight in order to mitigate the force of the problem of partial desire satisfaction? That is, couldn't one adopt this same strategy for making sense of loose talk of 'partially satisfied desires' without changing the letter of Classical ADS? Perhaps. Still, this would not reduce the plausibility of the solution as it is employed in Cloud Desire Satisfactionism.

CHAPTER 7

THE MAIN PROBLEM WITH DESIRE SATISFACTIONISM

My aim in this chapter is to argue that a certain partly response-independent version of the desire satisfaction theory, namely *Worthiness Adjusted Desire Satisfactionism*, is the most promising theory in the desire satisfaction family. This should not come as a surprise, considering the arguments I gave in chapter 2 for the claim that the true theory of welfare is to be found in the partly response independent category. But the argument of chapter 2 was sweeping, and in the present chapter, I will discuss in more detail my argument from chapter 2 as it applies to desire satisfaction theories in particular. I will discuss a number of problem cases that different philosophers have used to attack theories in this family, and I will argue that the entirely response dependent versions of the view cannot satisfactorily accommodate these cases. Instead, I argue, the desire satisfactionist can meet the objections based on these cases only by formulating a version of the theory that is partly response *independent*. The cases I appeal to here are akin to the ones that I argued in chapter 2 cast doubt on the entirely response dependent theories as a group, and so the argument of this chapter can be seen as providing additional support for the conclusion of chapter 2.

The order of business will be as follows. In section 7.1, I proceed to discuss Ideal Desire Satisfactionism and the motivation for it. In section 7.2, I present a range of cases that undermine both the best versions of Actual Desire Satisfactionism (i.e. Cloud Desire

Satisfactionism) and the best versions of Ideal Desire Satisfactionism. In section 7.3, I attempt to avoid these problem cases by incorporating objective restrictions into the desire satisfaction theory. This yields what I take to be the most promising version of the desire satisfactionist view, Worthiness Adjusted Desire Satisfactionism. It, too, is a cloud theory. Nonetheless, I argue in section 7.4 that this theory – the best one in the desire satisfaction family – also faces a serious problem. (This problem, incidentally, is analogous to the one that we saw in chapter 4 threatens Desert-Adjusted Intrinsic Attitudinal Hedonism.) Thus I conclude that despite the plausibility of Worthiness Adjusted Desire Satisfactionism, the view cannot capture the whole truth about well-being.

7.1 Ideal Desire Satisfactionism

I argued in chapter 6 that Cloud Desire Satisfactionism represents the best formulation of Actual Desire Satisfactionism. However, some philosophers think that *all* versions of Actual Desire Satisfactionism should be abandoned because of the problem that our actual desires may be based on mistaken or incomplete information. This problem of misinformation motivates the move to Ideal Desire Satisfactionism. I am willing to agree that Ideal Desire Satisfactionism, by avoiding the problem of misinformation, might be superior to Actual Desire Satisfactionism.

However, Ideal Desire Satisfactionism, too, needs to be formulated as a cloud theory in order to get around the technical problems discussed in chapter 6, viz. the problem of irrelevant desires, the problem of partial desire satisfaction and the problem of double-counting. For these problems threaten traditional versions of Ideal Desire Satisfactionism, just as they do traditional versions of Actual Desire Satisfactionism. In a moment, I'll present a cloud version of Ideal Desire Satisfactionism. But first, let us briefly look at the problem of misinformation, which motivates the move to Ideal Desire Satisfactionism in the first place.

The problem of misinformation is a well-known problem for actual desire satisfaction theories. Griffin, for example, puts it like this:

> Yet, notoriously, we mistake our own interests. It is depressingly common that when even some of our strongest and most central desires are fulfilled, we are no better, even worse, off. Since the notion we

> are after is the ordinary notion of ‘well-being’, what must matter for utility will have to be, not persons’ actual desires, but their desires in some way improved. (Griffin, 1986, p. 10)

As applied to Cloud Desire Satisfactionism, the problem is that a person might possess various projects (or else various particular desires inside of one’s various projects) only because one is missing important information, or makes some mistake of reasoning, or is not fully rational. If one were more informed or more rational, one might have had other projects entirely, or one’s projects might have been composed of very different particular desires. Thus we might think that one’s good lies not in satisfying the desires that *actually* make up one’s projects, but rather in satisfying the desires that *would* make up one’s projects if one were fully informed, fully rational, etc.

As Heathwood points out, problems for the actual desire satisfaction theory based on misinformation, irrationality, unimaginativeness, etc., have led some philosophers to adopt an *ideal* desire satisfaction theory.[1] I will keep my discussion of the ideal desire satisfaction theories brief because they have been explored in detail by many others.[2] Heathwood formulates a version of the ideal desire satisfaction theory that proceeds in terms that are similar to the ones used in Classical ADS, viz. the satisfaction or frustration of certain desires, the intensity of these desires, etc. But this means that Heathwood’s version of the ideal desire satisfaction theory faces analogs of the same three technical problems that were discussed in section 6.2, and which led us to abandon Classical ADS. In particular, traditional versions of Ideal Desire Satisfactionism also fails to incorporate degrees of desire satisfaction, and to address the problem of irrelevant desires and the problems concerning double-counting. Thus what we need is an ideal desire analog to Cloud Desire Satisfactionism.

I’m going to call this theory Ideal Cloud Desire Satisfactionism, or ICDS. The idea behind this theory is that what has intrinsic value for a person is success with respect to the projects that an idealized version of oneself would have. Similarly, what has intrinsic disvalue for a person, according to ICDS, is failure with respect to the projects that an idealized version of oneself would have. To state the theory in full detail, we would need an account of what is to count as an appropriately idealized version of oneself. This

[1] Cf. Heathwood, SDS, ms, pp. 6-8

[2] Dan Egonssen; David Sobel; Peter Railton, 2003; Griffin, 1986; Parfit, 1984; Kagan, 1998; Rosati, 1996.

question has received much discussion,[3] and I have nothing particularly to add. To state ICDS, I will just assume the following sketch of an account:

> S* is an ideal counterpart of some actual person, S, iff S* is fully rational and fully informed.

I assume that being fully rational requires, at a minimum that one commit no inferential errors.[4] Likewise, I assume that being fully informed requires at a minimum that one is not missing any *knowledge* such that if one were to obtain it, this would alter what one's desires are or what their strengths are.[5]

To state the view, I will rely on the accounts of levels of success with respect to a project and the importance of a project that were developed in the previous chapter. The only difference between the actualist version of CDS presented in chapter 6 and ICDS, presented here, is that what counts towards one's welfare is not one's actual projects, but the projects that an idealized counterpart of oneself would have (i.e. for one's actual self). The view, then, can be stated as follows:

> <u>Ideal Cloud Desire Satisfactionism (ICDS)</u>:
>
> The total amount of welfare contained in S's life is given by the following procedure:
>
> 1) For every time, *t*, during S's life, identify the desire clouds *that an idealized counterpart of S, viz. S*, would have at t.*
> 2) Determine which of S*'s desire clouds at *t* count as projects.
> 3) For each project, P, that S* has at *t*, determine two things: a) the overall importance of P for S* at *t*, and b) the level of success that S (i.e. the actual person, not the counterpart) has with respect to P at *t*.
> 4) The contribution to S's welfare made by P at *t* equals [S's level of success with respect to P at *t*] x [the overall importance of P for S* at *t*].
> 5) S's welfare level at *t* equals the sum of the contributions to S's welfare made by all the projects that S* would have at *t*.
> 6) The total amount of welfare contained in S's life equals the integral of S's welfare levels for all times in S's life.

I won't go into any more detail about how to understand ICDS. The basic idea should be clear. What's more, I am happy to admit that this theory avoids the problem of misinformation. For there are further substantive problems that threaten both ICDS and the actualist version of CDS discussed in chapter 6.

[3] See Shelly Kagan, 1998, p. 38. See Heathwood, SDS, ms, pp. 6-8. See Griffin, 1986, p. 12.

[4] See Griffin, 1986, p. 12

[5] This is similar to what Sumner says about what's required for a desire of yours to be authentic. (Cf. Sumner, 1996, ch. 6)

7.2 The Problem of Objectively Defective Projects

In this section I will argue that both CDS and ICDS are false because of the *problem of objectively defective projects*.[6] In particular, the problem is that both ADS and ICDS have counter-intuitive implications about a range of cases. They both imply that one's welfare would be enhanced by success with respect to projects that are in some sense objectively defective, even though intuitively it seems this should not enhance one's welfare.

7.2.1 Presenting the problem

This sort of objection has been offered by many philosophers, and a number of different cases have been used to make the point. Brink,[7] Heathwood,[8] Kraut,[9] and Parfit[10] all discuss several cases of this sort, for instance. However, their cases involve *single desires* that are allegedly defective (which is natural, of course, since these philosophers are concerned to argue against a view like Classical ADS). But I want to discuss this objection as it applies to what I take to be the best versions of the desire satisfaction theory – namely CDS and ICDS – and these theories are formulated in terms of projects (i.e. whole desire-clouds), not single desires. Thus I need to slightly re-describe the cases so that they are relevant to the theories I am concerned to discuss.

Heathwood offers a helpful taxonomy of 'defective desires,' i.e. desires whose satisfaction do not seem to enhance one's welfare. He mentions six types in all: ill-informed desires, base desires, poorly cultivated desires, pointless desires, artificially aroused desires, and the desire to be badly off.[11] I will not discuss ill-informed desires because this problem was already discussed in connection with the move to Ideal Desire

[6] I got this name from Heathwood's paper, 'The Problem of Defective Desires', (Cf. Heathwood, 2005)

[7] Cf. Brink mentions two examples of defective desires: the case of Ludwig (who has immoral desires) and the case of Zelda (who has trivial desires). Cf. Brink, 1989, p. 227.

[8] Heathwood, 2005, pp. 487-504.

[9] Kraut mentions, among others, the example of the person who desires his own misery (the self-punisher) and the icicle smashing fanatic. (Cf. Kraut, 1994, pp. 40-42)

[10] Parfit mentions Rawls' grass-counter example, as well as the desire to be badly off. Cf. Parfit, 1984, p. 500.

[11] Cf. Heathwood, 2005, p. 487-488

Satisfactionism, and I will not discuss poorly cultivated or artificially aroused desires because these cases do not (in my view) yield particularly powerful objections. Instead, I will focus just on the analogs of three of Heathwood's cases: what he calls base desires, pointless desires and the desire to be badly off. For each one of these types of defective desires, we can imagine a person who has an entire *project* that is similarly defective. This is how I will present the cases. (What's more, I will present them cases in such a way as to minimize the potentially distorting effects of extraneous facts about happiness and unhappiness.)

Begin with a paradigmatic case of a person with *base projects*. Fred Feldman offers a case (inspired by a passage from Moore[12]) that has been taken to pose a problem for Sensory Hedonism. This is the case of 'Porky', 'who spends all his time in the pigsty, engaging in the most obscene sexual activities imaginable.'[13] However, a modified version of Feldman's case can be used to threaten the desire satisfaction theories that we are interested in here:

> *Porky* – Porky has one project in life: to fornicate in the mud with the pigs as much as possible. This project of Porky's consists of all sorts of particular desires, all of which have a great deal of strength for him. Moreover, Porky's psychological make-up is such that he would continue to have this project (composed of these same particular desires) even if he were placed in 'ideal conditions' (e.g. if he were given full information, made fully rational and subject to extensive psychotherapy). As a matter of fact, Porky is highly successful with respect to this project. He spends all his waking hours rolling around in the mud with the pigs. What's more, he never feels any adverse effects from his chosen lifestyle. While he has no human friends, intellectual stimulation or significant challenges to overcome, this does not bother him. He never feels lonely or bored. Nor does he feel any sensory pain during his life. He never contracts any diseases from rolling in the mud, and he never sustains any injuries. Porky leads exactly the life he wants.

Other examples have been offered of people with base projects.[14] But the Porky example will suffice for our purposes.

[12] Moore, 1993, section 56

[13] Feldman, 2004, p. 40

[14] For instance, Brink mentions the example of Ludwig, a Nazi officer who dedicates his life to killing as many Jews as possible. (Cf. Brink, 1989, p. 227.) (Parfit briefly mentions a similar example. Cf. Parfit, 1984, p. 500.) However, this example does not provide a particularly compelling argument against desire satisfaction theories. In particular, the objection would be that the desire satisfaction theory implies that Ludwig has a good life, whereas, it's intuitive that a life like Ludwig's is not a good one. However, Heathwood gives a good answer to this objection. (Cf. Heathwood, 2005, p. 497-498) In particular, it seems that reason we think Ludwig does not have a good life is that his life is so fantastically immoral. But

Next, there are cases involving projects that are *pointless*. A good example is Rawls' case of the grass-counter,[15] who I am going to call Greg. For our purposes, the case may be described as follows:

> *Greg the Grass-counter* – Greg has one project during his life: to spend his life counting the blades of grass in the lawns in his neighborhood. This project of Greg's consists of all sorts of particular desires, all of which have a great deal of strength for him. Moreover, Greg's mind is so obsessed with counting blades of grass that he would continue to have this project even if he were placed in 'ideal conditions' (e.g. if he were given full information, made fully rational and subject to extensive psychotherapy). As a matter of fact, Greg is highly successful with respect to his grass-counting project. He spends all his waking hours on his hands and knees counting blades of grass in the different lawns in his neighborhood. What's more, he never feels any adverse effects from his chosen lifestyle. Although he has no human friends, intellectual stimulation or significant challenges to overcome, this does not bother him. He never feels lonely or bored. Nor does he feel any sensory pain during his life. He never contracts any diseases or sustains any injuries from crawling around on the ground all day. Greg leads exactly the life he wants.

The project that Greg dedicates his life to seems utterly pointless, and yet it is the project he wants to spend his life pursuing. Other philosophers also mention cases of this sort. Kraut, for instance, offers an example of a person who dedicates his life to knocking icicles off the roofs of the houses in his neighborhood.[16] Brink offers an example of a person who dedicates her life to developing the world's smallest handwriting.[17] These cases are, it seems to me, similar in every relevant respect to Rawls' grass-counter example. So I think we can make do with just the case of Greg.

Finally, I want to discuss a case that is in some respects similar to what Heathwood calls *the desire to be badly off*. Heathwood points out that some people with such a desire

this does not bear on the welfare value of the life. As Heathwood points out, we need to focus on the relevant scale of evaluation: namely how much welfare a life contains for the one who lives it, not how morally good a life is. It is beyond question that Ludwig's life is horribly immoral, but is not clear his life is such a bad one from the point of view of welfare. Thus if one has the intuition that Ludwig leads a bad life, this is arguably just because one has focused on the wrong scale of evaluation, viz. the moral scale as opposed to the welfare scale. Thus an objection to the desire satisfaction theory based on the case of Ludwig does not seem likely to succeed.

One might wonder whether a similar line of response would undermine the argument based on the cases of Porky and the others. In a moment, I will argue that this response is much less plausible in the cases of Porky and the rest than it is in the case of Ludwig. (In other words, I admit that it works for Ludwig, but not that it works for Porky, Greg and Max.)

[15] Cf. Rawls, 1971, p. 432. Parfit also discusses this case. (Cf. Parfit, 1984, p. 500.)

[16] Kraut, 1994, p. 42

[17] Brink, 1989, p. 227

might show desire satisfactionism to be paradoxical,[18] and Ben Bradley has argued for a similar conclusion.[19] However, Brad Skow has suggested an interesting way for the desire satisfaction theory to avoid the paradox.[20] Thus I am not going to focus specifically on the argument that desire satisfactionism is paradoxical. Instead, I will be concerned to argue that desire satisfaction theories give intuitively implausible consequences about people who desire something very close to being badly off.

Instead of focusing on a person who has an abstract desire for his own level of welfare to be negative (which is the sort of case that might show desire satisfactionism to be paradoxical), let us consider a person who dedicates himself to achieving something slightly more specific: viz. to make himself *sick, miserable, humiliated and pathetic*.[21]

> *Max the Masochist* – Max has one project in life: to achieve sickness, wretchedness, and humiliation. This project of Max's consists of all sorts of particular desires relating to his becoming pathetic and humiliated, and all of these desires have a great deal of strength for him. Moreover, Max's psychological make-up is such that no amount of information, rationally or psychotherapy could dislodge his pre-occupation with being sick, wretched, humiliated and pathetic. Thus this project of his would persist even if he were placed in 'ideal conditions.' As a matter of fact, Max gets just what he wants. He becomes extremely sick, and his sadistic doctors play cruel jokes on him until the very end. They experiment on him with drugs that rot his body and mind into a horrible, utterly pathetic state of waste and confusion. His existence is a perpetual state of weakness, humiliation and wretchedness (though it should be noted that he does not experience much sensory pain). This is exactly the life Max wants to lead. He judges that his life is going exactly according to plan.

Given this catalogue of cases in which people have various fundamental concerns that seem defective in one way or another, we can state the argument against CDS and ICDS.

[18] Heathwood, 2005, p. 501-503

[19] Bradley, 2008

[20] Skow, ms. My favored versions of the desire satisfaction theory, viz. CDS and ICDS, cannot appeal to Skow's solution to this paradox. The reason is that, while success with respect to a project comes in degrees, CDS and ICDS are not literally countenance the possibility of desires that can be partially satisfied. On CDS and ICDS, for every desire and every time, it is either satisfied or frustrated at that time. Thus if one of one's projects contains the desire to be badly off, it will be possible to construct cases in which CDS and ICDS imply a contradiction. So I am inclined to think that CDS and ICDS are in fact paradoxical. There's nothing I can do about that now. This problem has to be dealt with in depth and cannot be sufficiently addressed in just a footnote.

[21] The case of Max the Masochist has been suggested to me in conversation by Fred Feldman. A similar case is discussed by Parfit (cf. Parfit, 1984, p. 500). Also see Kraut's self-punisher (Kraut, 1994, p. 40). (Also see Heathwood, 2005; Bradley, 2008; and Skow, ms.)

The Argument from Objectively Defective Projects

1) If either CDS or ICDS are true, then Porky, Greg and Max lead lives that are very good for them.
2) But it's not the case that Porky, Greg and Max lead lives that are very good for them.
3) Therefore, it's not the case that either CDS or ICDS are true.

The rationale for line 1) here is that Porky, Greg and Max are all highly successful with regard to their respective projects. Porky's project is to spend his life rolling in the mud with the pigs, and he succeeds. Greg's project is to spend his life counting blades of grass, and he succeeds. Max's project is to become sick, weak, humiliated and pathetic, and he does. Thus CDS implies that all three have lives that are very good in terms of individual welfare. What's more, since the ideal versions of Porky, Greg and Max would have the same projects as their actual counterparts, it's also the case that ICDS implies that all three have lives that are very high in individual welfare.

However, the idea behind line 2) is that Porky's, Greg's and Max's respective projects are all objectively defective in some way or another. I find it intuitive that success with respect to projects like these would not enhance one's welfare very much. There are several grounds for this intuition. For one thing, the projects in all three cases seem to be pointless. There is nothing particularly worthwhile or valuable that is served by rolling in the mud with the pigs, counting blades of grass or sinking into a state of weakness and endless humiliation. The lives of Porky, Greg and Max have no significance or lasting impact. Their pursuits are trivial and inconsequential. Second, the intuition that Porky, Greg and Max do not have lives that are high in individual welfare is driven by the consideration that their lives are not in any way admirable or impressive. I would be surprised if there has ever been anyone who would find Porky, Greg or Max to be especially worthy of respect or admiration in virtue of the success they have with respect to their various projects.

But more importantly, the intuition behind line 2) is supported by the consideration that no one would be likely to wish the lives of Porky, Greg or Max on anybody that one cares about. Of course, very few people are going to have the same tastes, dispositions or psychological profiles as Porky, Greg or Max. So my point here is not that if there were some person, S, that you care deeply about, you would not be likely to wish upon S a lifestyle like Porky's, Greg's or Max's. After all, S is highly unlikely to care about the same sorts of things as Porky, Greg or Max. Rather, my point is this. Suppose you have a child about whom you care deeply. Moreover, suppose you find out (e.g. through some kind of advanced genetic testing) that your child will have, say, the same unalterable obsession with grass-counting as Greg. Would you in that case wish Greg's life on your child? Or would you wish that there were some therapy or treatment available to your child that could provide for him or her a less trivial, more normal and more admirable life? I submit that many – perhaps most – people would wish that such a treatment were available for their child in such a case.[22] A similar point can be made when it comes to Porky and Max as well. This, I think, provides yet more support for line 2).

In my view, the argument from objectively defective projects ultimately refutes CDS and ICDS. However, it is still too early to reject these theories on the basis of this argument. For there are some interesting replies available to a supporter of these theories. We must discuss these responses before we can take the argument from objectively defective concerns to be successful.

7.2.2 Responses on behalf of desire satisfaction theories

First (and perhaps least interestingly) one might claim that while CDS is refuted by this argument, ICDS is not. After all, one might think that if only Porky, Greg and Max were given enough information, were made fully rational[23] and were subject to extensive enough psychotherapy, they would abandon their defective projects in favor of 'better' ones. Thus one might think ICDS does not really imply that Porky, Greg and Max lead good lives.

However, such a response involves a misunderstanding of the cases as they were told. After all, it was stipulated that Porky, Greg and Max would not lose their defective projects even in the face of complete information, full rationality and extensive psychotherapy. Surely this stipulation does not make the cases impossible or incoherent.

[22] Moreover, even if the two lives were guaranteed to contain exactly the same total amount of desire satisfaction, we would wish the non-trivial, normal and admirable life on our children rather than a Greg-type life. All other things being equal, the pointlessness of one's pursuits makes one's life less good for one. I will argue for this in more detail in section 7.2.3.

[23] Procedural rationality is what is in question here. Not substantive rationality. That would be a question-begging way to make Ideal Desire Satisfactionism avoid the problem of objectively defective concerns.

So the move to Refined IDS does not succeed in avoiding the objection from objectively defective projects.

Heathwood, however, suggests two other responses that have much more plausibility. The first response depends on the important distinction between intrinsic goodness or badness for a person, on the one hand, and all-things-considered goodness or badness for a person, on the other. As Heathwood puts it,

> a state of affairs p is intrinsically bad for someone S iff given two lives exactly alike except with respect to p, the p-life is worse for S than the not-p life. (Heathwood, 2005, p. 491)

By contrast, something is all-things-considered bad for you, roughly speaking, if this thing would make your life go worse than it otherwise would have gone. As Heathwood puts it,

> 'a state of affairs p is *all-things-considered bad* for someone S iff the life S would lead if p were to obtain is worse for S than the life S would lead if p were not to obtain.' (Heathwood, 2005, p. 491)

Thus suppose that at a particular time, t, there are two different lives available to you that you might go on to lead: life A and life B. Suppose that the total amount of welfare for you contained in A is *less* than the total amount of welfare for you contained in B. If there is a state of affairs, p, such that i) you would lead life A (rather than B) if p were to obtain, and ii) you would lead life B (rather than A) if p does not obtain, then p is all-things-considered bad for you. Thus your mother's drinking heavily while she is pregnant with you is likely to be all-things-considered-bad for you, but arguably it is not intrinsically bad for you. All-things-considered goodness can be understood in an analogous way.

Heathwood's response to the argument from objectively defective projects, then, would go like this. Line 1) in the argument is to be interpreted in terms of intrinsic goodness. After all, CDS and ICDS are, first and foremost, theories about what is intrinsically good or bad for a person, not what is all-things-considered good or bad. However, line 2) in that argument seems plausible only because it is tacitly appealing to all-things-considered goodness instead of intrinsic goodness. After all, there clearly are *other* lives that Porky, Greg or Max could have led instead of their actual ones that would have been much better in terms of the degree to which they succeed at various projects. As the lives of Porky, Greg and Max were described, they have but one specific project each. But, for instance, suppose we brainwash them so that they come to have a broad range of additional projects: they become concerned to have meaningful relationships with other people, to lead a life of intellectual achievement, to make lots of money and to do good things for the world. Moreover, suppose they are quite successful in all these pursuits. Compared with the actual lives of Porky, Greg and Max, a life such as this one would clearly involve a much greater amount of success with respect to projects. Thus according to CDS (or ICDS) such a life would contain a great deal more welfare than the lives that Porky, Greg and Max actually lead. And this is the only reason why it might seem, as line 2) claims, that the lives of Porky, Greg and Max are not good for them – viz. these lives are not all-things-considered good for them. Now, the defender of CDS or ICDS can *grant* this claim about all-things-considered goodness, while at the same time insisting that the implications of CDS or ICDS are correct: namely, that Porky's, Greg's and Max's actual lives are indeed highly intrinsically good for them. Thus the defender of CDS or ICDS would be insisting that line 2) when interpreted in terms of intrinsic goodness is false. So goes the objection.

To avoid this sort of objection, however, it seems to me that all we have to do is stipulate that in the cases of Porky, Greg and Max, there is no other life available to them in which they enjoy a greater degree of success with respect to their projects than is contained in their actual lives. This is not an implausible stipulation to make. After all, given the unalterability of Porky's, Greg's and Max's respective interests, there is not much you can do – short of reprogramming them through comprehensive brain surgery – to get them to be committed to different projects than they actually are. Thus it is reasonable to suppose that if they lead any other life than their actual one, they are going to have *less* success in their various pursuits than they actually do.

Suppose that this is the way things are for them. This ensures that line 2) in the argument is not plausible if interpreted in terms of *all-things-considered* goodness. So line 2) would have to be interpreted in terms of *intrinsic* goodness instead. But does line 2) now suddenly seem entirely implausible? I don't think so. Even when interpreted in terms of intrinsic goodness rather than all-things-considered goodness, line 2) does not lose its intuitive force. The lives of Porky, Greg and Max still do not seem like particularly good ones for them. Thus I am not convinced by Heathwood's first response to the argument.

What about Heathwood's second response? With respect to the Porky example, for instance, Heathwood has this to say:

> Nor are we giving our approval of the indulger's behavior [i.e. Porky's]. It may turn out that what the indulger is doing is morally wrong, and if it is, this is consistent with the desire theory of welfare. Nor are we saying that the life of the perpetual indulger ranks high on the other scales on which we rank lives, such as the scales that measure virtue, dignity, or achievement. In short, it is perfectly consistent for an actual desire-satisfaction theorist to issue the following judgment about the perpetual indulger: 'What a pity! Sure, he's well off there in the barnyard, happy doing his thing, getting just what he wants, but his life is pathetic: he will achieve nothing; what he does is degrading; and his moral character is woefully underdeveloped. I would not wish this life upon anyone.' (Heathwood, 2005, p. 497)

What these comments of Heathwood's suggest is that line 2) in the argument from objectively defective projects seems plausible only because we are confusing welfare value with other types of value. Granted the lives of Porky, Greg and Max are very low on the morality scale, the achievement scale, the aesthetic value scale and a bunch of other evaluative scales as well. The defender of a theory like CDS or ICDS can accept all this. But when we focus on the scale of evaluation that is relevant – namely the *welfare* scale, on which lives are ranked according to how well they go for the people who lead them – we have no choice but take it that Porky's, Greg's and Max's lives actually score quite high. In other words, when we focus on the relevant scale of evaluation, the implications of CDS or ICDS are not implausible after all. When we're careful to focus specifically on welfare value, not some other type of value, we must reject line 2) in the argument.

I am not convinced by this response for two reasons. The first is autobiographical. I think I know what the concept of welfare is. I have read a lot of books and articles about it. I have encountered many theories about what things intrinsically enhance welfare. I understand the ways in which welfare value differs from moral value, aesthetic value, excellence, value for other people, and so on.[24] Nonetheless, I am still inclined to think that Porky's, Greg's and Max's lives are not high specifically in *welfare value*. (I don't think they are high in moral value, aesthetic value and all the rest either, of course.) And I do not think I am confusing the other kinds of value with specifically the welfare-type

[24] What's more, I wrote about this very question, viz. the question of what distinguishes welfare value from other types of value, at length in chapter 1 of this dissertation. I argued that welfare value has something to do with the kind of life that a generic person – perhaps a disembodied spirit – who is fully self-interested but not ideologically-motivated would want to lead when placed on earth.

value. My intuition is that Porky, Greg and Max do not lead lives that score high precisely on the welfare scale.

My second response to this objection of Heathwood's involves less navel-gazing. In particular, Heathwood seems to be appealing to a concept of well-being that is so narrow as to beg the question against those who are inclined to accept the argument from objectively defective projects. Heathwood writes:

> We therefore need a distinction among types of intrinsically defective desire – there are those that are welfare-defective, virtue-defective, dignity-defective, and achievement-defective. (There may be others for any additional scales on which a life can be ranked.) The lesson is that the Moorean argument must find an intrinsically welfare-defective desire, not merely an intrinsically defective desire or an all-things-considered-defective desire. (Heathwood, 2005, p. 498)

As this passage makes clear, Heathwood's response seems to assume that it is a conceptual mistake to think that things like one's virtue, dignity or achievement can in themselves impact one's welfare. After all, since the desire satisfaction theory is a theory about specifically welfare, what would be needed to refute this theory is 'a welfare-defective desire' – i.e. a desire whose satisfaction would not enhance one's welfare. But as Heathwood stresses, a desire that is, say, 'dignity-defective' or 'achievement-defective' (i.e. a desire whose satisfaction would not enhance one's dignity or level of achievement) will not suffice for this purpose. The assumption is that dignity-defective desires and achievement-defective desires cannot be welfare-defective desires. To think otherwise is to confuse the concept of welfare with other evaluative concepts. Thus Heathwood's response seems to be assuming that that it is a conceptual mistake to say that dignity, say, or achievement can directly impact one's welfare.

However, this seems to simply beg the question against theories of welfare according to which things like dignity or achievement can be directly relevant to one's welfare. Heathwood's response seems to commit him to ruling out the possibility of such a theory's being true on conceptual grounds alone. In order not to simply beg the question against a theory of welfare according to which achievement and dignity directly impact welfare, one's concept of welfare must at least allow it to be an *open question* whether achievement and dignity are directly relevant to welfare. If one's concept of welfare is so narrow that it's not even a theoretical possibility that things like achievement and dignity can impact welfare, then one seems to be begging the question against a range of substantive theories of welfare. Heathwood's second response to the argument from objectively defective concerns, it seems to me, is question-begging in precisely this way. And so I am not convinced by it.

Because I do not find any of these objections to the argument from objectively defective projects to be plausible (and because I don't know of any other objections), I am inclined to think that this argument is sound.

7.2.3 An argument in favor of counting the worthiness of your projects

In section 7.3, I will formulate a version of CDS that avoids the problem of objectively defective projects by incorporating certain objective constraints into the theory. But before I present this theory, let me briefly give an independent argument for the idea that the worthiness of the projects one succeeds with respect to intrinsically impacts one's welfare.

Compare the lives of two people, S1 and S2, that are similar in every respect except that S1's aims are more objectively worthy of pursuit than S2's aims. Let's say that both S1 and S2 have as their only project to smash the icicles on the roofs of the houses near them, and let's suppose that S1's icicle smashing project is just as important to S1 as S2's icicle smashing project is to S2. However, S1 has the justified true belief that, in his case, smashing icicles is very dangerous and requires great skill, and also that a lot of good for others hangs on his success at smashing the icicles. By contrast, in the case of S2, icicle smashing is an entirely pointless activity. It is easy, requires little skill and would benefit nobody. And S2 knows all this. Suppose both of these people have a psychological make-up such that they would retain their project of smashing icicles even if placed in idealized conditions of full information, full rationality, and so on. And as a final stipulation about this pair of lives, suppose that S1 and S2 are both successful to exactly the same high degree with regard to their respective projects.

Which life is better in terms of welfare, that of S1 or that of S2? CDS and ICDS both imply that S1's life contains exactly the same amount of welfare as S2's life. However, I suggest that this is counter-intuitive. Since S1's project is significantly more worthwhile than S2's project, and since they both succeed to the same high degree with regard to their respective projects, it seems to me that S1's life contains more welfare value than S2's life does. After all, we would admire S1's life more. We would tend to prefer this life for ourselves. We would tend to prefer for those we care about that they lead S1's life of success in icicle smashing, where this is a difficult and worthwhile task, rather than S2's life of success in icicle smashing, where this is an easy and entirely pointless task. These intuitions count, I suggest, as data against which theories of welfare are to be tested.[25] To my mind, they are sufficiently strong intuitions to warrant the rejection of any theory, e.g. CDS and ICDS, that conflict with them.

7.3 Worthiness Adjusted Cloud Desire Satisfactionism

To accommodate the intuitions that led to the rejection of CDS and ICDS, we need to formulate a version of the desire satisfaction theory that also allows the worthiness of one's projects to impact one's welfare. In this section, I formulate just such a theory. In my view, this theory is the most promising one in the desire satisfactionist family.

I am going to call this theory *Worthiness Adjusted Cloud Desire Satisfactionism*, or WACDS. This theory is similar in most respects to CDS, except that the intrinsic value for you of an episode of success with respect to one of your projects is going to be in part determined by the objective worthiness of the desires in that project. Thus the theory requires the assumption that for any particular desire that is found within some project that a person has, there is some degree of worthiness or intrinsic value that the object of this desire possesses.

I do not claim to know precisely what makes the object of one desire possess more worthiness than the object of some other desire. However, in chapter 4, in connection with Desert-Adjusted Intrinsic Attitudinal Hedonism, I discussed two good proposals about how to understand the notion of the worthiness of objects of some episode of attitudinal pleasure.[26] These proposals are also going to be live possibilities when it comes to understanding the worthiness of the objects of desires, as well. One option is to go with Adams' idea[27] and say that the object of some desire is objectively worthy to the extent that it is excellent (where excellence is to be understood in terms of resemblance to God). Alternatively, we might adopt my suggestion of an ideal observer account, and say

[25] This is exactly what I argued in chapter 1.

[26] See chapter 4, section 4.3.1-4.3.3.

[27] Cf. Adams 1999, p. 23

that the object of some desire is objectively worthy to the extent that an ideal observer would approve of that object's being realized.[28]

More work is needed on each of these proposals. Nonetheless, I will go ahead and formulate WACDS. I simply assume that for any desire, there is some degree of worthiness that the object of that desire possesses. What's more, I assume that this level of worthiness can be represented by a real number, which will be positive if the object of the desire is worthy all in all, but negative if the concern is unworthy all in all. I take it that there is in principle no upper limit on how worthy a given fundamental concern can be, and there is in principle no lower limit on how unworthy a given fundamental concern can be. To formulate WACDS, I need to introduce the notion of the overall worthiness of a project. Here is the procedure to follow in order to determine the overall worthiness of project P at t:

1) Find all the pairs of duplicate desires in the P-cloud at *t*.
2) For each such pair, remove one of them, so that the P-cloud contains no duplicate desires.
3) For all the remaining desires in the P-cloud, add up the levels of worthiness of their objects.
4) This number equals the overall worthiness of P at *t*.[29]

Given this notion of the overall worthiness of a project, we can say that the *contribution* to S's welfare made by project P at *t* = [S's level of success with respect to P at *t*] x [the overall importance of P for S at *t*] x [the overall worthiness of P at *t*].

With these assumptions in place, we can now in a position to state WACDS in full:

Worthiness Adjusted Cloud Desire Satisfactionism (WACDS)

> The total amount of welfare contained in S's life is given by the following procedure:
>
> 1) For every time, *t*, during S's life, identify the desire clouds that S has at *t*.
> 2) Determine which of S's desire clouds at *t* count as projects.

[28] In section 4.3.3, I suggested that those sympathetic to DAIAH should appeal not to the pleasure-worthiness of the *objects* of enjoyment, but rather to the worthiness of *episodes* of enjoyment. Perhaps we should do something similar here, viz. appeal to the worthiness not of the objects of desire, but rather of whole episodes of desire satisfaction. However, for simplicity I will ignore this possibility in what follows.

[29] There is another plausible way to conceive of the overall worthiness of a project at a time. However, I am not sure which is better. The idea, in any case is this:

> Find all the desires in the P-cloud at *t*, and consider the state of affairs consisting of as many of these desires being satisfied together as possible. The objective worthiness of this state of affairs is equal to the overall worthiness of P at *t*.

3) For each project, P, that S has at *t*, determine three things: a) S's level of success with respect to P at *t*, b) the overall importance of P for S at *t*, and c) the overall worthiness of P at *t*.
4) The contribution to S's welfare made by P at *t* equals [S's level of success with respect to P at *t*] x [the overall importance of P for S at *t*] x [the overall worthiness of P at *t*].
5) S's welfare level at *t* equals the sum of the contributions to S's welfare made by all the projects that S has at *t*.
6) The total amount of welfare contained in S's life equals the integral of S's welfare levels for all times in S's life.

I think WACDS is superior to both CDS and ICDS in that it gives the intuitively correct results about the lives of Porky, Greg and Max. Whatever the correct account of worthiness turns out to be, it will have to imply that the projects of Porky, Greg and Max are not worthy to a very high degree. Perhaps the worthiness of these projects is negative, or perhaps it is just a positive number that is close to zero. Either way, this would guarantee that WACDS implies that the lives of Porky, Greg and Max are not high in individual welfare. And this, I argued earlier, is the intuitively correct result. Thus WACDS seems to avoid the argument from objectively defective projects.

I think WACDS is the most plausible theory in the desire satisfactionist family. But notice that WACDS is a version of Actual Desire Satisfactionism. After all, it is one's success in with respect to the projects that one *actually has* that matters to welfare, according to WACDS. But what about the problem of desires based on misinformation, which motivated the move from CDS to ICDS? Might a worthiness adjusted form of *ideal* desire satisfactionism in fact be more plausible than WACDS, as I formulated it?

I don't think the advantages of such a theory – Worthiness Adjusted Ideal Cloud Desire Satisfactionism, or WAICDS – would be significant (if indeed there are any at all). For it seems to me that the main advantages offered by the move to ideal desires (i.e. the advantages that ICDS has over CDS) are also going to be provided by the sort of worthiness adjustment that WACDS involves. Suppose a person has adopted a project on the basis of some mistaken information. Suppose, say, Jason takes up the project of icicle smashing on the basis of the mistaken belief that icicles that are hanging from a roof are carcinogenic for the people inside the house. Since this is in fact not the case, then (assuming icicle smashing has no other independent significance) the worthiness of Jason's project is going to be quite low. And so even if Jason is very successful in smashing icicles, WACDS still will not imply that this will enhance his welfare by very much. And this seems to be the intuitively correct result. So in general, it seems to me that the main advantages of moving to ICDS can are going to be secured by building worthiness adjustment into the theory in the way that WACDS does.[30] Accordingly, my view is that WACDS represents the most promising theory in the desire satisfaction family.[31]

7.4 The Problem with WACDS

Despite its strengths, I do not think that WACDS represents the whole truth about welfare. In particular, I think it faces problems because it fails to recognize the intrinsic contribution to welfare that happiness and unhappiness seem to make. (Notice that this problem for WACDS is the exact analog of the problem that I claimed, in the last section of chapter 4, shows that DAIAH cannot represent the whole truth about welfare either.)

Consider a rather pleasant life such that the amount of success with respect to worthwhile projects that this life contains is exactly zero. For example, consider Bill, a person who is as apathetic as anyone can be. There is nothing that he cares much about. He does not aim to accomplish anything and he has no projects to speak of. He doesn't care about getting an education or making friends. He doesn't even care about having fun or being happy. Most of the time, Bill just sits around his house. It's not the case that he *desires* to just sit around the house all day, but it's also not the case that he desires *not* to sit around the house all day. Rather, he just doesn't care much either way. And this is how Bill feels about everything in life: utterly indifferent. Therefore, the amount of success with respect to worthwhile projects that is contained in Bill's life is zero. (If you think it's impossible for a person to have absolutely no projects or desires, then to ensure that Bill's life really does contain zero success with respect to projects, we can just add the stipulation about the case: if there are any desires that Bill *does* have, then every satisfied one is exactly matched by a frustrated one of precisely the same intensity,

[30] What's more, I suspect that by sticking with an actualist theory like WACDS, rather than an ideal desire theory, we are going to be in better shape with respect to the problem of irrelevant desires as well.

[31] WACDS is also going to be more plausible, I think, than a restricted version of desire satisfactionism, according to which it is only desires for things that meet some minimum requirement for worthiness that can count towards one's welfare. (Simon Keller offers some interesting objections to this sort of theory in 'Welfare and Achievement', 2003.) I think the desire satisfactionist is going to be better off allowing that success or failure in realizing *any* fundamental concern, no matter how pointless, can impact one's welfare *at least to some extent*.

duration and worthiness. Thus Bill's life really does contain zero success.) However, there is an oddity about Bill. He has a birth defect that mysteriously causes him to take pleasure in mundane activities. When he walks across the room he feels a pleasant tingle in the soles of his feet. When he flushes the toilet, he finds the sound oddly enjoyable. He gets a kick out of the yellow-ish tinge on the walls of his living room. His leg twitches in an amusing way whenever he sits on the couch for more than an hour at a time. And so on. Thus Bill's life actually contains quite a lot of pleasure. Nonetheless, Bill does not desire this pleasure. He does not seek it out, nor does he try to get more of it. It would not bother him in the least if the effects of his birth defect were cancelled, and all these little things pleased him no longer.

What does WACDS imply about the life of Bill? It implies that it is a worthless life. After all, Bill has no success whatsoever with respect to any project. (He does not even have any success with respect to the project of feeling happy, since he has no such project.) So Bill's life contains no welfare value at all, according to WACDS. But this implication seem counter-intuitive. For Bill actually experiences a decent amount of pleasure, even though he does not desire this or aim at it. Intuitively, Bill's life does seem to contain at least *some* positive amount of welfare. However, WACDS cannot yield this result, since Bill's life contains no success whatsoever with respect to any projects. Thus the theory seems to be unacceptable. It fails because it cannot account for the intuition that some lives that contain no success with respect to any projects can still contain a positive amount of welfare value (e.g. if these lives are particularly pleasant ones).[32]

So my view is that while WACDS gets a lot right, it does not capture the whole truth about welfare. In the next and final chapter of this dissertation, I will defend a theory that I think remedies this shortcoming. In particular, I will defend a theory combines WACDS with a version of Hedonism. This theory, I will argue, offers all the advantages of WACDS, while addressing the ways in which WACDS by itself falls short (i.e. the problems that are revealed by cases like that of Bill).

[32] A related problem with WACDS is that it implies that if there were a person who experienced no pleasure (or pain) in life, but who had a tremendous amount of success with respect to worthwhile projects, then this person's life would be a fantastically good one for him. This implication, too, seems counter-intuitive, however. For how can a life be fantastically good if it contains no pleasure whatsoever? Intuitively, a life must contain at least some pleasure, or happiness, if it is to be one that is fantastically high in welfare value. However, more about this problem in chapter 8, specifically section 8.2.2.

CHAPTER 8

THE HAPPINESS AND SUCCESS THEORY OF WELL-BEING

In this chapter, I discuss the question of how to develop a theory of well-being that fits with the intuitions that led us, in previous chapters, to reject other influential theories of well-being. I endorse a type of theory for which a fitting label is 'the Happiness and Success Theory'. What makes a theory belong to this type is that it makes welfare be a function of two things: how happy you are (i.e. how good you feel) and how successful you are in accomplishing worthwhile goals. Thus the Happiness and Success Theories are multi-component theories of welfare.

A number of philosophers have proposed multi-component theories of well-being of this sort. For instance, David Brink endorses a multi-component theory (one of whose components resembles WACDS in certain respects):

> Value must contain important objective components. This fact can be accommodated either within a purely objective theory or within a mixed theory. (…) I propose to discuss a theory of welfare that counts reflective pursuit and realization of agents' reasonable projects and certain personal and social relationships as the primary components of valuable lives. (Brink, 1989, p. 231)

T.M. Scanlon also seems to endorse a multi-component theory, though his theory has three components. In any case, the first two components look quite similar to the components of a Happiness and Success Theory:

> I conclude that any plausible theory of well-being would have to recognize at least the following fixed points. First, certain experiential states (such as various forms of satisfaction and enjoyment) contribute to well-being, but well-being is not determined solely by the quality of experience. Second, well-being depends to a large extent on a person's degree of success in achieving his or her main ends in life, provided that these are worth pursuing. (…) Third, many goods that contribute to a person's

> well-being depend on the person's aims but go beyond the good of success in achieving those aims. These include such things as friendship, other valuable personal relations, and the achievement of various forms of excellence, such as in art or science. (Scanlon, T.M. 1998, pp. 124-125)

Other philosophers have proposed other such multi-component views.[1] Multi-component theories are particularly promising, in my view. For such theories seem to be capable of avoiding the main problems of both the entirely response *in*dependent theories (e.g. certain Objective List Theories) and the entirely response *de*pendent theories (e.g. Hedonism and Desire Satisfactionism). Entirely response independent theories fail to account for the centrality of our attitudes to our welfare, while the entirely response dependent theories fail to account for the ways in which certain things can be good or bad for us independently of our attitudes. A multi-component theory that allows *both* happiness *and* success to count towards one's welfare, however, has the resources to avoid both sorts of objection.

The problem is that neither Brink, nor Scanlon, nor any other philosopher I know of has stated any such multi-component theory in full detail. Most worryingly, none of them discuss the question of how the *math* in such a theory should be worked out.[2] This question is an important one, however, because there are many ways in which the various components in a multi-component theory can be taken to be mathematically related to each other, and the theory will yield substantially different results depending on which way is picked. What's more, the many different ways in which the math can be done for multi-component theories provides a rich set of resources for dealing with problem cases. As we will see below, for instance, certain mathematical devices make it possible to avoid what seem to be knock-down objections to certain less mathematically sophisticated theories. However, because most traditional theories make welfare depend on only one component (i.e. are monistic), such mathematical resources have not, it seems to me, been sufficiently explored. This chapter will, I hope, take a step towards remedying this deficiency.

In particular, I will discuss the relative merits of three different ways to do the math when it comes to a particular type of two-component theory, namely the Happiness and

[1] See, for instance, Keller, forthcoming, (cf. especially sec. 2.9 and 2.10). Also see Raibley, 2007 ch. 4.

[2] Brink and Scanlon do not discuss this question. It not addressed in Keller (forthcoming) or Raibley (2007) either.

Success theory. My reason for focusing specifically on the Happiness and Success theory is that I think a strong case can be made for thinking that some version or other of this theory is likely to be true. The first version of the theory that I consider is mathematically simple, but this causes it to have a number of disadvantages. By contrast, the second and third versions of the theory that I discuss seek to solve these problems by taking there to be a more mathematically sophisticated function from happiness and success to welfare. Ultimately, I want to argue that the third version of the theory is superior to the other two versions. In fact, I am inclined to endorse this third version of the theory (which I call the Discount/Inflation Theory) because it seems to be able to avoid the main problems of virtually every other theory of welfare that I have considered in this dissertation. However, before getting to all this, I must begin by explaining the motivation for favoring some theory of the Happiness and Success type in the first place.

8.1 Motivation for Happiness and Success Theories

I think there is a fairly strong case to be made for thinking that some version of the Happiness and Success theory is likely to be true. The case is complicated and comprises many considerations about the relative merits of various theories of welfare. But in a nutshell, the thought is that the Happiness and Success theories seem to have the resources to avoid the big problems of most other theories of welfare that I have considered previously in this dissertation. Thus the motivation for the Happiness and Success theories depends on the results of the previous chapters. Two of these results are particularly important.

The first one, which I argued for in chapter two, is that there are major problems with both the entirely response *in*dependent theories and the entirely response *de*pendent theories. The entirely response independent theories were rejected on the grounds that they are incapable of accommodating the fact that some of our psychological responses (e.g. pleasure) are indeed relevant to determining one's welfare. The entirely response dependent theories, by contrast, were rejected on the grounds that our psychological responses do not seem to be *all* that determines welfare. This was illustrated by cases like those of Porky, Max the Masochist, Greg the grass-counter, and others. The entirely response dependent theories seemed to be incapable of capturing common intuitions about these cases. Thus, I concluded chapter 2 by claiming that there is good reason to think that the correct theory of welfare must belong to the partly response independent category.

The second crucial result of the foregoing chapters is that several of the most promising *monistic* theories of welfare in the partly response independent category face problems. In chapter 4, I discussed Objectively-Adjusted Enjoyment Theories like DAIAH. I argued that no such theory can account for the intuition that some lives that contain no enjoyment or pleasure whatsoever can still contain a positive amount of welfare value (e.g. if these lives are particularly successful ones). In chapters 5 through 7, I discussed Desire Satisfactionist theories of welfare, and I argued that Worthiness-Adjusted Cloud Desire Satisfactionism (WACDS) is the most plausible theory of the desire satisfaction type. However, I concluded by arguing that WACDS is undermined by an analog of the problem that threatens the Objectively-Adjusted Enjoyment Theories. In particular, WACDS cannot account for the intuition that some lives that contain no success with respect to worthwhile projects at all can still contain a positive amount of welfare value (e.g. if these lives are particularly pleasant).

The upshot is that although both DAIAH and WACDS are promising theories of welfare, neither one seems to capture the *whole truth* about welfare. DAIAH is problematic because there seems to be something that enhances welfare in addition to pleasure, namely some kind of objective success. WACDS is problematic because there seems to be something that enhances welfare in addition to the satisfaction of one’s worthwhile desires, namely happiness. An obvious solution would be to combine the two theories into a single multi-component theory. And doing this yields a theory of the Happiness and Success type.

8.2 A Simple Formulation of the Happiness and Success Theory

Given the respective limitations of DAIAH and WACDS, what seems to be called for is a two-component theory that takes the amount of welfare contained in one’s life to be determined, in some way or other, *both* by the amount of pleasure one takes in things during one’s life *and* by the amount of success that one has with respect to one’s worthwhile projects. In this section, I will present a simple version of the Happiness and Success Theory that is straightforwardly *additive* in a certain sense.[3] Then I will go on to draw attention to some of the less plausible features of this version of theory. In particular, these problems stem from the simple way in which the theory does the math. These problems provide the motivation for looking at more mathematically sophisticated versions of the theory, two of which I present (and one of which I endorse) in later sections.

8.2.1 Formulating the simple version

We saw in the previous section that there is some good motivation for adopting a theory that makes one's welfare be a function both of the pleasure in one's life and the net satisfaction of worthwhile desires in one's life. The simplest way in which such a theory can be constructed would be to simply take it that one's welfare is equal to the *sum* of the amount of happiness contained in one's life and the amount of success that one has with respect to worthwhile projects during one's life.[4]

Thus we would get the following theory of welfare:

(DAIAH+WACDS): The amount welfare value contained in P's life equals the sum of:

a) the net desert-adjusted attitudinal pleasure in P's life (where this is to be calculated in the way DAIAH specifies), and

[3] This is not the sense of 'additive' that for instance Velleman talks about in his well-known paper 'Well-being and Time.' (Cf. Velleman, 1991)

[4] Of course, this theory rests on the assumption that units of happiness and units of success with respect to worthwhile projects are commensurable in such a way that the former can be added to the latter. Some might be inclined to reject this assumption. However, I am not. At the very least, I think we would need *an argument* to think that it ought to be rejected. After all, there are straightforward ways in which to calibrate the units in question, so that one unit of happiness could be meaningfully compared with one unit of success with respect to worthwhile projects.

What I have in mind is an analogy with the way that the relative sizes of units of pleasure and pain are to be set. One good strategy for dealing with the incommensurability problem between pleasures and *pains* is to set the relative size of the units in such a way that a generic reasonable person would be indifferent between a) receiving one unit of pleasure and one unit of pain, and b) receiving no pleasure and no pain. A similar approach can be used, I think, to set the relative size of units of happiness and units of success with respect to worthwhile projects. In particular, the units should be set in such a way that a generic reasonable person would be indifferent between a) receiving exactly one unit of happiness and zero units of success, and b) receiving zero units of happiness and exactly one unit of success.

Because there is this obvious way of setting the relative sizes of the units, when it comes to happiness and success, it seems that the burden of proof is on those who would deny happiness can be meaningfully compared with success. Thus we would need to be given a compelling argument in order to reject this assumption. (See section 8.3.3.2 for more discussion of this point.)

b) the net amount of success that P has with respect to worthwhile projects in life (where this is to be calculated in the way that WACDS specifies).

This theory is a version of the Happiness and Success theory. DAIAH corresponds to the happiness component of welfare, while WACDS corresponds the success component.

DAIAH+WACDS does seem to have some appealing features. Since it makes one's welfare depend both on the pleasure in one's life and the success one has in one's worthwhile projects, the theory avoids the problems that DAIAH and WACDS faced when considered in isolation. DAIAH by itself was seen to be problematic because it does not allow that a life that is devoid of pleasure can still contain a positive amount of welfare. DAIAH+WACDS, by contrast, implies that this indeed is possible. For on DAIAH+WACDS, if one has some degree of success in achieving worthwhile goals, then one's life may contain a positive amount of welfare even if it contains no pleasure whatsoever. An analogous problem was seen to threaten WACDS. In particular, WACDS is problematic because it does not allow that a life devoid of success with respect to any projects can contain a positive amount of welfare. However, DAIAH+WACDS implies that this indeed is possible. For according to DAIAH+WACDS, if one experiences some pleasure in life, then one's life may contain a positive amount of welfare even if one has no success with respect to one's projects. Thus DAIAH+WACDS has some strengths that are lacked by DAIAH and WACDS by themselves.

8.2.2 Problems with the simple version

Nonetheless, DAIAH+WACDS has some unattractive features. I do not endorse it. It seems to me that the unattractive features of DAIAH+WACDS can be avoided, however, if we formulate a version of the Happiness and Success theory that does the math in a more sophisticated way. Or so I will argue in later sections.

To see the first problem with DAIAH+WACDS, consider what this theory implies about a life that contains an arbitrarily large amount of success with respect to worthwhile projects, but that contains no attitudinal pleasure (or pain) whatsoever. Is this a life high in welfare for the person who leads it? DAIAH+WACDS implies that it is. After all, the person in question here has been fantastically successful in pursuit of his worthwhile goals. However, this consequence of DAIAH+WACDS seems implausible.

The life in question is not pleasurable in the slightest. It is not directly *dis*pleasurable either, but since it contains no feelings of enjoyment for the one who leads it, the life seems cold and detached. Although I think we should take it that this life contains some small positive amount of welfare, it would be counter-intuitive to say that this is a remarkably good life for the one who lives it. Thus DAIAH+WACDS, in my view, is in trouble because of its implication that this is indeed a remarkably good life.

In other words, the problem is that, according to DAIAH+WACDS, a life containing no pleasure whatsoever can still be *arbitrarily high in welfare* value, as long as the life contains enough success with respect to worthwhile projects. This problem stems from the simple additive way in which DAIAH+WACDS does the math. DAIAH+WACDS is on the right track for taking welfare to be a function of both pleasure and success in worthwhile goals, but the theory seems to go wrong in taking this function to be simple *addition*. In later sections, we will see that this sort of problem can be avoided if we formulate a more sophisticated function that takes you from a person's happiness and success to that person's welfare.

DAIAH+WACDS faces a second problem of this same sort, though this time the issue is more complex. This second problem is based on a case that is the reverse of the one just discussed. Suppose there is a person – call him Jerry – who takes a huge amount of pleasure in things that are practically worthless. Suppose, for example, that Jerry spends the major portion of his life reading and greatly enjoying Trash Magazine, which (I am stipulating) is a source of pleasure that is as worthless as they come. Moreover, Jerry has no success whatsoever in achieving any worthwhile goals. Although he does have *some* desires – e.g. to read Trash Magazine – his desires are (we are supposing) for things that are not worthwhile. Thus according to DAIAH+WACDS the satisfaction of these desires provides no welfare benefit to Jerry. How high, then, is Jerry's life in terms of welfare?

For starters, it is clear that if one's theory simply took *quantity* of pleasure into account (understood in terms of intensity and duration only), as opposed to *desert-adjusted* pleasure, then Jerry would have an extremely good life. But this does not seem to be a plausible implication. As with the case of Porky, Jerry does *not* seem to have an astronomically high amount of welfare, even though he receives an astronomically high amount of pleasure. Nonetheless, one might think that this problem does not threaten DAIAH+WACDS because what is relevant to welfare, on this theory, is *desert-adjusted* pleasure. And since the things Jerry takes pleasure in are not very worthwhile, it seems that he does not have a whole lot of desert-adjusted pleasure. Thus one might think that DAIAH+WACDS gives the right result about the case of Jerry, viz. that Jerry does not have a life that is astronomically high in welfare value.

While I agree that taking into account specifically desert-adjusted pleasure may *help* the theory deal with this sort of problem case, it does not eliminate the problem entirely. To see why, first we need to answer this question: does the pleasure that Jerry receives from reading Trash Magazine enhance his welfare at all? It cannot be the case, it seems to me, that Jerry's pleasure has absolutely no positive impact on his welfare. It seems to me that all pleasure must enhance one's welfare at least somewhat.[5] Even in the case of Porky, it seems intuitive that his disgusting pleasures are good for him to *some* degree (though perhaps only a very small degree). I am inclined to think that no episode of pleasure can have absolutely *no* welfare benefit for the one who experiences it. To make DAIAH be consistent with this, it must be assumed that no object of pleasure can be pleasure-worthy to degree zero (even things as mind-numbing as reading Trash Magazine).

If I am right about this, then DAIAH+WACDS will still imply that a person like Jerry, who has no success to speak of in accomplishing worthwhile goals and is pleased only by practically worthless things, could still have an *arbitrarily* good life. Just imagine a string of cases involving Jerry-like people, where each person takes a progressively greater quantity of pleasure in the same nearly worthless objects (and where the amount of success in achieving worthwhile goals remains the same – i.e. zero or extremely close to it). Each life in this string of cases would be better in terms of welfare than the last. There is no principled limit to how much pleasure the possessor of one of these lives can experience, and so there is no limit to how well the life of this person can go in terms of

[5] Heathwood points out that '[i]t could be held that the objects of base desires or pleasures have a negative level of pleasure- or desire-worthiness. The theory would then have the implication that a base pleasure or satisfying a base desire actually makes a life go worse.' (Chris Heathwood, 2006, cf. p. 554) I acknowledge this as a theoretical possibility. However, I can't accept the view that pleasure taken in some objects might completely fail to enhance your welfare, or may even contribute negatively to it. I find this wholly counter-intuitive.

welfare – even though he has a vanishingly small amount of success in accomplishing worthwhile goals. Thus even DAIAH+WACDS allows that you can make Jerry's life be arbitrarily high in welfare just by making him take more and more pleasure in virtually worthless activities, like reading Trash Magazine.

This seems to be an implausible consequence. If there is a person who has no success to speak of in accomplishing worthwhile things, it would seem odd if this person's life could be made *arbitrarily good* in terms of welfare just by causing him to experience more and more low quality pleasure. I do not think that the mere addition of more and more pleasure can completely make up for utterly failing to accomplish anything worthwhile.[6] This intuition, however, is one that DAIAH+WACDS cannot accommodate.

Perhaps the present problem for DAIAH+WACDS can be brought out more forcefully if the point is made in terms of a comparison of lives. Consider a person like Jerry who has virtually no success in accomplishing anything worthwhile, but who takes a fantastically large amount of pleasure in things that are pleasure-worthy to a low degree. Thus the total amount of desert-adjusted pleasure in his life would be a very large number X (make X as large as you want), while the total amount of success in his life is something close to 0. Now compare Jerry's life with the more balanced life of Kerry. Kerry takes a moderately large amount of pleasure in exactly the same not-very pleasure-worthy things that Jerry enjoys. Thus suppose the total amount of desert-adjusted pleasure in Kerry's life is X/4. However, unlike Jerry, Kerry also has a moderately high degree of success in accomplishing worthwhile things. Suppose, in fact, that the amount of success in Kerry's life also happens to equal X/4.[7] Now, whose life is better in terms of welfare value? DAIAH+WACDS implies that Jerry's life is better than Kerry's. After all, according to this theory, Jerry has a huge amount of welfare, X, but Kerry's welfare is merely X/2. However, I would maintain that this is the wrong result. It is by no means clear that Jerry's life is *much better* than Kerry's life. For there is something important missing from Jerry's life that Kerry's life contains, namely some kind of accomplishment or successful activity. Kerry's success, it seems to me, might well make up for the

[6] Those who do not agree with this, however, may not be troubled by the line of objection to DAIAH+WADS that I am pursuing.

[7] And, again, we're supposing the units of pleasure are commensurable with the units of success. More about commensurability in section 8.3.

additional pleasure that Jerry has in his life. Since DAIAH+WACDS generates results that are inconsistent with my intuitions about this case, I find the theory to be problematic.

As was the case with the first problem, this problem, too, seems to stem from the fact that DAIAH+WACDS takes the amount of welfare in one's life to be simply the *sum* of the amount of pleasure it contains and the degree of success one has in accomplishing worthwhile goals. We will shortly see that these problems can be avoided if the theory is made to employ a more sophisticated function from pleasure and success to welfare. But first, let me briefly mention one more reason for being dissatisfied with DAIAH+WACDS.

In particular, DAIAH+WACDS seems to be awkward because it appeals to a response independent factor like desert or worthiness in two separate places. On the one hand, the DAIAH-component of the theory takes into consideration the pleasure-worthiness of the things one takes pleasure in. On the other, the WACDS-component of the theory takes into consideration the degree to which one's goals are worthwhile. This seems to me to be overkill. If possible, it would be better to have one's theory of welfare appeal to a response independent ('objective') factor like worthiness in only one place. Some think that appealing to worthiness in the first place is rather *ad hoc*,[8] and I want to minimize this concern to the extent possible by appealing to worthiness in only one place. I propose to do this by eliminating the appeal to pleasure-worthiness that figures into the DAIAH-component. That is, I propose that the contributions that episodes of pleasure make to one's welfare should not be desert-adjusted.

This amounts to replacing the DAIAH-component in the theory with an IAH-component (to be understood along the lines of 'IAH', as defined in chapter 4). Thus we would get a theory – we could call it 'IAH+WACDS' – according to which the amount of welfare contained in P's life would equal the net attitudinal pleasure contained in P's life *plus* the net success with respect to worthwhile projects in P's life. However, IAH+WACDS would face the familiar problems associated with the case of Porky (or Jerry). Porky experiences a fantastically large amount attitudinal pleasure in his life. Accordingly, even though Porky has no success in accomplishing worthwhile goals,

[8] Several philosophers have said so to me in conversation.

IAH+WACDS still implies that Porky has a fantastically good life. But intuitively his life is not fantastically good. Intuitively, Porky has at best a moderately good life. But IAH+WACDS cannot generate this result.

Nonetheless, also these problems can be avoided by formulating a theory that does the math in a more sophisticated way. That is, we can formulate a theory of welfare with two components – IAH and WACDS – that uses exclusively mathematical resources to deal with cases like that of Porky. And such a theory would have the additional advantage of appealing to a response independent factor like desert or worthiness in only one place. This, it seems to me, would be more elegant than a theory like DAIAH+WACDS, which appeals to a factor like desert or worthiness in two separate places. This concludes my explanation of the motivation for the more complex versions of the Happiness and Success Theory that I discuss in the remainder of the paper.

8.3 The Thresholds Version of the Happiness and Success Theory

The problems that seem to undermine DAIAH+WACDS can be avoided by versions of the Happiness and Success Theory that do the math in more sophisticated ways than mere addition. In this section, I will discuss one such version of the theory, and in section 8.4, I will discuss another. I will use ‘the thresholds version of the Happiness and Success Theory’ or ‘the Thresholds Theory’ to refer to the theory I present in this section. This theory takes the welfare-function – i.e. the function from a) one’s pleasure (happiness), and b) one’s success in accomplishing worthwhile goals, to c) one’s welfare – to involve thresholds in a certain way. In particular, this function ensures that it is impossible to attain a certain minimum amount of welfare unless one *both* has a sufficient amount of pleasure *and* a sufficient amount of success in accomplishing worthwhile goals.[9]

In section 8.3.1, I explain the intuitive idea behind the theory and explain how it gets around the problems for DAIAH+WACDS. In section 8.3.2, I state the theory in a more precise way, and in section 8.3.3 I discuss some of the problems facing the thresholds

[9] Thanks to Hilary Kornblith for a helpful conversation about how to conceptualize this version of the theory.

version of the Happiness and Success Theory. In Appendix II, I discuss the mathematical techniques needed to define the welfare-function that this theory employs.

8.3.1 The advantages of the use of thresholds

In order to avoid appealing to an entirely response-independent factor like desert or worthiness in more than one place, the thresholds version of the Happiness and Success Theory will not appeal to the notion of desert-adjusted pleasure. It, like IAH+WACDS, makes one's welfare be a function of two factors: the amount of attitudinal pleasure one experiences (but *not* desert-adjusted pleasure) and the amount of success one has in accomplishing worthwhile goals. However, unlike IAH+WACDS, this new theory does not do the math in a simple additive way.

Instead, this theory takes it that there is a certain minimum amount of welfare (which is greater than zero) that it is impossible to attain unless you have *both* a certain minimum amount of pleasure (happiness) *and* a certain minimum amount of success in accomplishing worthwhile goals. More specifically, suppose there is some (positive) amount of welfare that you must have in order for you to count as having a *minimally good life*. Any life containing less welfare than this is not a minimally good one.[10] Moreover, suppose that in order to obtain a minimally good life, there is a certain minimum amount of pleasure that you must have, and a certain minimum amount of accomplishment of worthwhile goals that you must have. Your life cannot be minimally good unless it contains *at least* a) this minimum amount of pleasure and b) this minimum amount of achievement of worthwhile goals. Thus if you have less than the minimum required amount of pleasure, then no matter how much achievement of worthwhile goals you have, you will not have a minimally good life. Similarly, if you don't have a certain minimum amount of achievement of worthwhile goals, then no matter how much pleasure you experience, you also will not have a minimally good life.

This is the intuitive idea behind the theory, which I will go on to state more precisely below. But even at this preliminary stage it should be apparent that a theory that is formulated along these lines will have many benefits. First, it adequately deals with the

[10] In order to formulate the theory, we do not need to know exactly where this cutoff goes, but only that there is some such cutoff point on the number line. More about this question below.

case of Porky. For although Porky's life contains much more than the minimum required amount of pleasure, it contains much less than the minimum required amount of achievement of worthwhile goals. So because it's not the case that Porky has the minimum required amount of *both* pleasure *and* achievement of worthwhile goals, he will not qualify as having a minimally good life. Thus using thresholds in this way to formulate the Happiness and Success Theory will make it yield the intuitive result that Porky does not have a fantastically good life.

For similar reasons, the thresholds version of the Happiness and Success Theory avoids the other problems faced by DAIAH+WACDS. Recall the person who had a tremendous amount of success but very little happiness. DAIAH+WACDS was problematic because it implied that such a person could have an arbitrarily good life, provided only that he has a high enough level of success. The thresholds version of the Happiness and Success Theory, however, does not give this result. Since this person's life does not contain the minimum required amount of pleasure, the theory implies that he cannot have a minimally good life – no matter how much additional success in accomplishing worthwhile goals he might have. This, I take it, is the intuitively correct judgment.

Similarly, the thresholds version of the Happiness and Success Theory generates the right result about people like Jerry. Recall that Jerry was a person who has no success to speak of in accomplishing worthwhile goals, but who takes a vast amount of pleasure in practically worthless objects (like reading Trash Magazine). DAIAH+WACDS had the problematic implication that such a person could have an *arbitrarily* good life as long as he gets enough pleasure from these practically worthless objects. By contrast, the thresholds version of the Happiness and Success Theory does not have this consequence. Since Jerry's life does not contain the minimum required amount of achievement of worthwhile goals, this theory implies that he cannot have a minimally good life – no matter how much additional pleasure he might receive. This, I take it, is also the intuitively correct judgment. And so the thresholds version of the Happiness and Success Theory seems to be preferable to DAIAH+WACDS on this score as well.

8.3.2 A more precise statement of the thresholds theory

Formulating the Happiness and Success Theory by appeal to thresholds thus seems to have many advantages. However, the thresholds theory can be spelled out in a number of ways. Here I present what I take to be (one of) the most plausible way(s) of doing it.

The Thresholds Theory takes it that there are two factors that, via a complex function involving thresholds, determine the amount of welfare contained in one's life. These factors are: a) the net amount of attitudinal pleasure – or happiness[11] – contained in one's life (as this is to be calculated according to IAH), and b) the net amount of satisfaction of worthwhile desires in one's life (as this is to be calculated according to WACDS).

A number of assumptions are required to state the theory. First of all, as with the other theories I have been discussing in this paper, what this theory tells us how to calculate is the total amount welfare value *contained in a person's life*. The theory assumes that there is in principle no upper or lower limit on how much welfare a life can contain. Similarly, the theory assumes that there is in principle no upper or lower limit on how much attitudinal pleasure or displeasure (i.e. happiness or unhappiness) your life can contain, or how much success or failure in accomplishing worthwhile goals you can have in your life. The numbers here could be positive or negative.

However, one thing that the theory assumes is that there is a certain salient cutoff point on the welfare scale. This cutoff is supposed to correspond to *the minimally good life*. Lives containing less welfare than this minimal amount do not qualify as being minimally good. In order to formulate the theory, we do not need to know exactly where this cutoff goes, but only that there is some such cutoff point on the number line. (The question of where the thresholds go will be discussed in more detail below.) Moreover, the theory also supposes that there is a salient cutoff point on the happiness scale. Call it *the minimum happiness level*. Likewise, there is a salient cutoff point on the scale that is concerned with net satisfaction of worthwhile desires. Call this *the minimum achievement level*. These cutoff points are related in the following way: your life cannot be a minimally good one unless it reaches at least a) the minimum happiness level and b) the minimum achievement level.

[11] Feldman, forthcoming, defends the view that the amount of intrinsic attitudinal pleasure in your life corresponds to the amount of happiness in your life. I accept this view as well.

For example, if the amount of success in accomplishing worthwhile projects that you have in life does not reach the minimum achievement level, then no amount of additional happiness can raise your total welfare score above what's needed to have a minimally good life. More happiness will take you *closer and closer* to the minimally good life point, but it will never take your welfare score over that point. Similarly, if the amount of happiness contained in your life does not reach the minimum happiness level, then no amount of additional success in worthwhile goals can take your welfare score above what's needed to have a minimally good life. More and more achievement may get you vanishingly close to the minimally good life cutoff, but it will never take your welfare score over that point. Thus, in order for you to have a life that is minimally good, your life must contain both the required minimum amount of happiness and the required minimum amount of achievement of worthwhile goals.

Finally, the theory also supposes that if your life contains *both* more happiness than the minimal happiness level *and* more success in accomplishing worthwhile goals than the minimum achievement level, then your welfare score should be roughly equal to the *sum* of the amounts of happiness and achievement you have in life. Similarly, if both your happiness and your achievement of worthwhile goals is *less* than the required levels – i.e. you have less happiness than the minimum happiness level and less achievement than the minimum achievement level – then your welfare score should again be roughly equal to the *sum* of the amounts of happiness and achievement you have in life.

Thus according to the sophisticated version of the Happiness and Success theory that we are considering, there is a welfare-function from the amount of happiness one has in life and the amount of success one has in accomplishing worthwhile goals in life to the amount of welfare value contained in one's life. The assumptions described in the previous paragraphs provide a number of formal conditions on how this welfare-function operates. To state the conditions formally, let me let me introduce some symbols:

- Let 'P' stand for an arbitrary person.
- Let '*h*' stand for the amount of happiness contained in P's life, as calculated by IAH.
- Let '*a*' stand for the amount of achievement (i.e. success with respect to worthwhile projects) in P's life, as calculated by WACDS.
- Let '*w*' stand for the amount of welfare contained in P's life.

- Let 'h_t' stand for the minimum cutoff point on the happiness scale - i.e. the minimum happiness level.
- Let 'a_t' stand for the minimum cutoff point on the achievement scale - i.e. the minimum achievement level.
- Let 'w_t' stand for the minimum cutoff point on the welfare scale - i.e. the minimally good life point.

The function from h and a to w should satisfy the following conditions:

1) $-\infty < h < \infty$
2) $-\infty < a < \infty$
3) $-\infty < w < \infty$
4) if either [$h < h_t$] or [$a < a_t$], then [$w < w_t$]
5) if both [$h \geq h_t$] and [$a \geq a_t$], then (roughly) [$w = h + a$]
6) if both [$h \leq h_t$] and [$a \leq a_t$], then (roughly) [$w = h + a$]
7) If [$h < h_t$], then increasing a will cause w to approach w_t, but w will never reach w_t.
8) If [$a < a_t$], then increasing h will cause w to approach w_t, but w will never reach w_t.

These are the conditions that, as far as the Thresholds theory goes, I think the welfare-function from h and a to w must meet. Condition 4) captures the idea that the welfare-function must incorporate thresholds that behave in the way described above. Conditions 7) and 8) are important because the welfare-function should capture the intuitive idea that that increasing one's score on either the happiness scale or the achievement scale will lead to an increase in one's overall welfare – even if either the minimum happiness level or the minimum achievement level is not met.

There are many ways to mathematically construct a welfare-function from h and a to w that meets conditions 1)-8). In Appendix II, I define a function that meets these conditions. However, the version of the Happiness and Success theory we are currently interested in can be formally stated even without a precise mathematical account of the welfare-function:

> The Thresholds Version of the Happiness and Success Theory: The amount welfare contained in P's life equals the value for w that is returned by the welfare-function, described by conditions 1)-8), when P's values for h and a are taken as input.

Note that conditions 1)-8) ensure that the thresholds version of the Happiness and Success Theory avoids the problems that cast doubt on the simpler, additive versions of the theory like DAIAH+WACDS. The thresholds theory does not have the problematic implication that a person who has a tremendously large amount of success in accomplishing worthwhile goals, but who experiences little or no happiness, could have an arbitrarily good life, provided only that his level of achievement is made high enough. This is ensured by condition 4). Since the life of this person does not contain enough happiness to meet the minimum happiness level, this person's life will not count as a minimally good one.

Similarly, the thresholds version of the theory does not have problematic implications about people, like Jerry or Porky, who have virtually no success in accomplishing any worthwhile goals, but who experience vast amounts of pleasure. DAIAH+WACDS had the problematic implication that as the amount of pleasure in one's life approaches infinity, one's welfare will approach infinity too – even if one accomplishes nothing worthwhile. By contrast, the thresholds theory does not have this problematic implication. Again, this is the result of condition 4). If one's life does not meet the minimum threshold when it comes to accomplishment of worthwhile goals, then the life cannot be a minimally good one – no matter how much additional pleasure the person in question might receive.

Perhaps one will worry that this way of dealing with problem cases like those of Jerry and Porky conflicts with an idea that I expressed my approval for earlier, viz. the idea that no episode of pleasure can completely fail to have a positive impact on one's welfare. However, there is in fact no such conflict. The thresholds version of the Happiness and Success theory is consistent with the idea that every episode of pleasure enhances welfare to some degree. This is ensured by conditions 7) and 8) above. For example, consider a life where h is above h_t but a is below a_t. For this life, w must be below w_t (according to condition 4)). But according to condition 7), if we keep a fixed and gradually increase h, then each increase in h will lead to an increase in w. However, w will never go above w_t. Since a remains below a_t, each increase in h will contribute less and less to w. Thus the thresholds version of the Happiness and Success Theory preserves the quite plausible idea that all additional pleasure has some positive impact on welfare.

I think that the thresholds version of the Happiness and Success theory is intuitively plausible and deserves serious consideration. Doing the math for the thresholds version of the theory is difficult, however, so I relegate the discussion of the math to Appendix II.

There I explain how a function from *h* and *a* to *w* might be constructed mathematically so as to be consistent with conditions 1)-8). (This allows us to provide a graphical representation of the welfare function as well.)

8.3.3 Objections to the thresholds version of the Happiness and Success Theory

At least five objections can be raised against the thresholds version of the Happiness and Success Theory. I discuss them in what I take to be descending order of seriousness.

8.3.3.1 Where do we set the thresholds?

Perhaps the most glaring challenge that the Thresholds theory must deal with is the question of where to set the thresholds it appeals to. The minimally good life point, w_t, is fairly easy to deal with. I have just been assuming that w_t equals the sum of the happiness threshold, h_t, and the achievement threshold, a_t. But this is not much progress. It just means that the real question is where to set the happiness threshold and the achievement threshold.

This is a difficult question, but it is not unanswerable, I think. It is a substantive evaluative question, and so we must appeal to the same methods we use to answer other substantive questions. In practice, this is going to be the method of Reflective Equilibrium. We need to find values for the happiness threshold and the success threshold such that, when plugged into the Thresholds version of the Happiness and Success theory, the theory gives intuitively plausible results – results that we find ourselves able to live with.

I do not have room to give a sufficiently detailed treatment of this question here, but I can say a few basic things. In particular, I take it that both the happiness threshold and the achievement threshold must be set somewhere not too high above the zero point on the relevant scale. The happiness threshold is supposed to represent the point at which a person can be said to have a minimally happy life. Intuitively, this requires not a whole lot of happiness, but it does require some. While an unhappy life is one that has a negative amount of net happiness, and a happy life, strictly speaking, is one that contains a positive amount of net happiness, the minimally happy life is a life that is at least 'good enough' in terms of happiness. It is a life that is at least minimally respectable when it comes to happiness. Similarly with achievement. The achievement threshold is supposed to represent the point at which a person can be said to have a minimally successful life. This is a life that contains a satisfactory or respectable amount success in achieving worthwhile goals, but not an impressive or overwhelming amount. Intuitively, a minimally successful life does not require a large amount of achievement, but it does require some moderate positive amount of achievement.

This is all I am able to say about this difficult question at present. More investigation is needed before a more satisfactory answer can be given. But I see no reason to think such an answer cannot be given.

8.3.3.2 How should the units be scaled?

Another difficult challenge in developing the thresholds version of the Happiness and Success Theory concerns the question of the relative value of happiness and success with respect to worthwhile projects. The Thresholds theory assumes that the amount of (net) happiness in your life can be represented by a number located in the interval between negative infinity and positive infinity. Likewise for the degree of success you have with respect to worthwhile projects. (See conditions 1 and 2 above.) But now the question arises as to how these two scales are related to each other. Consider one unit of happiness (i.e. the amount of happiness by which a life with a happiness level of, say, 1000 differs from another life with a happiness level of 1001). And consider one unit of success with respect to worthwhile projects. How much is this one unit of happiness worth, in terms of welfare, compared to one unit of success? Are a unit of happiness and a unit of success worth the same amount of welfare? Is a unit of one worth more than a unit of the other?

The first thing to notice is that Thresholds theory implies that the answer to this question will vary depending on whether one's happiness and one's success are above threshold or not. If one's happiness and one's success are both above threshold or are both below threshold, then the amount of welfare one would get per unit of happiness will equal 1 and the amount of welfare one would get per unit of success will equal 1. However, if one of the two variables is above threshold and the other is below, then things will be different. For instance, if one's happiness is above threshold and one's success is below, then the amount of welfare one would get per unit of happiness will be less than one.

However, this tells us only how the welfare value of a unit of happiness varies as a function of success, and how the welfare value of a unit of success varies as a function of happiness. We still do not have an answer to the underlying question of how the size of a unit on the happiness scale and the size of a unit on the success scale are to be fixed in the first place.

At first glance, one might think this problem can be brushed off easily. After all, it doesn't really matter *what* the size of the units are with which we measure anything. There is no mathematical reason why it is in principle better to measure my height in, say, feet than in meters (or in light-years, for that matter). We just need to settle on a convenient system of measurement so as to avoid confusion when talking to and working with others. However, the problem cannot be brushed aside so easily. For even if we do come up with some convenient system of measurement for happiness and some convenient system of measurement for achievement, the question still remains as to how the units in the one system are related to the units in the other system. We need to pick a system of measurement for each that allows for meaningful comparisons.

I think there is a natural way to go about answering this question, however. More work may need to be done to fully work out what the answer *is*, but it seems to me that we at least have some idea about how to proceed. To begin with, consider with an analogy. A Utilitarian might attempt to fix the relative value of pleasure and pain by assuming that one unit of pleasure and one unit of pain must be related in the following way: a rational person would be indifferent between a) receiving exactly one unit of pleasure and one unit of pain, and b) receiving no pleasure and no pain.[12] It doesn't matter what sizes a unit of pleasure and a unit of pain have, as long as one unit of the former is the exact amount it would take to adequately compensate for one unit of the latter.

It would be natural to think that we can take a similar approach when it comes to fixing the relative size of the units of happiness and the units of success. In particular, we

[12] That is, a rational person would have to be indifferent between preventing the receipt of exactly one unit of pain and receiving exactly one additional unit of pleasure (or, what amounts to the same thing: between receiving exactly one unit of pain and losing exactly one unit of pleasure).

can assume that one unit of happiness and one unit of success should be related in the following way: *a rational person would be indifferent between a) a scenario in which she receives exactly one (additional) unit of happiness and no (additional) units of success, and b) a scenario in which she receives exactly one (additional) unit of success and no (additional) units of happiness.* It doesn't matter what sizes a unit of happiness and a unit of pain have, as long as one unit of the former is the exact amount it would be rational to be indifferent about exchanging for a unit of the latter.[13]

The job is not done yet, however. After all, given a particular amount of happiness, X, and a particular amount of success, Y, how do we determine whether the rational person would be indifferent between X and Y? In the same way, I suggest, that we answer other substantive moral questions: by using the method of Reflective Equilibrium. That is, we must collect a representative sample of our intuitive judgments about various kinds of tradeoffs involving happiness and success, and then work towards finding a substantive theory about how much happiness it would be rational to trade for how much success that preserves our intuitions to the greatest extent possible. Once we arrive at such a theory with which we find ourselves in reflective equilibrium, then finally we will be in a position to fix the units on the happiness scale and the units on the success scale in such a way that it's rational to be indifferent between one unit of the former and one unit of the latter.

This is a long and hard process, to be sure. But I see no reason to think it can't in principle be done. A significant amount of work would be needed to actually fix the units of happiness and success in the way I've sketched, but I do not think it is an impossible job. It is no harder than answering other substantive evaluative questions, and we are not in general pessimistic about finding answers to those. Thus in order for the question of how to fix the relevant units of measurement to provide reason to *reject* the Thresholds theory, it seems we would need an *argument* for the idea that the process I described cannot be carried out in principle. I know of no such argument, however. And so I think

[13] Of course, this proposal would become circular if 'a rational person' were taken to be simply someone who, when all else is equal (e.g. when no others are affected), always prefers the option that would maximize her welfare, and who is indifferent between options that would have the same impact on her welfare. So that is not how 'a rational person' should be understood. An independent account of this notion would have to be given.

that the question of how to fix the units of measurement for happiness and success, at the very least, does not provide a knock-down objection to the Thresholds theory.[14]

8.3.3.3 Strange mathematical behavior

The previous problem is going to be a challenge that must be faced for any version of the Happiness and Success theory. The next problem is one that threatens specifically the Thresholds version of the theory. Because the theory employs thresholds in the way that it does, it can be expected to display some rather odd behavior under certain circumstances.

In particular, when you are above threshold on one scale but just below threshold on the other scale, then it will be possible for very small increases on the latter scale – increases just large enough to get you over threshold on that scale – to lead to dramatic increases in your welfare. Take an example. Suppose you are very high above threshold with respect to happiness, but just under threshold with respect to achievement. Now suppose we add just a tiny bit of achievement to your life, that is, just enough to get you to reach the achievement threshold. According to the Thresholds theory, this will cause a huge jump in your overall welfare. (For a graphical representation of this, see Fig. 8 in Appendix II.)

But perhaps this seems odd. It might seem contrived or arbitrary. However, while I grant that this is a strange consequence of the Thresholds version of the theory, it seems to me to be an unavoidable result of how the theory works mathematically. Nonetheless, there are other versions of the Happiness and Success Theory that do not allow small increases in either happiness or achievement to cause this sort of surprising jump in overall welfare. In particular, the theory I present in section 8.4 seems to avoid this problem.

8.3.3.4 Return of the experience machine problem?

One might wonder if the thresholds theory, since it is formulated in terms of IAH instead of DAIAH, falls prey to a particular version of the experience machine argument.

[14] Nor, for that matter, do I think that this worry about fixing the units of measurement provides a knock-down argument against the Discount/Inflation version of the Happiness and Success theory, which I go on to present in section 8.4.

In chapter 2, I appealed to experience machine arguments in order to reject a certain sub-group of the entirely response dependent theories, viz. the mental state theories. For a pair of internally indistinguishable lives whose only difference is that the one is lived in the real world, while the other is lived in the experience machine, the mental state theories will imply that these two lives contain the same amount of welfare. Since I find this to be a problematic implication, I am inclined to reject the mental state theories. One mental state theory is IAH. Accordingly, IAH will run into problems with experience machine scenarios too. By contrast, DAIAH is not a mental state theory and it has the resources to avoid the problems raised by experience machine scenarios. However, I appealed to the machinery of IAH, not DAIAH, to formulate the thresholds version of the Happiness and Success Theory. Thus, one might think that the thresholds version of the Happiness and Success Theory is vulnerable to experience machine objections.

In particular, the objection would go like this. Imagine a person in the real world who experiences a certain amount, X, of attitudinal pleasure that is above the happiness threshold h_t. What's more, this person has no success whatsoever in completing any worthwhile projects, and so his achievement level is below the achievement threshold a_t. Now imagine a second person who has an identical life in the experience machine. He too feels attitudinal pleasure to degree X and has no success whatsoever in completing any worthwhile projects. The thresholds version of the Happiness and Success Theory might seem to imply that these two lives contain the same amount of welfare. In particular, they both have a level of welfare that is below the minimally good life point. Neither person has a minimally good life. However, one might think this consequence is implausible. Since these two people have identical lives except for the fact that one life takes place in the real world, while the other is simulated in the experience machine, the first person should have a life that is at least somewhat higher in welfare. Thus the thresholds version of the Happiness and Success Theory is mistaken.

I do not find this to be a very worrying objection to the thresholds version of the Happiness and Success Theory. For the scenario required to generate the problem is so bizarre that it does not need to be taken very seriously, I think. Here is why. In order to generate the problem, the two people must literally have *no* success in achieving worthwhile goals at all. (That is, their values for *a* would have to be equal to zero.) But this could virtually never happen. Chances are that both people in question will have desires that certain things *actually* befall them, as opposed to merely *appear* to do so – e.g. that the person eats an apple (not a simulacrum of an apple), or that the person beats his neighbor at checkers (as opposed to a simulacrum of his neighbor), etc. Since the person who lives his life in the real world will have more of these desires satisfied than his counterpart in the experience machine, he will have more success with respect to his projects (i.e. more of the desires in his various desire-clouds will be satisfied) than his counterpart. As a result, the thresholds version of the Happiness and Success Theory will imply that the first person has somewhat higher amount of welfare (though it will presumably still be below w_t) than the second person. Thus the WACDS component of the Happiness and Success Theory allows it to circumvent the problems associated with all but the most bizarre experience machine scenarios.

However, could there be a version of the case that still causes problems for the Happiness and Success Theory? In particular, suppose there are two identical lives, one in the real world and one in the experience machine, where the person in each one has just one desire: the desire that it *seem to him* as though he eats an apple. Suppose the person in the real world actually eats an apple and it gives him a sizable amount of pleasure. Suppose the person in the experience machine has the experience as of eating an apple and it gives him the same sizable amount of pleasure. I admit that if the case is like this, then the thresholds version of the Happiness and Success Theory does imply that the two lives contain exactly the same (below threshold) amount of welfare. Each person gets the same (above threshold) amount of pleasure, and each person has the same (below threshold) amount of success in accomplishing worthwhile goals. However, now I lose the intuition that the person who lives the real life is better off in terms of welfare than the person who lives the life in the experience machine. Since they both aspire to merely have certain things merely appear to them to be the case (as opposed to really be the case), I am not sure why we should think the guy in the real world has a better life than the guy inside the experience machine. Thus I don't think the thresholds theory needs to accommodate the 'intuition' that the *real* life in question here contains more welfare than the *simulated* life.

8.3.3.5 Enjoyment of the excellent is best

A final objection to the thresholds version of the theory is that it cannot accommodate the idea that the best thing is to *enjoy the worthwhile projects* you are pursuing. Robert Adams, for instance, expresses this idea as follows:

> what is good for a person is a *life* characterized by *enjoyment of the excellent*. More precisely … the principal thing that can be noninstrumentally good for a person is a life that is hers, and … two criteria (perhaps not the only criteria) for a life being a good one for a person are that she should enjoy it, and that what she enjoys should be, in some objective sense, excellent. (Adams, 1999, p. 52.)

Derek Parfit seems to endorse a similar idea (though he puts the point in terms of 'wanting worthwhile things' rather than 'taking pleasure in worthwhile things').

> Pleasure with many other kinds of object has no value. And, if they are entirely devoid of pleasure, there is no value in knowledge, rational activity, love, or the awareness of beauty. What is of value, or is good for someone, is to have both; to be engaged in these activities, and to be strongly wanting to be so engaged. (…) We might claim, for example, that what is good or bad for someone is to have knowledge, to be engaged in rational activity, to experience mutual love, and to be aware of beauty, while strongly wanting just these things. (Parfit, 1984, p. 502.)

Darwall seems to throw his support behind such an idea as well:

> The normative claim I shall defend is that the best life for a person (in terms of welfare) is one involving activities that bring her into an appreciative rapport with various forms of agent-neutral value, such as beauty, the worth of living things, and so on. (Darwall, 2002, p. 17)

Thus the thought here seems to be that the most "powerful" source of welfare value is enjoying (or appreciating, or what have you) things that are worthwhile in themselves. More precisely, while enjoyment on its own may be good for you, to some extent, and while achieving things that are worthwhile may be good for you on *its* own as well, to some extent, what is better by far is to receive enjoyment *directly from* the worthwhile activities you engage in.

If you like this thought, you will have reason to be dissatisfied with the thresholds version of the Happiness and Success Theory. For this theory makes enjoyment and the achievement of worthwhile goals be *independent* sources of welfare value. Here is a way to bring out the potential problem. Consider a person, Jack, who takes some large amount of pleasure, X, in reading *Trash Magazine*. Moreover, suppose this person has some large degree of success, Y, in achieving worthwhile goals. But he is not pleased by his pursuit or attainment of these worthwhile goals in the least. Now compare this person with a second person, Jill. This person, too, receives an amount of pleasure that equals X and is successful to degree Y in achieving things that are worthwhile. However, Jill receives her pleasure *directly from* her pursuit and attainment of these worthwhile goals. The thresholds version of the Happiness and Success Theory implies that Jack has just as good a life as Jill. However, people who agree with the sentiment that Adams, Parfit and Darwall endorse would presumably not be comfortable with this implication. They would be likely to think that Jill has more welfare than Jack.

I grant that the thresholds version of the Happiness and Success Theory has this implication, which some may find to be problematic. However, I am willing to bite the bullet on this one. For one thing, I would not place much significance on the intuition that Jack (who enjoys to degree X reading Trash Magazine and is successful to degree Y) has less welfare than Jill (who enjoys to degree X her pursuit of worthwhile goals and is successful to degree Y in attaining these goals). It seems to me that the welfare value contained in these two lives is roughly the same. Perhaps if pressed, I would say that Jill's life is slightly better than Jack's in terms of welfare, but if so it's only by a small amount. Thus I would be willing to give up the offending intuition for theoretical reasons (which the method of reflective equilibrium allows us to do at times, at least with non-central intuitions).

However, there is another more weighty line of response available to the present objection. To say that a life like Jill's is better than a life like Jack's is simply to endorse the view that certain objects make for more welfare enhancing enjoyment than certain other objects do. But earlier (in section 8.2.2) I discussed the motivation for moving away from this kind of desert-adjustment of episodes enjoyment. Most importantly, considerations of simplicity suggested that we try to formulate our theory of welfare in such a way that it does not appeal to response-independent (objective) factors like desert-adjustment in more than one place. And so I proceeded to formulate the thresholds version of the Happiness and Success Theory in terms of IAH, not DAIAH. Now, in light of this new problem, we might want to go back on this. Thus we could formulate a version of the thresholds theory that proceeds in terms of DAIAH, not IAH. (In particular, see the definition of '*h*' in the statement of the theory, in section 8.3.2.) Doing this would allow us to capture the intuition that Jill's life is better than Jack's. So the thresholds theory could be modified to avoid the above consequence, which people like Adams, Parfit and Darwall may not like. Nonetheless, for the reasons of theoretical simplicity mentioned earlier, I prefer to bite the bullet and say that Jack's life and Jill's life are just as good in terms of welfare than to abandon my original formulation of the thresholds theory (which proceeds in terms of IAH, not DAIAH).

So we see that there are several problems for the thresholds version of the Happiness and Success theory. Some of them are answerable; others require more work to deal with adequately. In the next section I present a different version of the theory that seems fare better than the thresholds theory, at least with respect to some of these problems.

8.4. The Discount/Inflation Version of the Happiness and Success Theory

The Thresholds theory is not the only way in which a Happiness and Success theory can plausibly be developed. In this section, I will present another version of the Happiness and Success Theory, which has the same advantages as the Thresholds version, but which might be *even more* attractive in certain respects. For one thing, the version of the theory I present in this chapter does not appeal to the concept of 'a minimally good life.' On this new theory, there is no absolute cutoff point on the overall welfare scale. As a result, this version of the theory seems to require making fewer postulates than the thresholds version of the theory. An additional advantage of this theory, as we will see, is that it does not display some of the odd mathematical behavior that the Thresholds theory did. While the Thresholds theory allowed that a small increase in, say, happiness from just below threshold to just above could lead to a huge jump in your overall welfare. This seems like an odd consequence. But the new theory I present in this section does not entail the possibility that very small increases in either happiness or achievement can lead to such disproportionate jumps in overall welfare. So this is a second respect in which this new theory is preferable to the Thresholds theory.

8.4.1 The intuitive idea behind the theory

I call the theory to be presented in this section *the Discount/Inflation Version of the Happiness and Success Theory.*[15] The basic idea behind this theory is this. On the one hand, the degree to which your happiness contributes to your welfare is determined in

[15] Thanks to John Arthur Skard for helpful conversations about how to formulate this version of the theory.

part by the amount of success you have in achieving worthwhile goals. On the other hand, the degree to which your success in achieving worthwhile goals contributes to your welfare is determined in part by the amount of happiness you receive in life. However, the reason for this is not that the theory takes the welfare value of your life merely to be the result of multiplying the amount of happiness you get in life by the amount of success you have in life. (This would be a Multiplicative Version of the Happiness and Success Theory.) Rather, the Discount/Inflation theory takes it that your happiness makes an *independent* contribution to your welfare and that your achievement makes an *independent* contribution to your welfare.[16] However, the magnitude of the happiness-contribution to your welfare is in part a function of how much success in worthwhile projects you have, just as the magnitude of the achievement-contribution to your welfare is in part a function of how much happiness you experience.

More precisely, here is how it works. On the Discount/Inflation Version of the theory, your net happiness contributes a certain amount to your welfare – call this the *happiness-contribution* – and your success in achieving worthwhile goals contributes a certain amount to your welfare – call this the *achievement-contribution*.[17] The amount of welfare value in your life simply equals the happiness-contribution plus the achievement-contribution. But how are the sizes of the happiness-contribution and the achievement-contribution to be determined?

Let's focus on the happiness-contribution first. The happiness-contribution to your welfare is a function of two things, the raw amount of (net) happiness you experience in life and the raw amount of achievement you have in life. Your raw happiness equals the number of hedons minus dolors (of attitudinal pleasure) you get in life, while your raw achievement reflects the degree to which you are successful in achieving worthwhile goals in life (as determined by WACDS). We must also assume that there is a certain

[16] Here is what I mean by this. On the Discount/Inflation Theory, you can have zero happiness and still have a positive amount of welfare provided you have a positive amount of success, and you can have zero success and still have a positive amount of welfare provided you have a positive amount of happiness. This would not be possible on the Multiplication Version of the theory. Nor would it be possible on an adjustment theory like DAIAH. So here we have an advantage of the Discount/Inflation theory and the Thresholds theory.

[17] It is important to remember that the theory is supposed to tell us how to calculate the welfare value of a life after it is completed. Thus what we are interested in is the total amount of (net) happiness that you had in life and the total amount of success that you had in life – after it is over. These two things are what together determine the total amount of welfare value of your life – after it is over.

threshold on the achievement scale. This threshold represents the level at which a life can be said to be minimally successful. It is located somewhere above the zero point, but to see how the theory works it is not necessary to know exactly where. Now (assuming your raw happiness is positive, i.e. that you get more happiness than unhappiness in life) there are three cases. 1) What happens to the happiness-contribution if your raw achievement is *above* threshold? 2) What happens to the happiness-contribution if your raw achievement is below threshold? 3) What happens to the happiness-contribution if your raw achievement is *at* threshold?

Begin with the first case. If your raw achievement score is below threshold (i.e. you don't have even a minimally successful life), then the welfare value of the raw amount of happiness you have in your life gets progressively **discounted**. That is, for a given achievement level that is below threshold, increasing your raw happiness gives you progressively less benefit in terms of welfare. And as your raw achievement level is decreased farther and farther below threshold, then the welfare value of your happiness gets discounted more severely. But no matter how low your level of achievement is, it will never be the case that more happiness will give you *zero* benefit in terms of welfare. More happiness is always better; it is just that when your achievement is below threshold, more happiness does less and less for you. A simple diagram will help illustrate this:

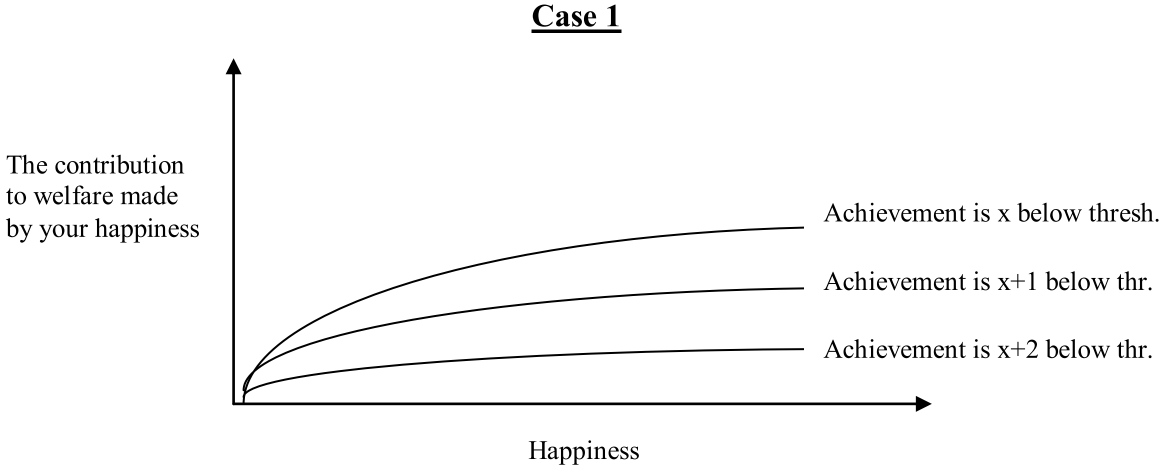

Figure 2: Discounting

Thus when we plot the welfare value of the happiness contained in your life, then as we increase the amount of happiness in your life, the derivative of the curve asymptotically approaches zero. This is what happens when your achievement level is below threshold.

Now for the second case. If your raw achievement score is above threshold (i.e. you have a life that is more than minimally successful), then the welfare value of the raw amount of happiness you have in your life gets progressively **inflated**. That is, for a given achievement level that is above threshold, increasing your raw happiness gives you progressively more benefit in terms of welfare. And as your raw achievement level is increased farther and farther above threshold, then the welfare value of your happiness gets inflated more dramatically. In other words, as your raw achievement level is increased above threshold, the amount of welfare you get for each additional unit of raw happiness gets larger and larger. Again, a diagram will help illustrate:

Case 2

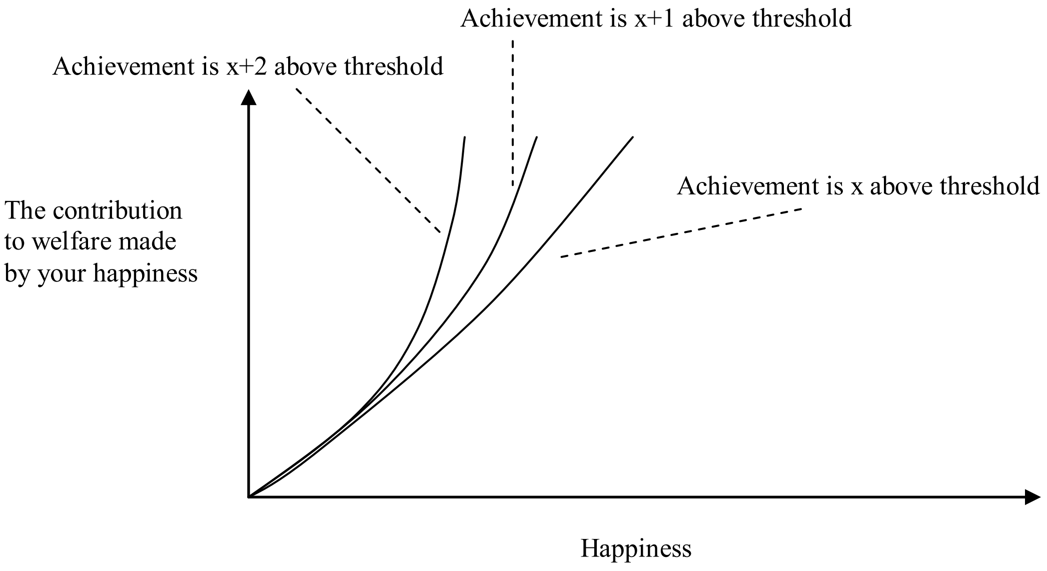

Figure 3: Inflation

So when your achievement level is above threshold, then as we increase the amount of happiness in your life, the derivative of the curve increases.[18]

[18] When achievement is above threshold, is there a limit to how inflated the welfare contribution of your happiness can get? More precisely, is there a limit to how high the derivative of the curve in this graph can get? You might think so since in Case 1, there was a limit to how low the derivative of the curve can go,

Finally, the third case. If your raw achievement is exactly *at* threshold, then the contribution to your welfare made by your happiness gets neither discounted nor inflated. In other words, no matter how much raw happiness you have in life, you get the same amount of welfare for each additional unit of raw happiness. A diagram to represent this:

Case 3

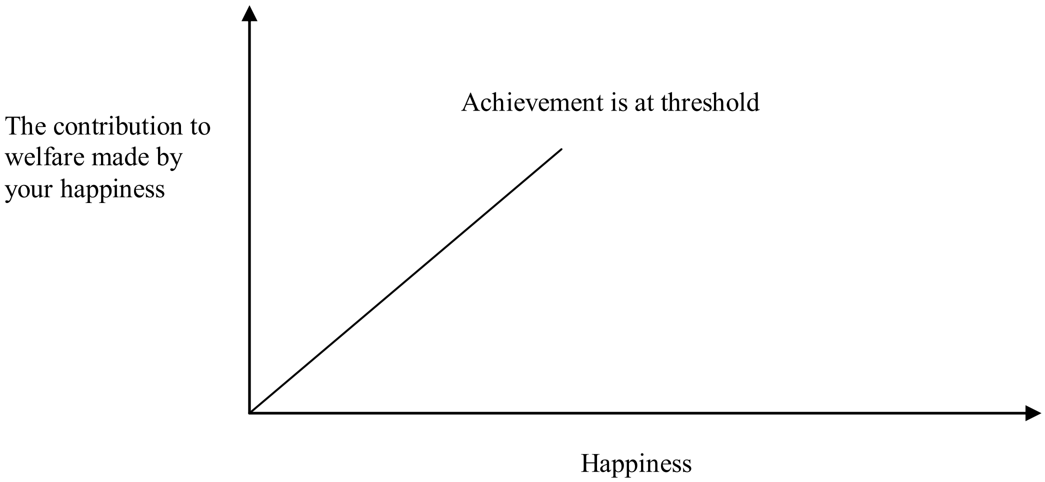

Figure 4: No Discount or Inflation

So when your achievement level is *at* threshold, then as we increase the amount of happiness in your life, the derivative stays the same.

These are the three cases that we encounter when the amount of raw (net) happiness you have in life is *positive*, i.e. when you have more happiness in life than unhappiness. But how do things work when your raw (net) happiness is *negative*, i.e. when you have more unhappiness in life than happiness? Again we encounter three cases.

To begin with, what happens when your raw happiness is negative and your raw achievement is *below* threshold? Well, the thought in general is that things go worse for you when your achievement level is below threshold. When your raw achievement is below threshold, this has an adverse effect on the contribution of happiness to your welfare. So when your raw happiness is negative and your raw achievement is below threshold, the result is that as your raw unhappiness is decreased (approaches minus infinity), you get hurt progressively more in terms of welfare. When your achievement is

viz. zero. That is, there was a limit to how much the welfare contribution of your happiness can be discounted. I will have more to say about this in the next section. In particular, see the discussion of condition v) below.

below threshold, more net unhappiness in your life will detract progressively greater amounts from your welfare. Visually represented:

Case 4

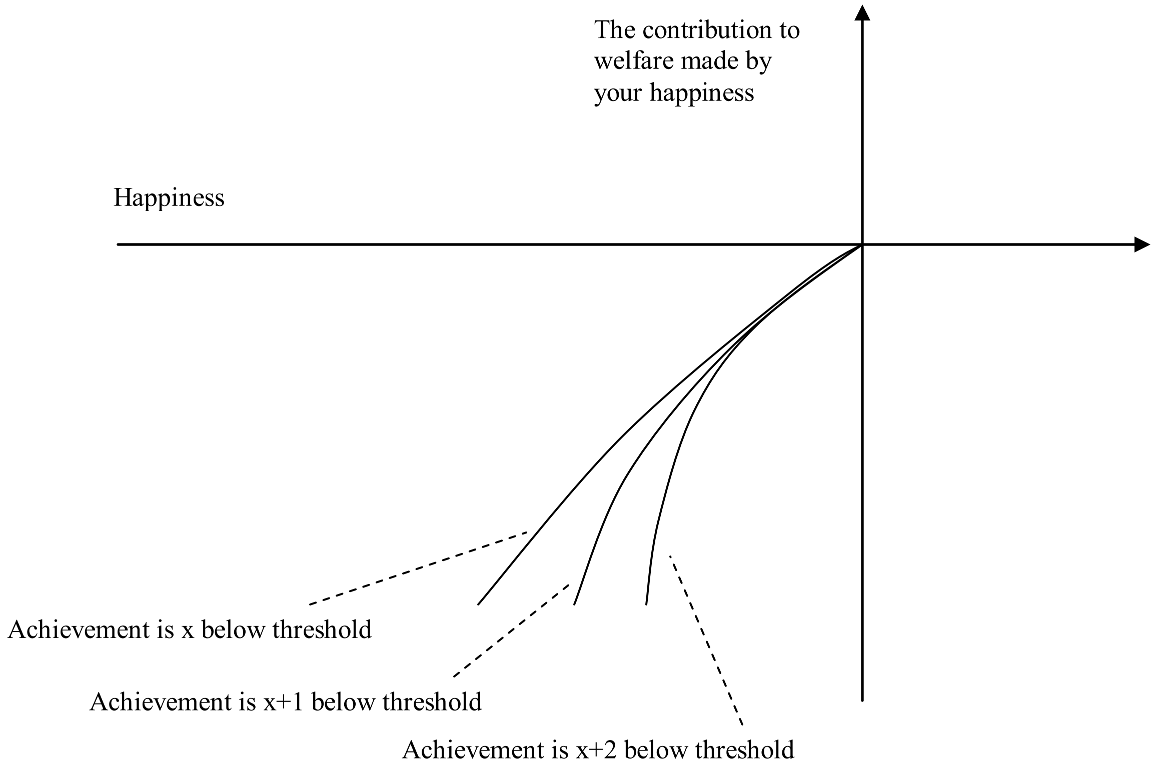
The contribution to welfare made by your happiness
Happiness
Achievement is x below threshold
Achievement is x+1 below threshold
Achievement is x+2 below threshold

Figure 5: Discounting

So when your achievement level is below threshold, then as we decrease the amount of happiness in your life farther and farther below the zero point, the derivative of the curve decreases as well.

Next, what happens when your raw happiness is negative and your raw achievement is *above* threshold? The thought in general is that things go better for you when your achievement level is above threshold. So when your raw achievement is above threshold, this has a beneficial, mitigating effect on the contribution of your unhappiness to your welfare. When your raw happiness is negative and your raw achievement is above threshold, the result is that you get hurt progressively less in terms of welfare for each additional unit of raw unhappiness you rack up in life. When your achievement is above threshold, more net unhappiness in your life detracts progressively smaller amounts from your welfare. But no matter how high your level of achievement is, it will never be the case that more unhappiness in life will cause you zero harm in terms of welfare. More unhappiness is always worse for you in terms of welfare. It is just that when your success is above threshold, more unhappiness hurts you less and less. Represented visually:

<u>Case 5</u>

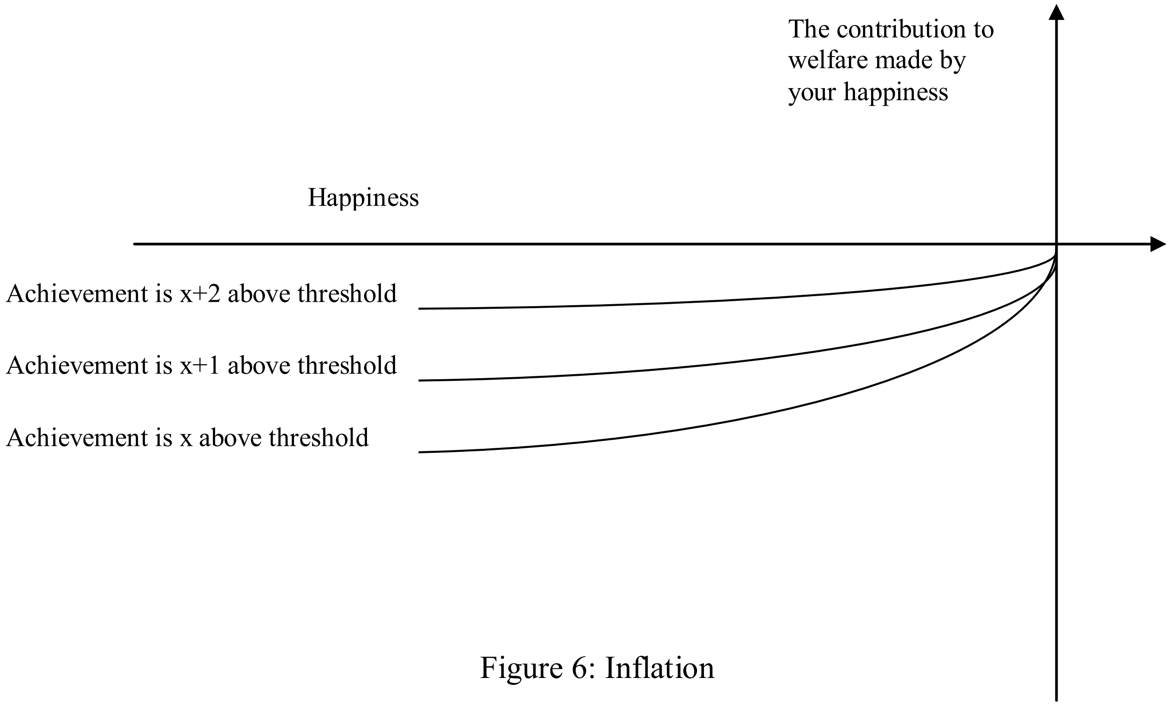
The contribution to welfare made by your happiness
Happiness
Achievement is x+2 above threshold
Achievement is x+1 above threshold
Achievement is x above threshold

Figure 6: Inflation

When your achievement level is above threshold, then as we decrease the amount of happiness in your life farther and farther below the zero point, the derivative of the curve approaches zero.

Finally, what happens when your raw happiness is negative and your raw achievement is *at* threshold? Just as before, the contribution to your welfare made by your net unhappiness gets neither discounted nor inflated. In other words, no matter how much net unhappiness you have in life, you get the same amount of welfare for each additional unit of unhappiness. Represented visually:

Case 6

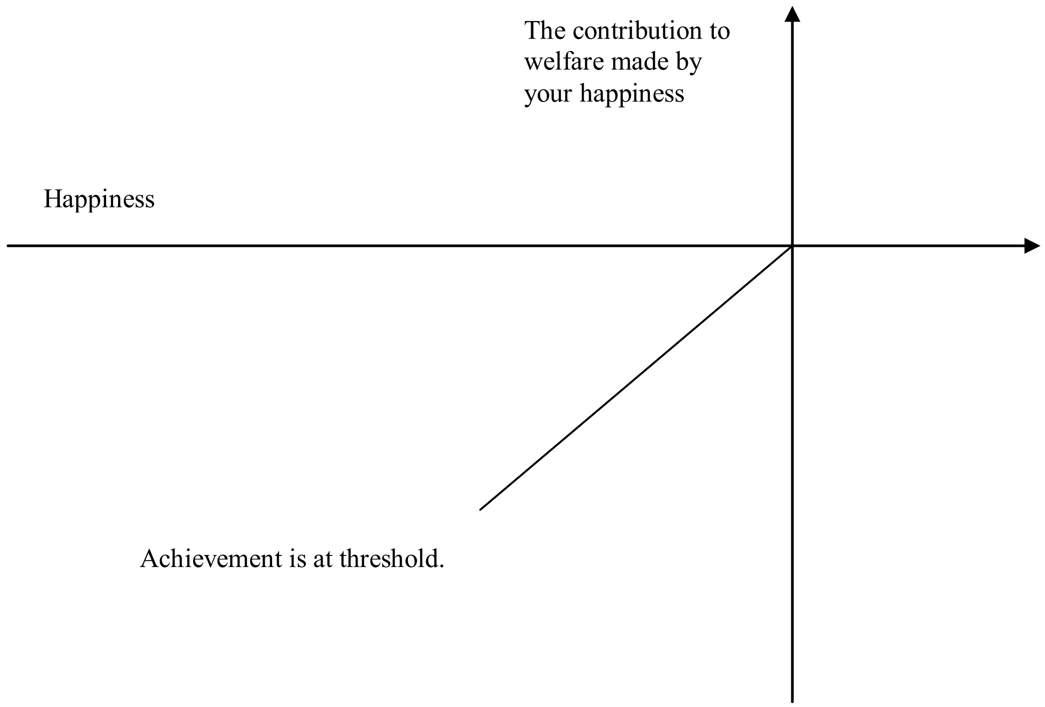
The contribution to welfare made by your happiness
Happiness
Achievement is at threshold.

Figure 7: No Discount or Inflation

As we increase the amount of net unhappiness in your life, the derivative stays the same. This is what happens when your achievement level is *at* threshold.

To this point, I have been sketching only how the size of the happiness-contribution to your welfare depends on raw achievement. But this is only half the story. We must also look at how the size of the *achievement-contribution* to welfare depends on raw happiness. This is straightforward, however, because things work in just the same way as with the happiness-contribution. Just as there is an achievement-threshold that is relevant for determining the size of the happiness-contribution to welfare, so too is there a happiness-threshold that is relevant for determining the size of the achievement-contribution to welfare. This threshold on the achievement scale represents the level at which a life can be said to be minimally successful. It is located somewhere above the zero point, but to understand how the theory works it is not necessary to know exactly where.

Now, just as before, we have a total of six cases – three for when raw achievement is positive and three for when raw achievement is negative. 1) When raw achievement is positive and raw happiness is *below* threshold, the result is that you will get less and less welfare out of each additional unit of achievement in your life. That is, the welfare value of your achievement gets progressively **discounted**. 2) When raw achievement is positive and raw happiness is *above* threshold, the result is that you will get more and more welfare out of each additional unit of achievement in your life. That is, the welfare value of your achievement will get progressively **inflated**. 3) When your raw achievement is positive and your raw happiness is *at* threshold, the result is that for each additional unit of achievement, you will receive exactly the same amount of additional welfare. 4) When your raw achievement is *negative*, having a raw amount of happiness that is *below* threshold will have an adverse effect. That is, the result is that you get hurt progressively *more* in terms of welfare for each additional unit of raw (net) unsuccessfulness (failure) you rack up in life. 5) When your raw achievement is negative, having a raw amount of happiness that is *above* threshold will have a mitigating effect. The result is that you get hurt progressively less in terms of welfare for each additional unit of raw (net) unsuccessfulness (failure) you rack up in life. 6) When your raw achievement is negative, having a raw happiness that is *at* threshold has neither an adverse nor a mitigating effect. The result is that the degree to which you are hurt by additional units of raw (net) unsuccessfulness (failure) is constant. No matter how much net unsuccessfulness (failure) you have in life, if your raw happiness is at threshold, each unit of your unsuccessfulness (failure) will hurt your welfare by the same amount.

So far I have been explaining in general terms how the size of the happiness contribution to welfare varies as a function of your achievement, and how the size of the achievement contribution to welfare varies as a function of your happiness. At this point an example would be helpful to illustrate how the welfare value of a life as a whole is to be calculated, according to the Discount/Inflation theory.

Consider the case of Porky. Recall that the Thresholds Version of the Happiness and Success Theory had an advantage over DAIAH+WADS (i.e. the additive version of the theory), in that the former can adequately accommodate intuitions about the case of Porky, while the latter cannot. The Discount/Inflation Version has this same advantage as well. Porky's life is one that is unusually high in happiness, but unusually low in achievement of worthwhile goals. To keep things simple, suppose Porky's level of achievement is just barely above the zero point. Thus his raw happiness score is above threshold, while his achievement score is below threshold. What contribution do his happiness and his achievement make to his welfare? On the one hand, since his level of achievement is below threshold, the contribution to his welfare made by his happiness will get *discounted*. Thus, the large amount of happiness he receives will not provide very much welfare for him. On the other hand, since his level of happiness is above threshold, the contribution to his welfare made by his achievement will be *inflated*. But Porky has next to zero achievement in the first place. So this inflation effect remains minor. Accordingly, the tiny amount of achievement he has makes a small positive contribution to his overall welfare. Now, to find the overall amount of welfare contained in Porky's life, we must add up the happiness-contribution, which is small, and the achievement -contribution, which is also small. When we do this, the result is that Porky's life contains only a small (or at best a mediocre) amount of welfare. And this, I have been assuming, is the intuitive result about the Porky case. Despite all his sensory pleasure, it should be intuitive that Porky does not have a fantastically good life in terms of welfare. And so the Discount/Inflation theory properly accommodates intuitions about the Porky case.

This example, then, illustrates how the welfare value contained in a life is determined according to the Discount/Inflation version of the Happiness and Success theory. But we do not yet have a complete, precise statement of the theory. This is the job to which I now turn.

8.4.2 A more precise statement of the theory

To provide a precise formulation of the Discount/Inflation version of the theory, let me begin with the terminology it employs. (Most of this was already introduced in formulating the Thresholds version of the theory.)

- Let 'P' stand for an arbitrary person.
- Let 'h' stand for the amount of raw happiness contained in P's life, as calculated by IAH.
- Let 'a' stand for the amount of raw achievement (i.e. the amount of success with respect to worthwhile projects) in P's life, as calculated by WACDS.
- Let 'W_h' stand for the contribution that P's raw happiness makes to P's welfare.
- Let 'W_a' stand for the contribution that P's raw achievement makes to P's welfare.
- Let 'w' stand for the amount of welfare contained in P's life.

- Let ‘h_t’ stand for the happiness threshold (the minimum happiness level).
- Let ‘a_t’ stand for the achievement threshold (the minimum achievement level).

Now I can begin stating the Discount/Inflation theory itself. The amount of welfare contained in P’s life equals the sum of the happiness-contribution and the achievement-contribution. We can capture this as follows:

(1) $$w = W_h + W_a$$

The happiness-contribution (W_h) is determined by a function from the amount of raw happiness in P’s life (h) and the amount of raw achievement in P’s life (a). Similarly, the achievement-contribution (W_a) is determined by a function from P’s raw achievement (a) and P’s raw happiness (h). Thus:

(2) $$W_h = f(h,a)$$

(3) $$W_a = g(a,h)$$

where ‘f(_)’ and ‘g(_)’ stand for two-place functions.

Now we need to ensure that W_h and W_a behave as described in the last section.[19] First, notice that for fixed values of achievement a, the happiness-contribution, W_h, looks very much like a simple power of h. For different values of a, the power will be different. But in each case, the graph of W_h looks like a simple power of h. (See Figs. 2, 3, and 4.) Similarly, the achievement-contribution, W_a, looks like a simple power of a, with the power being dependent on the value of h. Therefore, we can re-write the equations (2) and (3) as follows:

(2a) $W_h = h^{f(a)}$, when $h \geq 0$

(2b) $W_h = -(|h|^{f(a)})$, when $h < 0$ (where $|h|$ is the absolute value of h)

(3a) $W_a = a^{g(h)}$, when $a \geq 0$

(3b) $W_a = -(|a|^{g(h)})$, when $a < 0$ (where $|a|$ is the absolute value of a)

The absolute values in (2b) and (3b) are necessary in order to allow fractional exponents.

[19] Thanks to John Arthur Skard for help in developing these ideas in general, and especially in formulating the conditions stated below.

We can now ensure that W_h and W_a behave in the desired way by imposing certain simple conditions on *f(a)* and *g(h)*:

<u>Conditions on *f(a)*</u>

- When $h \geq 0$,
  - i) if $a = a_t$, then $f(a) = 1$
  - ii) if $a < a_t$, then $f(a) < 1$
  - iii) if $a > a_t$, then $f(a) > 1$
  - iv) as *a* decreases to -∞, *f(a)* decreases and approaches 0
  - v) as *a* increases to ∞, *f(a)* increases and approaches some constant (presumably 2)
- When $h < 0$,
  - vi) if $a = a_t$, then $f(a) = 1$
  - vii) if $a < a_t$, then $f(a) > 1$
  - viii) if $a > a_t$, then $f(a) < 1$
  - ix) as *a* decreases to -∞, *f(a)* increases and approaches some constant (presumably 2)
  - x) as *a* increases to ∞, *f(a)* decreases and approaches 0

<u>Conditions on *g(h)*</u>

- When $a \geq 0$,
  - xi) if $h = h_t$, then $g(h) = 1$
  - xii) if $h < h_t$, then $g(h) < 1$
  - xiii) if $h > h_t$, then $g(h) > 1$
  - xiv) as *h* decreases to -∞, *g(h)* decreases and approaches 0
  - xv) as *h* increases to ∞, *g(h)* increases and approaches some constant (presumably 2)
- When $a < 0$,
  - xvi) if $h = h_t$, then $g(h) = 1$
  - xvii) if $h < h_t$, then $g(h) > 1$
  - xviii) if $h > h_t$, then $g(h) < 1$
  - xix) as *h* decreases to -∞, *g(h)* increases and approaches some constant (presumably 2)
  - xx) as *h* increases to ∞, *g(h)* decreases and approaches 0

If *f(a)* and *g(h)* meet these conditions, then the welfare-function employed by the Discount/Inflation theory will be guaranteed to display the sort of behavior that was described in a qualitative way in section 8.4.1 (with one minor exception[20]). In Appendix

[20] The function W_h displays counter-intuitive behavior when *h* is a number between –1 and 1. Similarly, the function W_a displays counter-intuitive behavior when *a* is a number between –1 and 1. To see the problem, consider what happens to W_h when, say, $h = 0.2$, and a *is* above threshold. In this case, condition iii) guarantees that $f(a) > 1$. Let's suppose that *a* is so far above threshold that $f(a) = 1.8$. Thus $W_h = 0.2^{1.8} = 0.055$. But that is counter-intuitive. After all, since *a* is above threshold, the happiness contribution to

III, I present a mathematical function that meets all of these conditions. This allows us to construct a graph of the Discount/Inflation theory's welfare-function.

Let me give some explanation of these conditions. (I will only discuss the conditions on *f(a)*, however, because the rationales for the conditions on *g(h)* are exactly analogous.) Conditions i)-x) are needed to capture the behavior of the welfare function described in the six cases in section 8.4.1. I will go through each condition and explain what case it is needed to account for.

Condition i) is needed in order to capture the behavior of W_h in Case 3 (cf. Fig. 4). This was the case in which h is positive (i.e. you have a positive amount of net happiness in life) and a, your achievement level, is at threshold. When a is exactly *at* threshold, there is supposed to be no inflation or discount effect on the happiness-contribution to your welfare. Condition i) ensures that *f(a)* (i.e. the power that h is raised to) equals *1* in this case. Thus when a is at threshold, W_h will just be equal to h.

Conditions ii) and iv) are needed in order to capture the behavior of W_h in Case 1 (cf. Fig. 2). This was the case in which h is positive and a is below threshold. Discounting is supposed to occur in this case. In other words, in this case, the happiness-contribution to your welfare (W_h) is supposed to get discounted in such a way that for each additional unit of happiness, you get less and less welfare in return. Condition ii) tells us that whenever a is below threshold, the power to which h is raised will be less than one. What this means is that when a is below threshold (and h is positive), W_h will always be a number that is less than h. Moreover, because h is being raised to a power less than one, for each unit by which the total amount of happiness in your life is increased, you get progressively less welfare from it. However, condition iv) tells us that it will never be the

welfare should be inflated, i.e. W_h should be greater than h. But here we see it is not. So the problem is that when $0 < h < 1$ and a is above threshold, the happiness contribution to welfare actually gets *deflated.*

However, I think we can solve this problem. In particular, we need to add a third case to equations (2) and (3):

(2c) $$W_h = h^{1/f(a)} \text{ when } 0 < |h| < 1$$

(3c) $$W_a = g^{1/g(h)} \text{ when } 0 < |a| < 1$$

I think this solves the problem. Go back to the problematic example from before. Since $h < 1$, we get that $W_h = 0.2^{(1/1.8)} = 0.2^{0.55} = 0.41$. And this seems correct. Here we see inflation taking place, i.e. $W_h > h$. And this is just as it should be, given that a is above threshold.

case that adding additional units of happiness (h) affords you absolutely *no* welfare benefit whatsoever (or has a *negative* impact on your welfare). You will always get some additional welfare benefit from each added unit of happiness. Thus conditions ii) and iv) together capture the way W_h is supposed to behave in Case 1.

Next, when it comes to conditions iii) and v), they are needed in order to capture the behavior of W_h in Case 2 (cf. Fig. 3). This was the case in which h is positive and a is above threshold. Inflation is supposed to occur in this case. In other words, in this case the happiness contribution to your welfare (W_h) is supposed to get inflated in such a way that for each added unit of happiness, you get progressively more welfare in return. Condition iii) tells us that whenever a is above threshold, the power to which h is raised will be greater than one. What this means is that when a is above threshold (and h is positive), W_h will always be a number that is greater than h. Moreover, because h is being raised to a power greater than one, for each unit by which the total happiness in your life is increased, you get more and more welfare from it. However, there needs to be a limit on how quickly your happiness can be inflated. And this is what condition v) ensures. In particular, it tells us that as a increases to infinity, the power to which h is raised will grow to approach 2. (That is, as a approaches infinity, the curve for W_h will look more and more like the curve for h^2.)

At this point, two questions might arise. In particular, why do we need to impose any limit at all on how high $f(a)$ can get? And why should *2* be picked for this upper limit? Begin with the first question. If *no* upper limit were placed on the degree to which the happiness contribution to your welfare could be inflated as your achievement level goes to infinity, then this would lead to counter-intuitive results. Recall the person who had a tremendous amount of achievement but very little happiness. DAIAH+WACDS was seen to be problematic precisely because it implied that such a person could have an arbitrarily good life, provided he just had enough achievement. A similar problem would threaten the Discount/Inflation version of the theory if there were no top limit on how much one's happiness could be inflated, as a is increased to infinity. For if there were no such top limit, then for a person who has a tiny positive amount happiness but an arbitrarily large amount of achievement, W_h would become arbitrarily high too. And in that case, the contribution to welfare made by this person's tiny amount of happiness would be arbitrarily high. But I submit that this is counter-intuitive. Intuitively, you cannot have an arbitrarily good life if you only have a tiny amount of happiness – no matter how much achievement you have. To prevent this counter-intuitive result, we cannot allow that *f(a)* can become arbitrarily high as *a* is increased to infinity. Thus we need *f(a)* to asymptotically approach some constant as *a* gets bigger and bigger.

Now, why should this constant be set at *2*, in particular? Considerations of symmetry suggest that the constant should be 2. Condition iv) implies that as *a* is decreased from the threshold level, a_t, to minus infinity, then *f(a)* will begin at 1 and then asymptotically approach 0. Thus *f(a)* (i.e. the power to which *h* is raised) can change by a maximum of 1 as *a* is decreased from threshold to minus infinity. Accordingly, symmetry would suggest that same thing should happen when *a* is increased *above* threshold. That is, *f(a)* should be allowed to change by a maximum of 1 in the *positive* direction, as *a* is *increased* from the threshold level, a_t, to *plus* infinity. Thus we get condition v), saying that as *a* is increased from the threshold level, a_t, to infinity, *f(a)* (i.e. the power to which *h* is raised) should begin at 1 and then asymptotically approach 2.

So far we have seen the intuitive motivation for conditions i)-v). They are needed to describe the behavior of W_h in Cases 1-3. By contrast, conditions vi)-x) are needed to capture the behavior of W_h in Cases 4-6, i.e. the cases in which *h* is a negative number. Condition vi) is needed to capture the behavior of W_h in Case 6 (cf. Fig. 7). Similarly, conditions vii) and ix) are needed to capture the behavior of W_h in Case 4 (cf. Fig. 5). And conditions viii) and x) are needed to capture the behavior of W_h in Case 5 (cf. Fig. 6). However, I won't go through these conditions separately, because the rationales for each of them are directly analogous to the rationales just discussed for conditions i)-v). So it should be fairly obvious why conditions vi)-x) are needed. (And for similar reasons, I will not explicitly discuss the rationales for the conditions on *g(a,h)* either.)

Having said something in support of the conditions imposed on *f(h,a)* and *g(a,h)*, we are now in a position to formulate the Discount/Inflation theory itself:

> <u>The Discount/Inflation Version of the Happiness and Success Theory</u>: The amount welfare contained in P's life equals the value for *w* that is returned by the welfare-function, described by equations (1)-(3) and conditions i)-xx), when P's values for *h* and *a* are taken as input.

This completes my presentation of the theory.[21] This theory seems to have at least as many advantages as the Thresholds version of the theory. We saw that the Discount/Inflation theory adequately accommodates the case of Porky. We also saw that it does not have the problematic implication that a person who has a tiny amount of happiness can have an arbitrarily good life, provided only that he has enough achievement. For similar reasons, the Discount/Inflation theory does not have the problematic implication that a person who has a tiny amount of success can have an arbitrarily good life, provided only that he has enough happiness. On this theory, to have an extremely good life in terms of welfare, one must have *both* a very high amount of happiness *and* a very high amount of success in achieving worthwhile goals.

What’s more, the Discount/Inflation theory in fact seems to be *superior* to the Thresholds theory. For one thing, it does not require postulating any cutoff point on the overall welfare scale above which you simply cannot get as long as either your happiness or your achievement is below threshold. Second, recall that the Thresholds theory implied that sometimes very small increases in one of your scales will lead to oddly large increases in overall welfare. Suppose you are very high above threshold with respect to happiness, but just under threshold with respect to achievement. Now suppose we add just a tiny bit of achievement to your life, just enough to get you to reach the achievement threshold. This will cause a huge jump in your overall welfare, according to the Thresholds theory.

[21] In previous chapters, I have made a fuss about how certain theories of welfare (e.g. Desire Satisfactionism) must be consistent with the principle that the intrinsic value of something must depend only on its intrinsic features. Does the Discount/Inflation Theory conflict with this principle? Perhaps one thinks so, given that, e.g., the welfare-contribution made by your happiness is supposed to depend on whether your achievement is above threshold or not.

However, the Discount/Inflation Theory can indeed be understood in a way that is consistent with this principle. In particular, the theory should be taken to say that the fundamental bearers of welfare value are complex states consisting of your having a given happiness score and your having a given achievement score. The theory can't be taken to say that happiness by itself is intrinsically good for a person, or that achievement by itself is intrinsically good for a person. If the theory said this, it would indeed conflict with the principle that intrinsic value must depend on intrinsic features. Therefore, the theory should instead be understood so that what has welfare value for you are states of your-being-happy-to-some-degree-and-your-being-successful-to-some-degree. For such a state, everything that is needed to determine the amount of welfare value that it has for you, according to the Discount/Inflation Theory, would be contained within that state itself. [More specifically, the theory would have to say this: A state of affairs, S, has (a positive amount of) intrinsic welfare value for P iff S is a state of P's being happy to some (positive) degree and P's being successful with respect to worthwhile projects to some (positive) degree.] If the theory is understood in this way, it would not conflict with the principle that intrinsic value must depend on intrinsic features.

But this oddness of the Thresholds theory is not present in the Discount/Inflation theory. After all, if your happiness is above threshold and your achievement is just below threshold, this will cause the happiness-contribution to your welfare to be discounted just a little bit, while the achievement-contribution to your welfare will get inflated. And adding just enough achievement to your life to get you up to threshold will not cause a strange jump in your welfare. Instead, as your achievement level approaches threshold, the happiness-contribution to your welfare will smoothly cease to be discounted. This is because (assuming h is greater than 0) conditions ii) and v) guarantee that as a approaches a_t from below, $f(a)$ will approach 1, and so the value of W_h will approach being equivalent to h. Then once a_t goes slightly above threshold, W_h will be equal to a number that is just slightly greater than h. This is behavior is generated by conditions iii) and v). Thus the Discount/Inflation theory will not allow small increases in either happiness or achievement to cause strange jumps in welfare, which is a problematic consequence of the Thresholds theory. Of course, closer study of these two theories may reveal problems with the Discount/Inflation theory. But I think the Discount/Inflation theory is likely to prove to be superior to the Thresholds theory.

In Appendix III, I present a set of mathematical functions corresponding to $f(a)$ and $g(h)$ that meet conditions i)-xx). There are, however, infinitely many functions that meet these conditions. But the functions I describe in Appendix III are as good as one can hope for, I think. Thus Appendix III should be seen as containing the most fully worked out version of the Discount/Inflation Theory that I am able to offer.

8.5 Conclusion

Most traditional theories of welfare have been monistic. They make welfare be a function of one item. In this chapter, I have argued that theories of the Happiness and Success type are particularly attractive because they (or some of them, at least) are capable of avoiding most defects of the monistic theories. Since Happiness and Success theories make welfare a function of two components, such theories offer a rich set of resources for dealing with problem cases. I presented several versions of the Happiness and Success theory, and I argued that the Thresholds version, and especially the

Discount/Inflation version, are plausible theories that deserve further attention. Pending further investigation, I am inclined to think that the Discount/Inflation version of the Happiness and Success Theory is true. If nothing else, my discussion of the Thresholds theory and the Discount/Inflation theory should show that the mathematical resources that multi-component theories of welfare make available are powerful tools that can aid significantly in the defense of one's preferred theory of welfare.

APPENDIX I

INTERNALISM ABOUT A PERSON'S GOOD: DON'T BELIEVE IT[*]

Internalism about a person's good is roughly the view that in order for something to intrinsically enhance a person's well-being, that person must be capable of caring about that thing. This is a view that a number of contemporary moral philosophers seem to accept, in one form or another. Peter Railton, for instance, expresses his support for such a view as follows:

> it does seem to me to capture an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware. It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him. (Railton 2003, p. 47)

David Velleman, too, endorses a version of internalism that is 'restricted to statements about what's intrinsically good for a person.' (Velleman 2000, p. 85) In his view, internalism specifically about one's good is more likely to be true than analogous internalist views about other normative concepts, because while 'the norms of morality and rationality aren't tailored to suit individual tempers,'

> a person's own good is indeed tailored to him, and we cannot imagine him saying that his good is not for him, since its being for him seems essential to its being specifically his good. (Velleman 2000, p. 85)

Internalism about a person's good is not always endorsed as explicitly as this, however. A tacit commitment to the view seems to underlie many attempts to argue for 'subjective' theories of well-being. In the view of L.W. Sumner, for instance, the correct theory of

[*] Thanks to Fred Feldman, Ernesto Garcia, Scott Hill, Kristian Olsen, James Patten and Michael Rubin for many helpful comments and conversations about this paper.

well-being is going to have to be subjective in the sense that it ‘make[s] your well-being depend on your own concerns: the things you care about, attach importance to, regard as mattering, and so on.’ (Sumner 1996, p. 42)

Support for internalism about a person’s good thus seems to be fairly widespread, but it is less common to find direct arguments given in favor of the view. While Railton offers a bit of intuitive motivation for the view,[1] and Velleman gives an overt argument for it,[2] the most sophisticated and comprehensive defense of the view is given by Connie Rosati. She not only incorporates the insights of Railton and Velleman, but goes well beyond them to provide a clear and *prima facie* compelling case for internalism about a person’s good (Rosati 1996).

I argue in this paper that internalism about a person’s good should not be believed. My focus here will be Rosati’s views, since she provides the most comprehensive case in favor of internalism specifically about a person’s good. Rosati’s defense of the view consists mainly in offering five independent arguments to think that at least some form of internalism about one’s good is true. But I argue that, on closer inspection, not one of these arguments succeeds. The problems don’t end there, however. While Rosati offers good reasons to think that what she calls ‘two-tier internalism’ would be the best way to formulate the intuition behind internalism about one’s good, I argue that two-tier internalism is actually false. In particular, the problem is that no substantive theory of well-being is consistent with two-tier internalism. Accordingly, there is reason to think that even the best version of internalism about one’s good is in fact false. Thus, I conclude, the prospects for internalism about a person’s good do not look promising.

1. How to Formulate Internalism

It will be useful to begin with a brief explanation of Rosati’s reasons for thinking that two-tier internalism is the most plausible formulation of internalism about a person’s good. What underlies this sort of internalism is the intuition that ‘an individual’s good

[1] Railton supports internalism about a person’s good by pointing to the counter-intuitiveness of the idea that something that is ‘highly alien’ to a person can nonetheless be part of that person’s good. He also seems to hint at something like what Rosati calls the argument from metaphysics. (Cf. Railton, 2003, p. 47)

[2] In ’Is Motivation Internal to Value?’, Velleman presents a version of the argument that Rosati calls the argument from ‘ought’ implies ‘can.’ (Cf. Velleman 2000, pp. 93-96, and Rosati 1996, pp. 320-322)

must not be something alien – it must be "made for" or "suited to" her.' (Rosati, p. 298) But how is internalism to be formulated precisely? Rosati offers four successive formulations of the internalist thesis, each one supposedly superior to the previous one. The first formulation Rosati calls 'simple internalism':

> Let's understand simple internalism, then, as follows: something X can be good for a person A only if A is capable of caring about X. X, to be good for A, must be a possible object of her concern. (p. 301)

The problem with this formulation of internalism is not that it is obviously false, but rather, as Rosati points out,[3] that it is too weak to capture the internalist's basic intuition. That is, it cannot capture the intuition that things that are 'completely alien' to one (e.g. things one could be brought to care about only by being radically altered by hypnosis or brain surgery) are not capable of being good for one. To adequately capture this intuition, internalism must be formulated so as to rule out the possibility of things being 'completely alien' to one and yet still good for one. But simple internalism does not rule out this possibility. It just amounts to the claim that for X to be good for one, there must be some possible world in which one cares about X. Thus simple internalism does not rule out the possibility of a person who is incapable of caring about, say, making money except under the most extreme circumstances (e.g. post-lobotomy), and yet for whom making money is still good. Simple internalism allows that making money can be good for this person because there is indeed a possible world in which he cares about making money (viz. the world in which he is lobotomized). Thus simple internalism is too weak to adequately capture the basic internalist intuition.

By contrast, the next formulation of internalism that Rosati mentions is not overly weak. This view, which she calls 'strict internalism', states that

> Something X can be good for a person A only if she can care about X without any marked alteration of her present condition. (p. 303)

Nonetheless, strict internalism suffers from the opposite problem: it is too restrictive to be plausible. After all, one's actual conditions might be impaired. Thus, there could be things that intuitively are part of one's good, but that one is not capable of caring about in

[3] Cf. Rosati, pp. 301-302

one's *actual* conditions.[4] To avoid this problem for strict internalism, Rosati proposes a 'stronger' version of the view:

> We might begin to characterize a stronger internalist constraint as follows: something X can be good for a person A only if A would care about X for her actual self, were A under appropriate conditions and contemplating the situation of her actual self as someone about to assumer her position. (pp. 303-304)

This 'stronger internalism' seems to avoid the problem for strict internalism. For if X really is good for you, but your actual conditions are defective in such a way that you cannot actually care about X, then you might well still be able to care about X under 'appropriate conditions' (e.g. conditions of full information and rationality).

However, stronger internalism is too vague to be satisfactory formulation of the view. We still need to pin down what the 'appropriate conditions' are supposed to be. As Rosati explains,

> How are we to constrain what shall count as "appropriate" conditions? To satisfy the intuition supporting internalism, we must insure that these conditions do not themselves strike us as alien. (…) We need a way to rule out alienated conditions, without allowing the appropriateness of conditions to depend on how they now strike a person, whatever her present state might be. (pp. 304-305)

The thought here, then, is that if I simply would not care what my counterfactual self would want, then to take my counterfactual self's desires to be determinative of my good could make my good seem to be highly 'alien.' For instance, I do not care what a counterpart of me who has been indoctrinated into some religious cult would desire as someone who is about to be put into my actual situation. Thus the intuition underlying internalism suggests that the desires of my indoctrinated counterpart should not be taken to help determine the good of my actual self. Internalism must be formulated so as to prevent my good from being determined by the desires of counterparts who are such that I do not care what they would think or want.

To do this, Rosati suggests that we take the 'appropriate conditions' to be the ones that the agent would reflectively endorse in normal conditions. To capture this idea, Rosati introduces the notion of *ordinary optimal conditions*, which 'would include that a person not be sleeping, drugged, or hypnotized, that she be thinking calmly and rationally, and that she not be overlooking any readily available information.' (p. 305)

[4] Rosati puts the point this way: 'a person may be unable in her present condition to care about something that she would care about under other conditions and that we are perfectly prepared to regard as part of her good. Her unaltered condition, after all, might itself be one in which she is seriously impaired.' (cf. p. 303)

Using this notion, Rosati can now formulate her preferred version of internalism, which she calls *two-tier internalism*:

> [S]omething X can be good for a person A only if two conditions are met:
>
> 1. Were A under conditions C and contemplating the circumstances of her actual self as someone about to assumer her actual self's position, A would care about X for her actual self;
> 2. conditions C are such that the facts about what A would care about for her actual self while under C are something A would care about when under ordinary optimal conditions. (p. 307)

The idea here is this. Internalism in general requires that something must be 'suited to you' if it is to be part of your good. Two-tier internalism spells out this notion of what it is for something to be 'suited to you.' Consider a set of idealized circumstances – call them 'C_{IDEAL}' – that have the following feature: were you in ordinary optimal conditions (e.g. not drugged, distracted or sleeping), you would care about what a counterpart of you in C_{IDEAL} would desire. (For many people, these idealized circumstances, C_{IDEAL}, might include full information and rationality.) If a counterpart of you in C_{IDEAL}, having been told that he or she is about to be placed in your *actual* circumstances, would desire X, then X is 'suited to you.' Two-tier internalism requires that something must be suited to you in this sense in order for it to be good for you.

Two-tier internalism has the most going for it, Rosati thinks, and I am inclined to concur. Unlike simple internalism, it is not too weak to capture the internalist's underlying intuition. Unlike strict internalism, it is not too restrictive to be plausible. And unlike stronger internalism, it is not too vague to be of interest. Rosati thinks that this is the version of internalism that should be accepted by those who already accept the basic internalist intuition. But why should one accept the basic internalist intuition in the first place?

2. Why Rosati's Arguments for Internalism Fail

In response to the question of why one should accept the internalist intuition at all, Rosati presents five arguments for believing some version of internalism about a person's good. Some of these arguments seem capable of supporting two-tier internalism in particular; others seem to be capable only of supporting some weaker version of internalism like simple internalism. I want to consider the most defensible versions of Rosati's arguments, and so in the discussion to follow, I will take her arguments to be aimed at establishing only something very weak: viz. the claim that *some* version of internalism is true (perhaps merely simple internalism). I argue that the five arguments Rosati offers cannot establish even this very weak conclusion. So the case for internalism about a person's good in general seems unconvincing.

2.1 The Argument from Judgment Internalism

Rosati's first argument begins with a well-known view called *judgment internalism*. This is roughly the view that 'it is a necessary condition on sincere judgment about a person's good that the speaker normally have some inclination, not necessarily overriding, to promote or to care about that thing. A person cannot sincerely judge that something is good for herself unless she has some tendency to approve of or pursue that thing.' (p. 310) Rosati argues that if the necessary condition on sincere judgment that judgment internalism imposes is correct, then some version of internalism about a person's good[5] is also true. Rosati explains her argument as follows:

> The truth of judgment internalism might seem to support the claim that a plausible account of the good for a person must satisfy existence internalism, at least in the form of simple internalism. An account of the good for a person must permit judgments about a person's good to serve their characteristic action-guiding functions. It must be able to explain how it is that, at least normally, judgments about a person's good motivate, and it must also preserve their characteristic recommending and expressive functions or normative force. An account can succeed in this, without embracing noncognitivism and its antirealist implications, only if it satisfied simple internalism. By limiting a person's good to some subset of those things that can matter to her, an account insures that it will at least be possible for judgments about a person's good to perform their characteristic functions. (p. 310)

The argument here does not seem to be that the truth of judgment internalism *entails* a version of internalism about a person's good. Instead, I think Rosati's argument is most plausibly understood as an inference to an explanation. Accordingly, I suggest that her argument, explicitly spelled out, is this:

<u>The Argument from Judgment Internalism:</u>

1) An explanation is needed of the truth of judgment internalism.
2) One way to explain it would be to suppose that non-cognitivism is true.
3) Another way to explain it would be to suppose that some form of internalism about a person's good, in particular two-tier internalism, is true.
4) These are the only two explanations available.
5) The explanation involving non-cognitivism is not an acceptable explanation.
6) If 1)-5), then 7).

[5] Rosati often uses 'existence internalism' to refer to internalism about a person's good, in order to distinguish it from judgment internalism. I will continue to use the phrase 'internalism about a person's good', however.

7) Therefore, some form of internalism about a person's good, in particular two-tier internalism, is true.

What reason do we have to accept the premises in this argument? Premise 1) rests on the idea that there is intuitive plausibility to the idea that your judgments about your good must guide your actions in a certain sense. Specifically, your sincerely judging that something is good for you requires that you are at least somewhat motivated to pursue it (provided you are not irrational). While many philosophers have debated the truth of judgment internalism, what Rosati is primarily concerned to argue is that *if* the view is true, then this would support some form of internalism about one's good. Thus it seems charitable to simply accept judgment internalism for the sake of argument. I will raise no objections to it here, anyway. Now, supposing that judgment internalism *is* true, we would want some explanation of *why* it is true. After all, it would be mysterious if the truth of judgment internalism were merely an unexplainable, brute fact. Hence premise 1).

One possible explanation of the truth of judgment internalism, it seems, might be offered by non-cognitivism. This is roughly the view that part of the function of uttering an evaluative judgment is to express one's attitudes of approval or disapproval. This is simply part of their meaning. So if there is indeed a necessary connection, as judgment internalism posits, between sincerely uttering a judgment about one's good and being motivated to act in accordance with that judgment, then one neat way to account for this would be to make the non-cognitivist claim that it is part of the *meaning* of evaluative judgments that they serve to express one's attitudes of approval and disapproval. Such attitudes, after all, are directly connected with one's motivational states. (Perhaps they constitute motivational states, or else they might be direct causes of these states.) So if non-cognitivism is true, judgment internalism would seem to be true as well. Hence premise 2) in the argument.

Non-cognitivism might not be the only available explanation for the truth of judgment internalism, however. In particular, one might accept the claim, made in premise 3), that if some form of internalism about one's good is true, then this would secure the truth of judgment internalism. Rosati points out that simple internalism is not strong enough a thesis to establish judgment internalism,[6] but she thinks that two-tier internalism can do the job. In a moment, I will argue against this claim. But first, here is my best attempt to explain why someone might be inclined to accept premise 3).

According to simple internalism, X could be good for you even if you would care about X only under some far-out conditions – say, after your lobotomy – that you are in fact completely unconcerned with. Thus Rosati thinks that if it were only simple internalism that were true, your judging that something is good for you would not guarantee that you actually have any motivation to pursue that thing. So judgment internalism would not be established. By contrast, Rosati thinks that two-tier internalism does a better job of guaranteeing the motivational force of your judgments about your good. After all, two-tier internalism limits the things that can be good for you to the objects that quite plausibly are objects of your concern. If it's true that X can be good for you only if you would desire it under idealized conditions of the sort that matter to you, then when you judge 'X is good for you' you would most likely be motivated to go get X. For, if X meets the conditions of two-tier internalism, X is likely to be something that is a real concern of yours. Thus when you judge that X is good for you, you would thereby be *admitting* that X matters to you, and so you would have to be somewhat motivated to go out and get X. Thus Rosati thinks two-tier internalism does a decent job of explaining why your judgments about your good motivate. This, I take it, is the idea behind premise 3).

When it comes to premise 4), the thought would presumably be that non-cognitivism and two-tier internalism are the only plausible explanations that have been offered of the truth of judgment internalism. Rosati does not argue that there can be no other explanations; rather the idea would seem to be just that there are no other good explanations on offer.[7] Premise 5) captures the idea that non-cognitivism is an unattractive option. Not only does it have familiar problems (e.g. the embedding

[6] Cf. Rosati, p. 311

[7] Perhaps premise 4) would be more plausible if it stated that non-cognitivism and internalism about a person's good are the only available *philosophical* explanations of the truth of judgment internalism. However, then premise 1) would have to be modified accordingly, which might lead one to doubt its plausibility. In particular, one might wonder why *only* a philosophical explanation of the truth of judgment internalism would be adequate. Perhaps a psychological explanation might do as well.

problem), but it seems to require the truth of anti-realism.[8] Considering the counter-intuitive commitments of anti-realism, the cost of adopting non-cognitivism thus seems to be high.

However, this leaves only one option on the table for explaining the truth of judgment internalism: namely, two-tier internalism. And this is the thought behind premise 6). This premise may be taken to embody a principle about inference to the best explanation. In particular, the thought would be that two-tier internalism would seem to be the best available explanation of judgment internalism, and since we are at least epistemically permitted to believe that the best available explanation, it follows that we may take two-tier internalism to be true.

This concludes my best attempt to spell out and explain Rosati's Argument from Judgment Internalism. If I have got the argument right, however, it is not persuasive. Several of the premises it employs are problematic. I suspect that many will have doubts about lines 4) and 6). However, here I will focus on line 3). For this seems to be the keystone of the argument. Perhaps the argument could be reformulated so that the problems afflicting the other lines are avoided. But if premise 3) is indefensible, then it seems no version of the argument can succeed.

The problem is that internalism about a person's good does not explain the truth of judgment internalism. Let us focus on two-tier internalism. Rosati thinks that since two-tier internalism limits the things that can be good for you to what quite plausibly are genuinely objects of your concern, it would follow that when you judge that X is good for you, X must be something you are motivated to obtain. However, this is a mistake. The reason is that people frequently make *incorrect* judgments about what is good for them. Granted, if we assume two-tier internalism, then it follows that when you *correctly* judge that X is good for you, X would be something that matters to you in the special way two-tier internalism requires. But this by itself is not enough to establish judgment internalism. For judgment internalism states that sincere judgments about what is good for you will *in general* motivate (provided you are rational). Suppose you judge that Y is good for you, but you are *mistaken*. That is, despite your judgment, Y is in fact not good

[8] Of course, some non-cognitivists have taken steps towards addressing this second concern by showing how some of the problematic features of anti-realism might be avoided. (See, for instance, Blackburn, 1993.)

for you. Even assuming two-tier internalism, there will be no guarantee that Y is something you are concerned with or would be motivated to obtain. After all, since Y is not good for you, two-tier internalism does not guarantee that Y is something that matters to you in any way. Thus even if two-tier internalism is true, it would not follow that your sincere judgments about what is good for you will *in general* have motivating force. On the assumption of two-tier internalism, it is perhaps true that you will be motivated in cases when you correctly judge that something is good for you; but when you incorrectly judge that something is good for you (and surely this happens all the time), motivation will not be guaranteed. Accordingly, two-tier internalism cannot establish judgment internalism.[9]

Perhaps Rosati would try to respond by insisting that if you *know* that two-tier internalism is true, then you could not judge something to be good for you and yet remain entirely unmoved to pursue it. Even if we grant this claim, however, it will not save the argument. For one thing, we can grant that a belief in two-tier internalism would guarantee the motivating force of judgments about one's good *even if two-tier internalism turns out to be false*. Thus even if we grant that judgment internalism would follow from the fact that everybody *believes* two-tier internalism, it would not be the case that judgment internalism follows from the truth of two-tier internalism *itself*. What's more, even if two-tier internalism *were* true, there could still be people who either believe that two-tier internalism is outright false, or else have no beliefs about the matter whatsoever. (In fact, there surely are *actual* people of both these kinds.) Now, even if two-tier internalism *were* true, people like this still would not be guaranteed have any motivation to pursue the things that they judge to be good for them (for reasons presented in the last paragraph). Thus, the present line of response does not do anything to save the argument. It seems that the truth of two-tier internalism cannot explain judgment internalism.

[9] Could one save premise 3) by pointing out that judgment internalism does not have to be the strong claim that *all* sincere judgments about one's good motivate, but could rather merely be the weaker claim that sincere judgments about one's good *normally* motivate? This will not save premise 3). The same problem persists. After all, even under normal conditions, most of us will frequently make sincere but mistaken judgments about what is good for us. And in such cases (even though conditions are normal) two-tier internalism still will not guarantee that these mistaken judgments will motivate. Thus from the truth of two-tier internalism, not even this weaker version of judgment internalism (i.e. the claim that sincere judgments about one's good *normally* motivate) would follow. (Thanks to Kristian Olsen for pressing me on this point.)

2.2 The Argument(s) from the Metaphysics of Value

The next argument Rosati presents in favor of internalism about a person's good rests on some considerations about the metaphysics of value.[10] In particular, the idea is that internalism about a person's good must be true since 'value exists only in virtue of subjectivity.' (p. 313) Rosati explains the argument as follows:

> If value can exist only if there are creatures who can be affected by and react to their world, then value, and more specifically, goodness for a person, must be a motivational property. What else, after all, could it be? The only alternative might seem to be that the property of being good for a person is a Moorean, non-natural property, but this alternative introduces special metaphysical and epistemological problems. We thus arrive at the suggestion that not only must a person's good be something that she can care about, but that the very goodness of her good is *constituted* by her being disposed to care about it, at least under ideal conditions. Considerations about the metaphysics of value show that a plausible account of a person's good must satisfy existence internalism. (p. 313-315)

I see two distinct, but perhaps related arguments in this passage, and I am not sure which one Rosati means to endorse. Perhaps she endorses both. In any case, the two arguments may be stated as follows:

First Metaphysical Argument

1) Goodness for a person is either a motivational property, or else it is a Moorean non-natural property.
2) It's not the case that goodness for a person is a Moorean non-natural property.
3) Lemma: Goodness for a person is a motivational property.
4) If goodness for a person is a motivational property, then internalism about a person's good is true.
5) Therefore, internalism about a person's good is true.

Second Metaphysical Argument

1) Value can exist only if there are creatures who can be affected by and react to their world.
2) If 1), then goodness for a person is a motivational property.
3) If goodness for a person is a motivational property, then internalism about a person's good is true.
4) Therefore, internalism about a person's good is true.

I will explain each of these arguments[11] in turn and point out the difficulties they face.

[10] Note that this metaphysical argument is prefigured, but not explicitly developed, in Velleman 2000. (See p. 86)

[11] Note that neither of these formulations of the Argument from the Metaphysics of Value issue in the conclusion that two-tier internalism is true; they merely purport to establish that that some form of internalism about a person's good is true. This is because Rosati does not offer much explanation of how the considerations she mentions about the metaphysics of value would support specifically two-tier

Begin with the First Argument from the Metaphysics of Value. Why think, as the first premise states, that goodness for a person must either be a motivational property or a non-natural property? The rationale would seem to be that there are only two options when it comes to the question of what the metaphysical status of goodness for a person is: it must be either a natural property or a non-natural property. What is a natural property? David Copp offers a sufficiently a clear and plausible account:

> a property is natural if and only if – leaving aside analytic truths, if there are any – any proposition about the instantiation of that property that can be known, can be known only empirically, or by means of empirical observation and standard modes of inductive inference. (Copp, 2004. pp. 12-13)

Now, on the one hand, G.E. Moore (1903) and some philosophers inspired by him[12] defend the view goodness cannot be a natural property in this sense. On this view, propositions about the instantiation of goodness for a person cannot be known solely by means of empirical observation. Appeals to some sort of intuitions are necessary in order to come to know propositions of this sort. On the other hand, goodness for a person might be a natural property. If it is, then its instantiation would have to have empirically observable consequences. These observable consequences would presumably have to involve people's motivations, actions and/or sentiments. So if goodness for a person is a natural property, then presumably it is a 'complex motivational property' of some sort. Hence premise 1).

The second premise in the First Metaphysical Argument embodies the not uncommon view that goodness for a person cannot be a non-natural property. In support of this idea, Rosati cites Mackie's well-known metaphysical and epistemological objections.[13] The metaphysical objection is roughly that if goodness were a non-natural property, it would be 'metaphysically queer' and utterly unlike every other property in the universe. On the

internalism. She claims that 'the argument also supports two-tier internalism, albeit more indirectly.' (p. 315) But when she goes on to explain how this indirect argument is supposed to work, she just appeals to the Argument from Judgment Internalism. As she puts it, 'this argument for [two-tier] internalism cannot be entirely independent of the first [i.e. the argument from judgment internalism].' (p. 315) But we have already seen that the Argument from Judgment Internalism fails. I am going to assume that Rosati is right that the Argument from the Metaphysics of Value can support two-tier internalism only if the Argument from Judgment Internalism goes through. As a result, I focus here only on the Argument from the Metaphysics of Value as it applies to internalism about a person's good in general, not as it applies to two-tier internalism.

[12] For a recent defense of the view that goodness is a non-natural property, see for instance Parfit 1997. (However, note that Parfit states his view primarily in terms of reasons, not goodness.) Also see Huemer 2005.

[13] See Mackie 1977, ch. 1.

other hand, the epistemological objection is that if non-naturalism were true, we would require a special mental faculty of intuition in order to obtain evaluative knowledge; but there is no evidence that people are imbued with any such faculty. Citing arguments like these, Rosati rejects the idea that goodness for a person could be a non-natural property.

Of course, few philosophers, even ones with naturalistic commitments[14], are convinced by Mackie's arguments. However, the issue of naturalism about goodness is a complicated one. I cannot embark on a serious treatment of it here, so let us assume for the sake of argument that it is an untenable view that goodness for a person is a non-natural property. Accordingly, we would get line 3) in the argument: goodness for a person must be a motivational property.

Premise 4) asserts that if goodness for a person is a motivational property, then internalism is about a person's good must be true. Why think this? At first glance it might not be clear because Rosati does not tell us what is meant by 'complex motivational property'. But I think we may understand this phrase roughly as follows:

> D1. P is a 'motivational property' for a creature C =df. C would, at least under some circumstances, be motivated to obtain things that possess P.

If this is what 'motivational property' means, then internalism about a person's good really would follow from the fact that goodness for a person is a motivational property.[15] So this is presumably how premise 4) in the First Motivational Argument should be understood in order to be plausible.

However, now a problem arises. If D1 captures more or less what is meant by 'motivational property,' then there seems to be no reason to accept premise 1) in the argument. That is, there is no reason to think that if goodness for a person is not a non-natural property, then the only other option is that it is a motivational property. There are many ways to understand goodness for a person so that it is both a natural property and yet not a motivational property, i.e. a property that the person in question would be attracted to or motivated to obtain. To take just one example, one might propose a naturalistic account of the good on which the facts about what is good for you are

[14] See, for instance, Brink 1984, pp. 111-25. (Also see Brink 1989.)

[15] There are other ways to understand 'motivational property,' of course (e.g. as a property that is had by a creature's motivational system). But it seems that if any of these alternative interpretations of 'motivational property' are adopted, internalism about a person's good would not follow from the fact that goodness for a person is a motivational property. So premise 4) would become false.

determined by the nature of your species (e.g. its characteristic capacities and functions).[16] On this theory, goodness for a person would be both a natural property and yet not a motivational property in the sense of D1.[17] And so premise 1) in the argument would be false. It is not the case that goodness for a person is either a motivational property or else a Moorean, non-natural property. Thus the First Metaphysical Argument fails.

In that case, might the Second Metaphysical Argument be more likely to succeed? In fact, it seems to be less so. Begin with premise 1). The most plausible way to understand this premise would be to take it to be the claim that the property of goodness for a person *would not be instantiated* if there were no creatures around for whom things can matter.[18] This claim would seem to be true, no matter whether the true theory about goodness for a person were a paradigmatically internalist one like Desire Satisfactionism or a paradigmatically externalist one like an Objective List theory. Suppose Desire Satisfactionism is true. If there were no creatures who are capable of desiring things, then there would be no episodes of desire satisfaction. And so the property of goodness for a person would be uninstantiated. Similarly, suppose a version of the Objective List Theory were true according to which getting more money is the only thing that enhances a person's welfare. If there were no people around, then nobody would be able to acquire more money, and so nothing would instantiate the property of being good for a person. Thus, premise 1) on this understanding of it seems plausible.

[16] Philippa Foot (2001) has defended an account of this sort.

[17] Rosati herself mentions another kind of naturalistic account on which goodness for a person is not a motivational property in the sense of D1. As she puts it:

> 'the "sensibility theories' proposed by John McDowell and David Wiggins might hold that goodness for a person is a sui generis secondary property, akin to color properties, that can only be grasped by those with the proper ethical sensibilities. While such a property and the corresponding sensibility are mutually dependent, in the way that color and color sense are or that humorousness and a sense of humor are, the property is not constituted by our dispositions to care about things.' (p. 314)

I did not mention this theory in the main body of the text because Rosati thinks that the 'sensibility theory' is an internalist theory. Thus the Metaphysical Argument could be reformulated so as to take account of the sensibility theory. In particular, the first premise would state not 'Either goodness for a person is a motivational property or else it is a Moorean, non-natural property,' but rather it would have to say 'Either *internalism about a person's good is true*, or else goodness for a person is a Moorean, non-natural property.' This reformulation of the argument would accommodate the possibility of an internalist theory of goodness for a person according to which goodness is not a motivational property. The problem I raise in the main body of the text still threatens this revised version of the argument, however. For there are some versions of naturalism about goodness for a person, like the Aristotelian account, on which internalism is not obviously true.

[18] Other interpretations of this premise are possible. In particular, see the next footnote.

Nonetheless, there seems to be no reason to believe premise 2) in the Second Metaphysical Argument. Why think that, from the fact that the property of goodness for a person would be uninstantiated if there were no people, it follows that the property of goodness for a person must be a motivational property? We would not make such an inference about any other property. For instance, consider the property of being salami. If there existed no people or other creatures with sausage making capabilities, then the property of being salami would not be instantiated. However, we clearly would not conclude from this that the property of being salami is a motivational property. It seems that all sorts of properties would be uninstantiated if there existed no creatures that could be affected by or react to their world. But this gives us no reason to think that these properties are motivational properties. So why draw this conclusion when it comes specifically to the property of being good for a person? I cannot see any reason to. Thus I see no reason to accept premise 2) in the Second Metaphysical Argument.[19] [20]

2.3 The Epistemological Argument

The basic idea behind Rosati's third argument is that '[w]e can justify to a person the claim that something is good for her… only if her alleged good satisfies internalism.' (p. 316) Rosati's single clearest presentation of the argument is this:

[19] Another way to understand premise 1) in the second metaphysical argument would be to take it to be the claim that the property of goodness for a person *itself* would not exist if there were no creatures for whom things can matter. I think this understanding of premise 1) is less plausible than the one mentioned above, since I am inclined to accept a metaphysics on which properties can exist even if they are never instantiated. However, this alternative understanding of premise 1) would not help one salvage the second metaphysical argument. After all, there seems to be no reason to think that from the fact that the property of goodness for a person would not exist if there were no people around, it follows that goodness for a person in the actual world, where it does exist, is a motivational property. According to a metaphysics on which properties that are never instantiated do not exist, the property of being salami would not exist if sausage-making creatures did not exist. But why would it follow from this that the property of being salami in the actual world, where this property does exist, is a motivational property? There seems to be no reason to think this.

[20] Perhaps Rosati could respond to this argument by pointing out that the property of being salami is an *artifactual* property. She may grant that it is a mistake to infer that a given property, P, is a motivational property from the fact that P would not be instantiated if there were no creatures around. After all, artifactual properties (like that of being salami) are like this too. But Rosati may nonetheless insist that one *can* legitimately infer that P is a motivational property from the fact that *both* i) P would not be instantiated if there were no creatures around *and* ii) P is a non-artifactual property. (Thanks to Scott Hill for pointing out this line of response to me.)

However, this too would be a mistaken inference. After all, consider the property of being hairy. Or a phenomenal property like that of appearing to someone to be bluish. Both of these are non-artifactual properties that would not be instantiated if there were no creatures around. Still, neither one is a motivational property. Thus this line of response will not save the argument.

> Unless a person *could* care about the thing in question it cannot be justified as a part of her good, because the possibility of her caring about the thing is necessary evidence of its being good for her. But why think that it is necessary evidence? Consider the following thought experiment. Suppose that a person A could not be brought to care about a thing X under any conditions and so concluded that it is not good for her. What counterevidence could be produced to subvert A's conclusion? We have no picture, the argument might go, of what such evidence could be. (p. 316)

The idea, then, is that since it seems to be impossible to justify that X is good for one unless one is capable of caring about X, it follows that simple internalism is true. The argument may be succinctly stated as follows:

The Epistemological Argument

1) The claim that X is good for one cannot be justified unless one can care about X.
2) If the claim that X is good for one cannot be justified unless one can care about X, then X cannot be good for one unless one can care about X.
3) Therefore, X cannot be good for one unless one can care about X.

The conclusion here amounts to simple internalism. For it is just a statement of the view in the contrapositive. When it comes to premise 1), Rosati's reasons for accepting it are clear from the longer passage just cited. She says there that the fact that a given person can care about X is 'necessary evidence' for the claim that X is good for that person. After all, if it is impossible for a person to care about X, then how could one ever successfully justify to that person that X is good for her? Hence premise 1). Premise 2) is necessary in order to get the argument to be valid. It seems to be a plausible claim considered in its own right. After all, it seems correct that if it is impossible to *justify* a certain claim unless a given state of affairs obtains, then it would follow that this claim cannot be *true* unless that state of affairs obtains. Premise 2) is an instance of this general principle.

Before I argue that the Epistemological Argument is unsuccessful, I should note that Rosati presents a version of the argument specifically in favor of two-tier internalism.[21] I

[21] Rosati puts it this way:

> 'Suppose that a person could be brought to care about something X, but only under conditions C. And suppose that she cannot care about the fact that she would care about X under C, even when she is under ordinary optimal conditions. What counterevidence could be produced to subvert her conclusion that X is not good for her?' (p. 319)

These considerations are supposed to establish two-tier internalism, in particular. There is one main difference between this argument and the original Epistemological Argument in favor of simple internalism. Premise 1) in the original version of the argument stated that the claim that X is good for one cannot be justified unless one can care about X. This premise, in an argument for two-tier internalism, would have to say something like this:

will not discuss this version of the argument here, though, because the problems that undermine the epistemological argument for simple internalism carry over to the epistemological argument for two-tier internalism as well.

The Epistemological Argument is not convincing because it commits the fallacy of equivocation. In particular, the argument seems to equivocate on the notion of justification. Both premises appeal to the idea of 'justifying' the claim that X is good for one. But this might mean one of two things, and there are problems in either case. On the one hand, to justify a claim might mean to justify it *to certain people*, i.e. to *convince* them that it is true. If this is how to understand Rosati's talk of justifying a claim, then the most plausible interpretation of premise 1) would be this: *one cannot be convinced that the claim that X is good for one is true unless one can care about X.* There is some plausibility to this claim.[22] But the problem is that if this is how premise 1) is to be understood, then premise 2) becomes dubious. To maintain the validity of the argument, premise 2) would have to become the following claim: *if one cannot be convinced that the claim that X is good for one is true unless one can care about X, then X cannot be good for one unless one can care about X*. But I see no reason to accept premise 2) on this interpretation of it. After all, the question of what is required for people to be *convinced* of a given claim has no bearing on the question of what is required for that claim to be *true*. After all, a given claim might be true even though nobody can be convinced of it, just as people might be convinced of all sorts of claims that are not true. So from the fact that a given person cannot be convinced of the claim that X is good for her (or, for that matter, from the fact that no one can), it does not follow that X is not good for that person. Thus, premise 2) on this interpretation is implausible.

On the other hand, to justify a given claim might mean to provide *evidence for the truth* of that claim. On this interpretation, premise 2) does indeed seem to be quite

> *Premise 1*):* The claim that X is good for one cannot be justified unless there are some conditions, C, such that i) one would care about X under C and ii) this is a fact that one would care about under ordinary optimal conditions.

The same problems that afflict the original version of the argument, in favor of simple internalism, will afflict this modified version of the argument, in favor of two-tier internalism, as well.

[22] However, I am still inclined to think that premise 1), even on this interpretation, is false. What people can be convinced of (like what they are motivated to do) seems to depend to a large extent on what they believe. If one explicitly believes that externalism is true (i.e. that certain things can be good for one even if one does not care about them), then it might well be possible to convince this person that X is good for her even if she cannot care about X. Thus premise 1), even on this interpretation of it, would be false.

plausible. However, if this is how to understand what it is to justify a claim, then a different problem arises. In particular, the Epistemological Argument becomes question-begging. If justifying a claim is to be understood as providing evidence for the truth of that claim, then premise 1) would become the following: *no evidence can be provided for the truth of the claim that X is good for one unless one can care about X*. But this simply amounts to a statement of simple internalism. For this interpretation of the premise is equivalent to the thought that it cannot be the case that X is good for one unless one can care about X. Thus the first premise on this interpretation just amounts to the conclusion that the Epistemological Argument is intended to establish.

This in itself would not be a major problem if Rosati provided an independent reason for believing premise 1) on this interpretation. But she does not provide any such reason. After all, what Rosati says to back up premise 1) – to the effect that if somebody cannot care about X, then there is no conceivable evidence you could present that person with to cannot 'justify to her' that X is part of her good – does not support the present interpretation of premise 1). That is, what Rosati says does not support the idea that no evidence can be provided *for the truth* of the claim that X is good for one unless one can care about X. Instead, her claims support only premise 1) on the *first* interpretation, i.e. the thought that one cannot be *convinced* that something is good for one unless one can care about that thing. Thus Rosati does not provide any independent rationale for premise 1) on the second interpretation. And so the Epistemological Argument on this second interpretation seems to be straightforwardly question-begging.

Thus there are serious problems for the Epistemological Argument on either interpretation of it: either one of its premises is implausible or else the argument is question-begging. The argument initially seemed promising only because it equivocated on the notion of what it is to justify a claim. Once the equivocation is removed, it becomes clear that the argument the argument fails.

2.4 The Argument from 'Ought' Implies 'Can'

The fourth argument for internalism about a person's good that Rosati develops is based on an ought-implies-can principle. Rosati describes the argument (which she has adopted from David Velleman[23]) as follows:

[23] Cf. Velleman 2000, pp. 85-98

> We think of our good, [Velleman] suggests, as being that which we ought, at least prima facie, to care about. Yet it cannot be that we ought to care about something if we are incapable of caring about it. We can be prima facie obligated to care about something only if it is at least prima facie an option (…). And something can only be prima facie an option for a person, if she is capable of caring about it. Simple internalism thus derives from a plausible rendering of the principle that 'ought' implies 'can'. (p. 320)

The argument Rosati is presenting here seems to be this:

The Argument from 'Ought' Implies 'Can'

1) If X is good for P, then P prima facie ought to care about X.
2) If P prima facie ought to care about X, then P can care about X.[24]
3) Therefore, if X is good for P, then P can care about X.

This argument, the conclusion of which amounts to a statement of simple internalism,[25] does not have much plausibility. But this is because of premise 2), not premise 1). It is reasonable to think that we prima facie ought to care about our own good, since there does seem to be some reason for us to care about our own good. After all, if we didn't care about our own good, we would probably be less happy and successful than if we did care about our own good. So premise 1) in the Argument from 'Ought' Implies 'Can' seems acceptable.

Nonetheless, premise 2) in the argument is false. That one is able to do something is not implied by the fact that one prima facie ought to do it.[26] Here is a simple case that shows why. Suppose Jack has promised to care about something that it is simply impossible for him to care about. Jack's girlfriend really likes Wagner's operas, but Jack does not. He finds them to be tedious. This makes Jack's girlfriend unhappy; she wishes that they liked more of the same things. Since Jack wants to improve his relationship with

[24] In the above passage, Rosati says 'we can be prima facie obligated to care about something only if it is at least *prima facie an option*…' (my italics). Thus one might think premise 2) really should say this instead:

2*) If P prima facie ought to care about X, then P *prima face* can care about X.

However, I do not think this is the most charitable way to interpret the argument. For one thing, this notion of 'prima facie can' seems deeply obscure. What does it mean for one to be prima facie able to do something? I, for one, don't know. Second, if one takes the second premise in the argument to be 2*), then to preserve the validity of the argument, the conclusion would have be this:

3*) Therefore, if X is good for P, then P *prima facie* can care about X.

However, this conclusion does not amount to any version of internalism. For internalism is not formulated in terms of 'prima facie can'. For these reasons, I have chosen to formulate the epistemological argument not in terms of 'prima facie can', but rather in terms of just plain 'can'. What's more, a version of the epistemological argument that proceeds in terms of 2*) and 3*) would still fall prey to the same problems as the standard version (discussed in the body of the text).

[25] Rosati presents a version of this argument in favor of two-tier internalism as well. But it suffers from the same problems that undermine the basic version of the argument, in favor of simple internalism. So there is no need to explicitly discuss the more complicated version of the argument, in favor of two-tier internalism.

[26] This point is argued in detail by Pete Graham in 'Some Thoughts on Prima Facie Moral Obligation' *ms*.

his girlfriend, he promises her that he will care more about opera. As a result of his promise, he incurs a prima facie obligation to care more about opera. After all, promise-making is a paradigmatic source of prima facie obligations.[27] However, suppose that Jack is constitutionally incapable of caring about opera – under *any* circumstances. Accordingly, we have a case in which a person is prima facie obligated to care about something that he simply cannot under any circumstances care about. And so premise 2) in the Argument from 'Ought' Implies 'Can' is false. It is very easy to incur prima facie obligations and it seems that many times we simply will not be able to fulfill the ones we have acquired.

Perhaps this argument could be strengthened if it were modified so that it does not appeal to the notion of prima facie obligation. For instance, one might be tempted to formulate the argument in terms of all-things-considered obligation instead. After all, surely the standard 'ought' implies 'can' principle (OIC) is more plausible than the more exotic 'prima facie ought' implies 'can' principle. Unfortunately, this will not help. If OIC is employed in premise 2), then premise 1) would end up saying this: 'if X is good for P, then P *ought* to care about X.' But this new version of premise 1) is implausible. If something is good for you, that may well be *some* reason for you to care about it, but it clearly does not follow that you would have *most* or *all-things-considered* reason to care about it. Even if X's being good for you provides some reason for you to care about X, there could easily be other more weighty reasons for you not to care about X. In such a case, you would have no all-things-considered obligation to care about X. And so premise 1) would be false.

A more plausible version of argument from 'ought' implies 'can' might appeal to *prudential obligation* instead. What you have a prudential obligation to do is a matter of what would be best for you. The concept of prudential obligation is such that you prudentially ought to bring about the state of affairs that, out of all the ones you are capable of bringing about, would be best for you. Thus 'prudential ought' does plausibly imply 'can.' Now suppose that if X would be best for you, you prudentially ought to care about X. From this (together with the premise that prudential obligation implies can) it

[27] Cf. Ross 1930 (see especially ch. 2, 'What Makes Right Acts Right?').

would follow that if X would be best for you, then you can care about X. Thus a version of internalism about one's good would be true.[28]

The problem with this version of the argument, which appeals to prudential obligation, is that it's not the case that you have a prudential obligation to *care* about what would be best for you. It is true that if a state of affairs would be best for you, then you have a prudential obligation to *bring it about*. But it needn't be the case that you prudentially ought to *care* about it. The 'Hedonic Paradox', a widely discussed phenomenon in the psychological literature on happiness, shows why.[29] The 'Hedonic Paradox' refers to the peculiar psychological fact that if you strive for happiness, you often will not get it. It might be the case that if you care intensely about something, X, that would make you very happy, you will fail to obtain X. Perhaps you will be able to obtain X, thereby increasing your happiness, only if you forget about it and focus on something else instead (e.g. your job, the welfare of other people, etc.). Still, obtaining X might really be best for you. This shows that even if a given thing would be best for you, it does not follow that you prudentially ought to *care* about it. Caring about the thing in question may often lead to a worse outcome for you than not caring about it. Thus formulating the argument from 'ought' implies 'can' in terms of prudential obligation will not save the argument either. For one of the crucial premises in this version of the argument is false.

2.5 The Argument from Autonomy

Rosati thinks that an appeal to autonomy supports internalism about a person's good. Rosati explains the argument as follows:

> The "autonomy-based argument" for internalism is an instance of a more general intuition, namely, that the good of a creature must suit its own nature. In the case of persons or autonomous agents, their nature most centrally includes the capacity for rational self-governance. Their good must thus suit them as creatures with this capacity. (…) Something cannot be a part of a person's good if it cannot

[28] Thanks to Fred Feldman for pointing out this way of formulating the argument and for offering helpful comments about its problems.

[29] See, for example, Nettle 2005, p. 154. John Stuart Mill also recognized this phenomenon: 'I never, indeed, wavered in the conviction that happiness is the test of all rules of conduct, and the end of life. But I now thought that this end 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; on the happiness of others, on the improvement of mankind, even on some art or pursuit, followed not as a means, but as itself an ideal end. Aiming thus at something else, they find happiness by the way.' (Cf. Mill, J.S. *Autobiography*, 1873, Ch. 5)

> enter into her rational self-governance. And it can enter into her self-governance only if she is capable of caring about it. If she is not capable of caring about it, she cannot of her own accord rationally pursue it, promote it, or simply cherish it." (p. 323-324)

I am not entirely sure what Rosati means in talking about things 'entering into someone's rational self-governance.' Still, I think that a charitable way to interpret this phrase would be to take it that something enters into a person's rational self-governance when that person has decided for herself, without undue external influence, to pursue that thing. Assuming that this interpretation is fair, we may state Rosati's argument as follows:

The Argument from Autonomy

1) For any creature, C, if X is good for C, then X 'suits' C's nature.
2) If X 'suits' the nature of a rational, autonomous creature, P, then P could decide for himself without undue external influence to pursue X.
3) If P could decide for himself without undue external influence to pursue X, then P is capable of caring about X.
4) Therefore, if X is good for a rational, autonomous creature, P, then P is capable of caring about X.

The conclusion here amounts to simple internalism, at least as applied to rational, autonomous creatures.[30] It seems to me that if any of Rosati's arguments for internalism has a chance at succeeding, this is the one that does. Those who are committed to the idea that autonomy is tightly connected to a person's good may indeed find internalism to be an attractive view. Nonetheless, I am inclined to think that the Argument from Autonomy fails. While premise 3) seems fairly plausible and I have certain doubts about premise 2),[31] the main problem is that premise 1) seems to be false.

[30] Rosati attempts to extend this argument to cover two-tier internalism as well, but again I focus primarily on the basic version of the argument because the expanded version suffers from the same problems as the basic version.

[31] In particular, there seem to be cases in which something suits the nature of a rational and autonomous person, but where this person cannot *deliberately choose* to pursue this thing. For instance, although (let's suppose) I am a rational and autonomous person, it seems that I cannot deliberately choose to be hungry when I wake up. After all, it is not open to me to *not* be hungry when I wake up in the morning. This is simply something that happens to me because of the way my body works. And yet, being hungry in the morning is surely something that suits my nature to a high degree. Thus there seem to be counter-examples to premise 2).

However, some may not find this sort of counter-example convincing. After all, it relies on the assumption that one cannot deliberately choose to do something unless it is also possible for one to deliberately choose *not* to do it. But I suspect that not everyone will accept this assumption. So perhaps my counter-example will not be a conclusive objection to premise 2) in the Argument from Autonomy.

Insofar as I have an intuitive understanding of the notion of something's *suiting* the nature of a given creature,[32] it seems to me to be false that something can be good for a creature only if it suits that creature's nature. Accordingly, I reject premise 1). To see why, consider the case of Mark. Mark is a very disturbed guy who has an incurable fascination with torturing little animals. It gives him large amounts of pleasure to torture squirrels, cats and the like. His habit gets him into a lot of trouble with the other members of his community who frown on this sort of thing. Mark has been given extensive therapy and all sorts of drugs, but nothing is successful in ridding him of his penchant for torturing furry animals. In this scenario, it seems that Mark would have been much better off if he had not enjoyed torturing animals so much. It seems that not having an incurable fascination with watching animals writhe in pain would be both *intrinsically* and *instrumentally* good for Mark. Nonetheless, not torturing animals would fit very poorly with Mark's nature. It would not 'suit' his nature to refrain from torturing them. After all, this is what he enjoys and chooses to do. It is what he would do, absent the interference of others. No amount of psychotherapy or drugs can eliminate Mark's fascination with animal torturing. In fact, the only way to get him to refrain from doing it would be to force him to undergo massive brain surgery. So we seem to have a case in which something would be very good for a person, both instrumentally and intrinsically, but that completely fails to suit that person's nature. Because of cases like this, I think line 1) in the Argument from Autonomy is false.

3. An Argument against Internalism

Not one of the five arguments that Rosati discusses succeeds in establishing that some form of internalism about one's good is true. Of course, this by itself does not establish that internalism about one's good is *false*. Nonetheless, I will argue in this section that one ought to be pessimistic about the prospects for finding a satisfactory version of internalism about one's good. We saw in section 1) of this paper that Rosati presents a

[32] Clearly it would not suit the nature of a fish to find itself out of water. Being in water would surely be better suited to the nature of a fish. But beyond such obvious cases, how are we to decide in a systematic way whether something is well suited or poorly suited to a given creature's nature? I suspect that we will have a strong intuition that X suits the nature of creature C if and only if we have a strong intuition that X is good for C. Thus I am not convinced that one can argue in a non-question-begging way from intuitions about what suits the nature of a creature to claims about what can or cannot be good for that creature.

plausible case for the claim that *if* one accepts the basic internalist intuition, then one would do well to prefer some strong version of internalism, like two-tier internalism. But there is in fact good reason to think that two-tier internalism is false. In particular, I will argue that two-tier internalism is not consistent with any first-order theory of welfare.[33] Thus if Rosati is right that those who accept the internalist intuition should prefer something like two-tier internalism, then my argument should cast significant doubt on the prospects for internalism about a person's good, in general.[34]

If two-tier internalism were true, it would place a constraint on the true theory of welfare. In particular, the true theory of welfare would have to be consistent with two-tier internalism. Rosati acknowledges this herself by saying that '[i]t is important to see that two-tier internalism most directly tests theories rather than alleged goods.' (p. 308) It might be natural to think that the constraint on the true theory of welfare imposed by two-tier internalism would tend to favor theories of welfare that are traditionally taken to be *subjective*, like Hedonism or Desire Satisfactionism. Roughly, subjective theories of a

[33] I suspect that a problem similar to the one I raise for two-tier internalism will threaten most other formulations of 'stronger internalism' as well (i.e. a version that is restrictive enough to adequately capture the basic internalist intuition). I will not explicitly argue for this here, however. My discussion of the problem for two-tier internalism should be sufficient to indicate how the problem would carry over to the other strong versions of internalism on offer. Note, however, that the problem I raise for two-tier internalism will *not* afflict something as weak as *simple internalism*. After all, *every* theory of welfare is going to be consistent with simple internalism. Nonetheless, we saw before (in section 1) that simple internalism was unsatisfactory for other reasons. In particular, as Rosati argues, it is too permissive to adequately capture the basic internalist intuition. Thus the prospects for internalism about a person's good in general do not seem promising.

[34] Note that two-tier internalism has some technical problems as well, in addition to the substantive problem I present in the body of the text. However, I think that more work could probably yield a version of two-tier internalism that avoids these technical problems. (Thanks to Fred Feldman for pointing these problems out to me.)

What are these technical problems? The second condition of two-tier internalism is intended to pick out the 'appropriate conditions,' i.e. the counterfactual conditions in which your preferences would be relevant to determining your good. However, it does not seem that the second 'tier' in two-tier internalism succeeds in picking out the 'appropriate conditions.' For one thing, suppose you have not given any thought to the question of which conditions are relevant for determining your good. In that case, there might be no counterfactual conditions, C, such that in ordinary optimal conditions you would care what your preferences would be in C. As a result, nothing would be good for you. Second, suppose you are deeply confused and believe that the ideal conditions for determining your good are conditions in which you have fasted for ten days. You think that when you are delirious with hunger, you are afforded a special insight into the nature of the universe so that your preferences are optimally attuned to your good. Suppose you would go on thinking this even if you were in ordinary optimal conditions. In this case, two-tier internalism would imply that the 'appropriate conditions' for you are conditions in which you are delirious with hunger. But intuitively this does not seem plausible. It is unlikely that conditions of extreme hunger are appropriate for determining what your good is – even if you would care about your preferences in these conditions, while in ordinary optimal conditions.

person's good are the ones according to which the question of whether a given state of affairs is good for one is determined by the facts about what one's psychological responses to things are (or would be). The objective theories are the ones that deny this. Now, if two-tier internalism is true, then the true theory of welfare would have to imply that one must care about or have some positive response to a given thing (at least under certain ideal circumstances) in order for that thing to be good for one. Thus two-tier internalism might be thought to favor the subjective theories, which make one's good depend on one's responses, over the objective theories, which do not. However, it would be a mistake to think this. For on closer inspection, there in fact seem to be *no* substantive theories of welfare that are consistent with two-tier internalism.

Granted, the objective theories do not satisfy the constraint imposed by two-tier internalism. To see this, consider a simple objective theory like the Money Theory, according to which the only thing that is intrinsically good for a person is money. Now consider Siddhartha who not only is completely unconcerned with money in his *actual* circumstances, but would remain entirely unconcerned with money even if he were transported into the set of privileged counterfactual conditions that he cares about under ordinary optimal conditions. As a result, if two-tier internalism is true, getting more money would not enhance Siddhartha's welfare. But that means that the Money Theory cannot be true. For the Money Theory entails that getting more money *does* enhance Siddhartha's welfare. Since there clearly could be people like Siddhartha, two-tier internalism entails that the Money Theory cannot be the true theory of welfare.

Not only are objective theories like the Money Theory inconsistent with two-tier internalism, but many (perhaps all) subjective theories are inconsistent with it as well. I will run through three examples, which I hope will be a representative sample. First, consider Sensory Hedonism, according to which the only thing that is intrinsically good for a person is to experience pleasure. Sensory Hedonism is not consistent with two-tier internalism because there could be people who do not actually desire pleasure, nor would do so even if placed in appropriate counterfactual conditions.

To see an example of such a case, consider Fred Feldman's example of Stoicus.[35] Stoicus, after much soul-searching, decides that he doesn't ever want to experience

[35] Cf. Feldman 2004, pp. 49-50.

sensory pleasure again. He practices hard and is never again tempted by the prospect of experiencing sensory pleasure. Stoicus is fully rational. He is intelligent and cool and calculated and reflected. He never flies off the handle irrationally. Thus Stoicus is in ordinary optimal conditions. Suppose just for simplicity that under these ordinary optimal conditions, Stoicus cares about what a more intelligent and informed version of himself would care about as someone about to assume his actual position. So the second condition in two-tier internalism is fulfilled. However, Stoicus is a guy who not only *actually* has no desire for sensory pleasure, but also would not have any desire for pleasure if placed in conditions of enhanced intelligence and information. Nothing you can do to Stoicus (short of subjecting him to hypnosis or brain surgery) could make him want to experience pleasure. Accordingly, if two-tier internalism is true, sensory pleasure could not be part of Stoicus' good. However, Sensory Hedonism implies that pleasure would indeed be good for Stoicus. So it follows that if two-tier internalism is true, Sensory Hedonism is false. Thus two-tier internalism rules out Sensory Hedonism as a candidate for the true theory of welfare.

An analogous case shows that two-tier internalism is also inconsistent with Desire Satisfactionism. Desire Satisfactionism is roughly the theory that what intrinsically enhances your welfare is getting the things that you desire. According to this theory, the items that are the fundamental bearers of intrinsic welfare value are states of desire satisfaction. Two-tier internalism implies that Desire Satisfactionism is false because there could be a person who would not have a desire for episodes of desire satisfaction, even if placed in appropriate counterfactual conditions.

To see this, consider Stoicus Jr. He is a person who, after much soul-searching, decides that he doesn't want to have any of his desires satisfied ever again. In other words, he develops an overwhelming second-order desire not to have any of his first-order desires satisfied. He practices hard and is never again tempted by the prospect of getting his first-order desires satisfied. Stoicus Jr. is intelligent and cool and calculated and reflected. He never flies off the handle irrationally. Thus Stoicus Jr. is in ordinary optimal conditions. Suppose again for simplicity that under these ordinary optimal conditions, Stoicus Jr. cares about what a more intelligent and informed version of himself would care about as someone about to assume his actual position. So condition 2 in two-tier internalism is fulfilled. However, not only does Stoicus Jr. have no *actual* desire for episodes of desire satisfaction, but he would not have a desire for desire satisfaction even if he were placed in conditions of enhanced intelligence and information. Stoicus Jr.'s aversion to the idea of getting his desires satisfied is so strong that even if he were placed in the appropriate counterfactual conditions – i.e. the sort that he actually cares about – and then told that he's about to be placed into the position of his actual self, he would still have no desire whatsoever for episodes of desire satisfaction. Nothing you can do to Stoicus Jr. (short of subjecting him to hypnosis or brain surgery) could make him desire to get his desires satisfied. The thought of it disgusts him. Accordingly, if two-tier internalism is true, episodes of desire satisfaction could not be part of Stoicus Jr.'s good. But Desire Satisfactionism implies that this indeed is constitutive of his good. So if two-tier internalism is true, then Desire Satisfactionism is false. Two-tier internalism rules out Desire Satisfactionism as a candidate for the true theory of welfare.

To this, some might object that I have misunderstood Desire Satisfactionism. I took it that Desire Satisfactionism entails that the bearers of intrinsic welfare value are episodes of desire satisfaction (i.e. complex states of one's desiring that p is true and p's really being true). But perhaps one thinks that Desire Satisfactionism should be understood in a different way. In particular, one might want to take the theory to state that the bearers of intrinsic welfare value are the *things* that you desire (provided you obtain them). If one understands Desire Satisfactionism in this way, then Desire Satisfactionism might not be inconsistent with two-tier internalism. After all, while Stoicus Jr. would not desire episodes of desire satisfaction even under ideal conditions, he does desire some *things*. Sometimes he wants to sleep. Sometimes he wants to meditate. Sometimes he wants to not be desiring anything. Desire Satisfactionism, understood in this new way, would indeed entail that these things can themselves be good for Stoicus Jr. provided he gets them. The idea would be that while episodes of desire satisfaction can't be good for Stoicus Jr. since he doesn't want them, other things that he desires may still be good for him. Now, two-tier internalism also allows that these other things may be good for Stoicus Jr. After all, a more enlightened version of himself would desire meditation and to not be desiring anything. Thus two-tier internalism would be consistent with the consequences of Desire Satisfactionism understood in this new way.

However, this new way of understanding Desire Satisfactionism is implausible. We must take the bearers of intrinsic welfare value to be episodes of desire satisfaction. After all, if one takes it that the bearers of intrinsic welfare value are the *things* that are desired and obtained, then Desire Satisfactionism would conflict with the widely accepted axiological assumption that the intrinsic value of something can depend only on its intrinsic features.[36] Suppose Desire Satisfactionism were taken to state that the bearers of intrinsic welfare value are the *things* you desire. In that case, when you desire an apple and get it, what would be good for you is the apple. But in that case, the intrinsic value of the apple for you would *not* depend only on the *intrinsic* features on the apple. It would also depend on its relation to you and your desires. But this is an extrinsic feature of the apple. And so the axiological principle that intrinsic value must depend on intrinsic features would be violated. Thus Desire Satisfactionism cannot be taken to state that the bearers of intrinsic welfare value are the *things* that are desired. Instead, the theory must be taken to state that the bearers of welfare value are episodes of desire satisfaction. For the theory thus understood does not violate the intuitive axiological principle that the intrinsic value of something can depend only on its intrinsic features. Thus the only plausible interpretation of Desire Satisfactionism would be the one that I argued above is inconsistent with two-tier internalism.

So I take it that some common subjective theories of welfare (viz. Sensory Hedonism and Desire Satisfactionism) are inconsistent with two-tier internalism. But are they all? Perhaps there is only one sort of theory that is consistent with two-tier internalism. In

[36] See, for example, Feldman 2004, p. 73 and Bradely 2009, p. 19. Bradley offers the following argument for the principle (which he calls SUP) that the intrinsic value of something depends solely on its intrinsic properties:

> 'SUP is a requirement of any acceptable theory of well-being. This is because, as noted above, the value atoms should be *instantiations of the fundamental good- or bad-making properties* – the properties that are fundamentally and completely responsible for how well a world (or a life, or …) goes. Suppose SUP were false. Then there could be two properties, F and G, such that the only intrinsically good states of affairs are those involving the instantiation of F alone, but whose values are determined by whether there are any instantiations of G. But if that were true, F would fail to be a fundamental good- or bad-making property, for instantiations of F would fail to completely determine what value there is. The fundamental good- or bad-making property would involve both F and G, contrary to our assumption. Once we are committed to the project of finding the fundamental good- and bad-making properties – the fundamental project of axiology, and of the theory of well-being – we are immediately committed to SUP…' (Bradley, p. 19)

particular, this would be the theory that something is good for a person if and only if it satisfies the conditions specified by two-tier internalism. To state the theory precisely, let me introduce the notion of a *counterfactually sanctioned desire*. Consider the counterfactual conditions, C, that you would care about under ordinary optimal conditions. Suppose you are placed in C and then told that you are about to assume your actual self's position. Any desire that you would have in that case would be a *counterfactually sanctioned desire*. Now, the theory in question, which I will call *Counterfactually Sanctioned Desire Satisfactionism* (CSDS), is that what intrinsically enhances your welfare is getting your counterfactually sanctioned desires satisfied. The intrinsic bearers of welfare value, on this theory, are states of counterfactually sanctioned desire satisfaction.

CSDS is specifically designed to meet the conditions of two-tier internalism. If any theory can accommodate two-tier internalism, it seems CSDS would have to be it. However, (perhaps surprisingly) CSDS does not satisfy the requirements of TTI either. The reason is that there could easily be a person who would have *no counterfactually sanctioned desire* for episodes of counterfactually sanctioned desire satisfaction. To put it more simply, imagine a person like the following. Call him Jerry. Under ordinary optimal conditions, Jerry cares about what an ideal counterpart of himself would desire under a certain privileged set of counterfactual conditions, C. Now suppose Jerry is placed in C and is told that he is about to assume his actual self's position. As a result Jerry, while in C, comes to desire just four things: love, knowledge, health and happiness. But notice that nowhere on this list do we find a desire for the satisfaction of counterfactually sanctioned desires. What does this mean? It means that if two-tier internalism is true, then while love, knowledge, health and happiness might be good for Jerry, the satisfaction of counterfactually sanctioned desires cannot be good for Jerry. Thus if two-tier internalism is true, the very thing that CSDS says is intrinsically good for a person – i.e. episodes of counterfactually sanctioned desire satisfaction – cannot be good for Jerry. So because cases like that of Jerry are possible, if two-tier internalism is true, then CSDS is not true. Thus not even CSDS satisfies the constraint imposed by two-tier internalism.

I have tried to think of a theory of welfare that is consistent with two-tier internalism, but I could not do it. In general, it seems that for any theory of welfare one could dream up, there could be a person who simply would not be concerned (in the sense that two-tier internalism specifies) with the thing(s) that this theory identifies as the fundamental bearer(s) of welfare value. So I am inclined to think that no theory of welfare is consistent with two-tier internalism. Thus we get the following argument against two-tier internalism:

The Argument from Insatiability

1) If two-tier internalism is true, then no theory of welfare is true.
2) It's not the case that no theory of welfare is true.
3) Therefore, it's not the case that two-tier internalism is true.

If this argument is sound, then not only does there not seem to be any successful argument in favor of internalism about a person's good, but there is also a good positive reason to think that the version of internalism that has most going for it, viz. two-tier internalism, is false. Rosati argued convincingly that two-tier internalism, or something like it, is the version of internalism that should be endorsed by those who accept the basic internalist intuition. But if she is right about this, then my Argument from Insatiability would cast significant doubt on the prospects for internalism about a person's good, in general.

APPENDIX II

THE MATHEMATICS OF THE THRESHOLDS THEORY'S WELFARE-FUNCTION

The welfare function employed by the Thresholds version of the Happiness and Success Theory must meet the eight conditions stated in section 8.3.2. However, there are many ways (infinitely many, in fact) to define a function mathematically that meets conditions 1)-8). In this appendix, I discuss what seems to be one good way to do this.

To begin with, recall the behavior that the Thresholds theory is supposed to display. First, whenever both your happiness, h, and your achievement, a, are above the respective thresholds, your welfare, w, is supposed to equal $h + a$. (Cf. condition 5.) Similarly, whenever both h and a are below their respective thresholds, w is supposed to equal $h + a$. (Cf. condition 6.) The surprising behavior occurs when one variable is above threshold and the other is below. So if h is above threshold and a is below, then w cannot be greater than the minimally good life point, w_t. Similarly, if a is above threshold and h is below, then w cannot be greater than w_t in this case either. (Cf. condition 4.) Finally, if h is below threshold, increasing a will cause w to approach w_t, but w will never reach w_t. (Cf. condition 7.) Similarly if a is below threshold, then increasing h will cause w to approach w_t, but w will never reach w_t.

The following function meets all of these specifications.[1] (Moreover, it also avoids some problems with other preliminary functions that were tried before this one. We needn't go into these problems here, though.)

[1] Again, I am heavily indebted to John Arthur Skard designing the function presented here, as well as for programming the graphs in Excel. (Thanks also to Chris Meachem for some helpful preliminary proposals that eventually led to the development of the final version of the function, which is presented here.)

(8)

- IF ($a \geq a_t$ and $h \geq h_t$) OR ($a < a_t$ and $h < h_t$):

$$W = h + a$$

- ELSE:

$$W = h_t + a_t + \frac{h - h_t}{2} + \frac{a - a_t}{2} + \frac{1}{2}\sqrt{(h - h_t)^2 + (a - a_t)^2}$$

When graphically represented, the function looks like this. (The happiness threshold, h_t, and the achievement threshold, a_t, are both set at 1000. The minimally good life point, w_t, is set at $h_t + a_t = 2000.$)

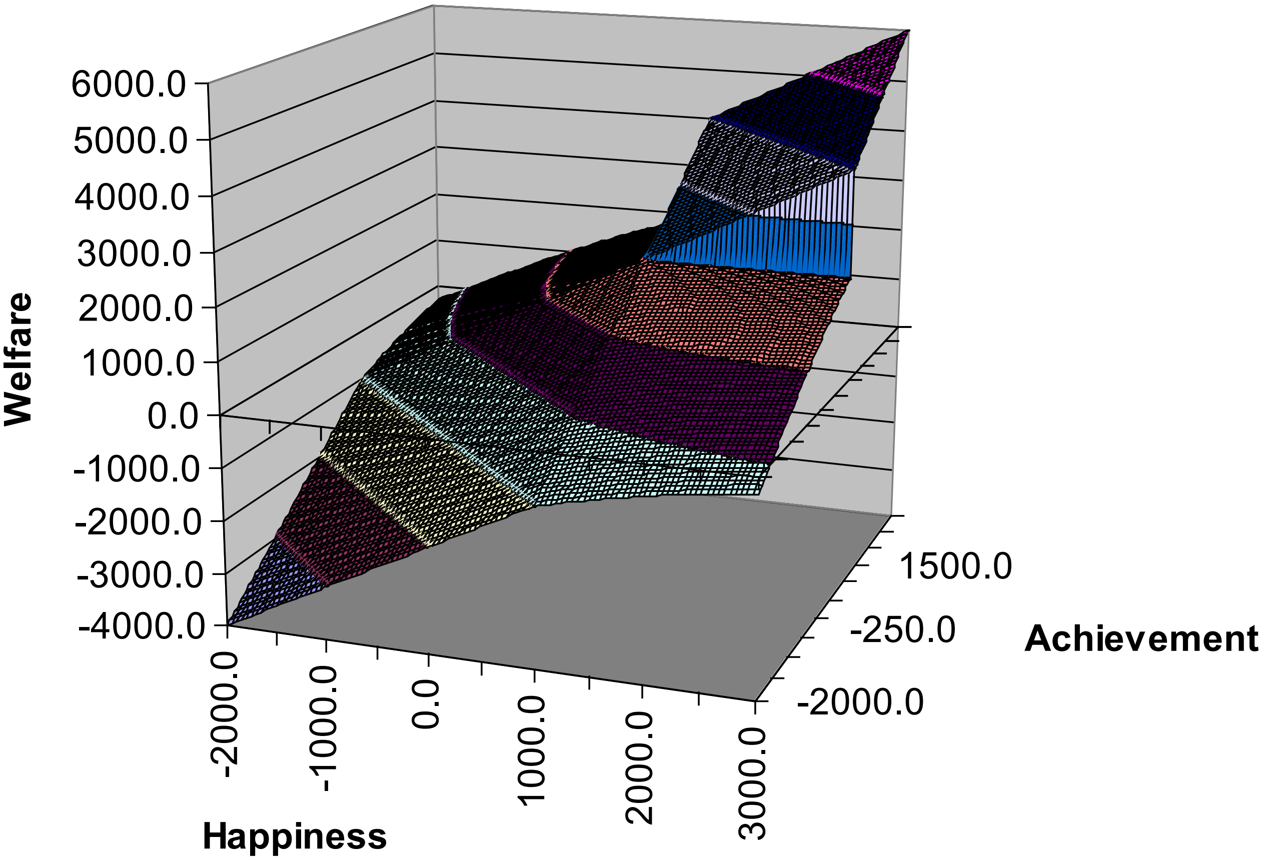

Figure 8: The Thresholds Theory

Function (8) is the best representation of the Thresholds theory I know of. It displays pure h+a behavior whenever both variables are above threshold, or both are below threshold. In the regions where either h or a is above threshold, but not both, then welfare, w, increases asymptotically towards w_t when one of the variables is increased to the threshold level. Thus if, say, h is above threshold and a is below, then w increases to w_t when a approaches a_t from below. Similarly, if a is above threshold and h is below, then w increases to w_t when h approaches h_t from below. This is exactly the sort of behavior that the Thresholds theory is supposed to display. What's more, this version of the function avoids various discontinuities or 'kinks' that were generated by other functions that were tried.

However, this function let us clearly see the major flaw in the Thresholds theory, namely the implausibly big 'jumps' in welfare that it permits. In particular, when you are above threshold on one scale but just below threshold on the other scale, then it will be possible for very small increases on the latter scale – increases just large enough to get you over the threshold on that scale – to lead to dramatic increases in your welfare. This can be clearly seen from the graph. In particular, look for the 'vertical wall' that is visible in the upper right hand corner. All mathematical representations of the Thresholds theory will display this odd behavior. The Discount/Inflation Theory does not generate these strange jumps in welfare, however.

APPENDIX III

THE MATHEMATICS OF THE DISCOUNT/INFLATION THEORY'S WELFARE-FUNCTION

There are many ways (infinitely many, in fact) to define a function mathematically that meets conditions i-xx) stated in section 8.4.2. In this appendix, I present what I take to be one very good attempt to do this.[1]

We saw in 8.4.2 that the basics of the welfare-function involved in the Discount/Inflation theory can be captured by the following equations:

(1) $w = W_h + W_a$

(2a) $W_h = h^{f(a)}$, when $h \geq 0$

(2b) $W_h = -(|h|^{f(a)})$, when $h < 0$

(3a) $W_a = a^{g(h)}$, when $a \geq 0$

(3b) $W_a = -(|a|^{g(h)})$, when $a < 0$

But what functions should be substituted for '*f(a)*' and '*g(h)*' here? They must be such that they meet conditions i)-x) and xi)-xx), respectively.

In the antecedents of the conditions that *f(a)* must meet, mention is made of whether a is above or below threshold a_t. Thus it will be convenient to let *f(a)* be a function not of a in isolation, but rather of the *distance* between a and a_t. With that proviso, here is a function that meets conditions i)-v), and which therefore can be used in the cases when h is a positive number:

[1] The function presented in this section was constructed by John Arthur Skard. I am extremely grateful to him for all his help.

$$f(a) = \left(\frac{e^{k_1(a-a_t)} - 1}{e^{k_1(a-a_t)} + 1} + 1 \right) \qquad (4)$$

This function uses the natural logarithm base *e* just for convenience; some other constant could just as easily have been used instead. Moreover, this function uses a scale factor k_1. This is necessary in order to control how quickly *f(a)* will change when the variable *a* changes. Since the only scale known from the outset here is the threshold value, a_t, the change should happen over a scale that is comparable to this value. It seems reasonable to let $k_1 \approx 1/a_t$. The larger value of k_1, the more 'violent' the function's behavior becomes. For a threshold value of $a_t = 1000$, for example, this means $k_1 = 0.01$, approximately. However, we can let the scale-factor k_1 be whatever is necessary to make *f(a)* vary on a scale that is commensurate with whatever value is chosen for the threshold value a_t.

Function (4) meets conditions i)-v). When $a = a_t$, then $a – a_t = 0$, and so the whole function becomes equal to 1. When *a* is greater than a_t, then the function returns a value greater than 1. But when *a* is less than a_t, then the function returns a value less than 1. What's more, the function makes use of the term $e^{(a-at)}$, which is good for generating a function that asymptotically approach some value. Thus this function approaches 0 as *a* goes to minus infinity, and it approaches 2 as *a* approaches plus infinity. Thus all of conditions i)-v) are met. The function looks like this when graphically represented, where *a* is plotted on the x-axis and the value for *f(a)* is plotted on the y-axis:

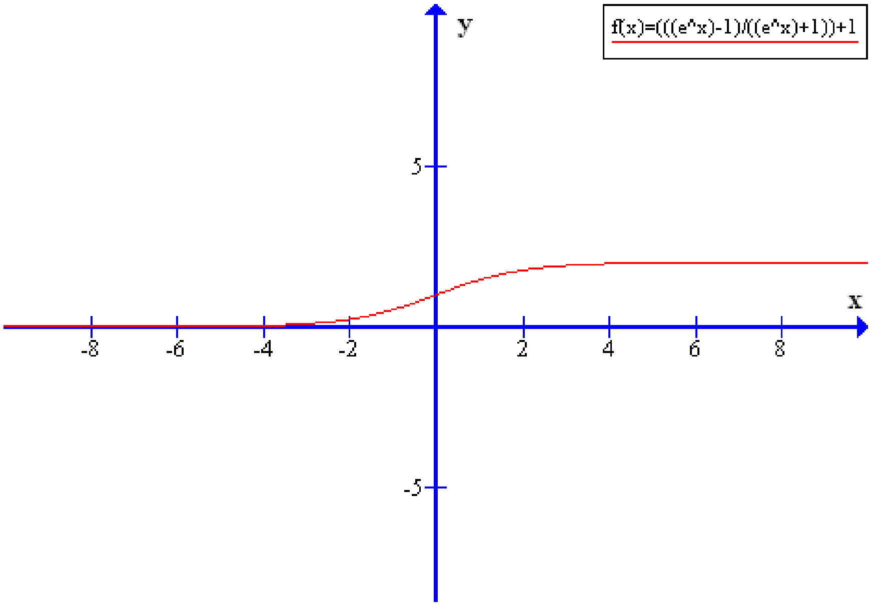

Figure 9: *f(a)* when *h* > 0

However, the function in (4) gives us only half of the required behavior for *f(a)*. We also need a function that meets conditions vi)-x), which can be used to capture the behavior of *f(a)* when *h* is less than zero. Here is a function that does the job:

$$f(a)=\left(\frac{e^{-k_1(a-a_t)}-1}{e^{-k_1(a-a_t)}+1}+1\right)$$

(5)

This function meets conditions vi)-x). When $a=a_t$, then $a–a_t = 0$, and so the whole function becomes equal to 1. When *a* is greater than a_t, then the function returns a value less than 1. But when *a* is less than a_t, then the function returns a value greater than 1. What's more, the function approaches 0 as *a* goes to infinity, and it approaches 2 as *a* approaches minus infinity. Thus all of conditions vi)-x) are met. The function looks like this when graphically represented, where *a* is plotted on the x-axis and the value for *f(a)* is plotted on the y-axis:

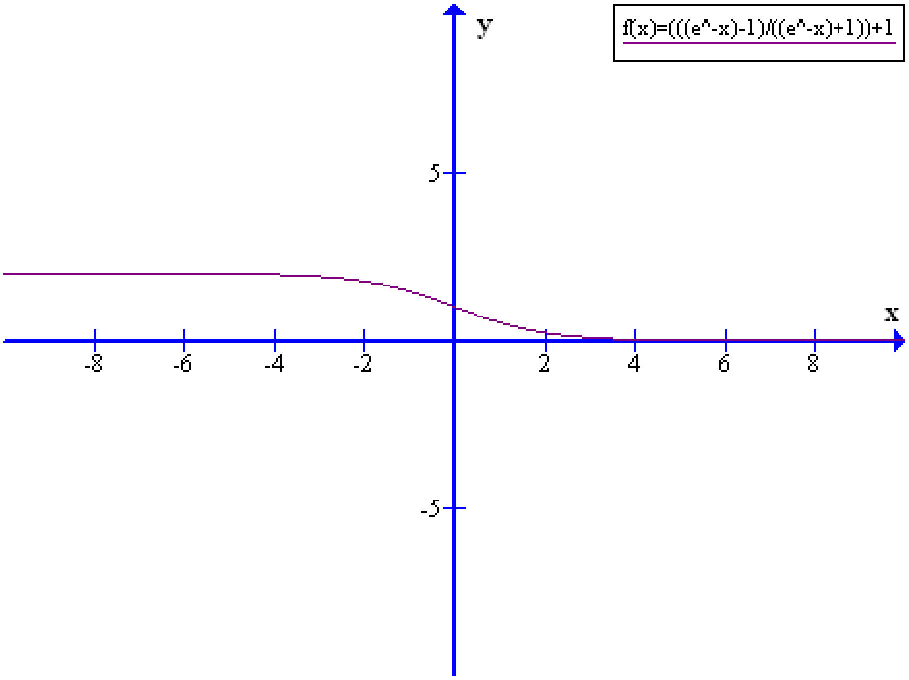

Figure 10: *f(a)* when *h* < 0

The function in (4) can be combined with the function in (5) to fully capture the required behavior for *f(a)*. Two analogous functions can easily be constructed to capture the required behavior of *g(h)*, as specified by conditions xi)-xx).

Thus we are in a position to fully formulate a satisfactory version of the welfare function that the Discount/Inflation theory employs. (Here W_h and W_a are written out in full, instead of employing the simplifying devices of *f(a)* and *g(h)*.)

(1) $w = W_h + W_a$

(6) If $h \geq 0$, then

$$W_h = h^{\left(\frac{e^{k_1(a-a_t)}-1}{e^{k_1(a-a_t)}+1}+1\right)}$$

If $h < 0$, then

$$W_h = -(-h)^{\left(\frac{e^{-k_1(a-a_t)}-1}{e^{-k_1(a-a_t)}+1}+1\right)}$$

(7) If $a \geq 0$, then

$$W_a = a^{\left(\frac{e^{k_2(h-h_t)}-1}{e^{k_2(h-h_t)}+1}+1\right)}$$

If $a < 0$, then

$$W_a = -(-a)^{\left(\frac{e^{-k_2(h-h_t)}-1}{e^{-k_2(h-h_t)}+1}+1\right)}$$

(Depending on what sort of units one picks to represent amounts of happiness and achievement, one have to pick appropriate values for k_1 and k_2.)

It will perhaps be helpful to see a graphic representation of this welfare function. Since the function takes welfare to depend on two variables, happiness and achievement, the function describes a surface in three dimensions. Here is a representation of the welfare function, which has been appropriately scaled.[2] The happiness threshold and the achievement threshold are both set at 1000. (An appropriate scale factor has also been used: $k_1 = k_2 = 0.01$)

The Discount/Inflation Theory:
welfare as a function of happiness and achievement; thresholds set at 1000

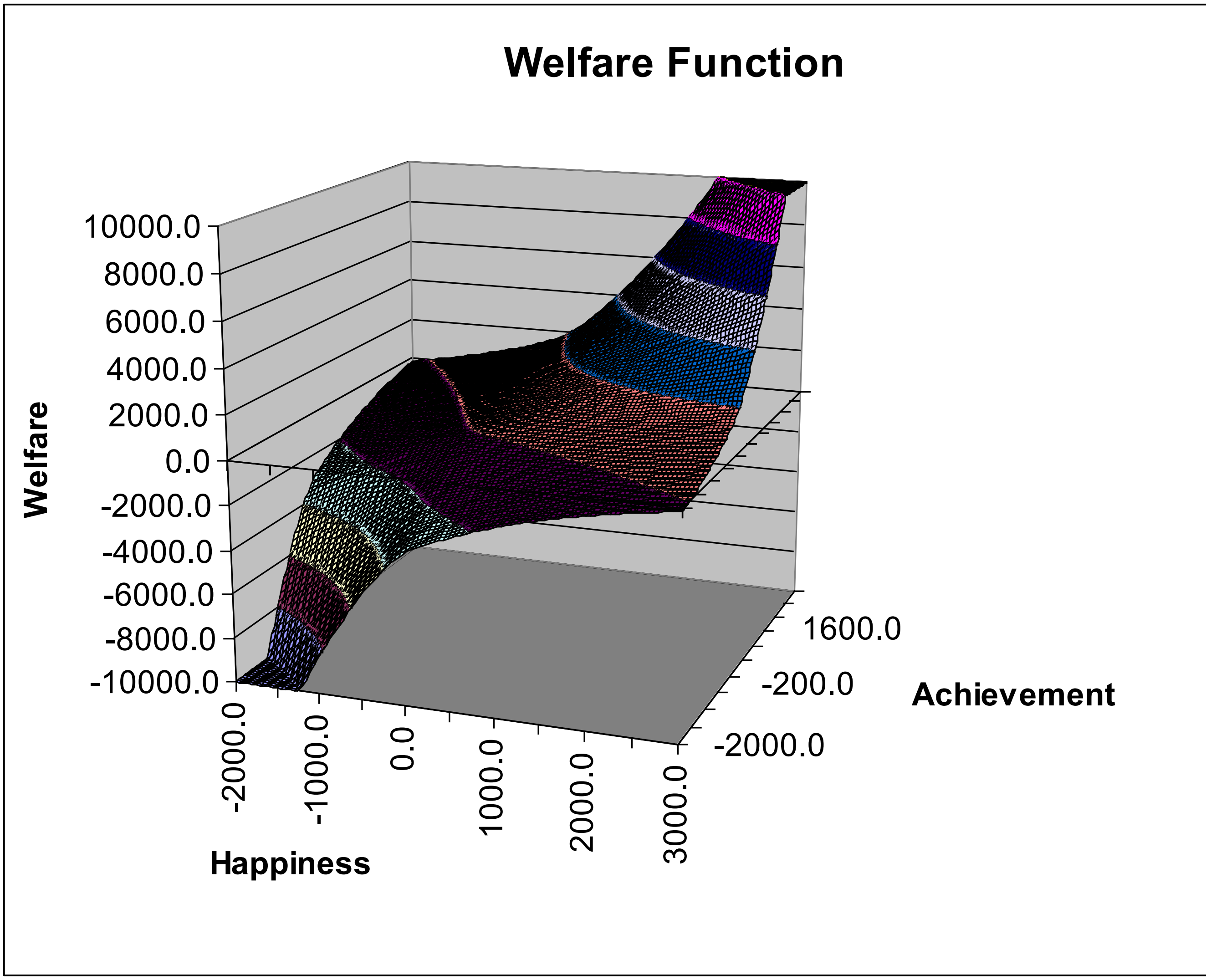

Figure 11: The Discount/Inflation Theory

[2] Thanks again to John Arthur Skard for programming this diagram in Excel.

BIBLIOGRAPHY

Adams, R.M. 1999, *Finite and infinite goods : a framework for ethics,* Oxford University Press, New York.

Bealer, G. 1999, "A Theory of the A Priori (Volume 13: Epistemology)", *Nous-Supplement: Philosophical Perspectives,* vol. 13, pp. 29-55.

Bealer, G. 1998, "Intuition and the Autonomy of Philosophy" in *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry* Rowman & Littlefield, Lanham.

Bealer, G. 1996, "'A Priori' Knowledge and the Scope of Philosophy", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* vol. 81, no. 2-3, pp. 121-142.

Blackburn, Simon. 1993, *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Boyd, R. 1988, "How to be a Moral Realist" in *Moral Realism*, ed. G. Sayre-McCord, Cornell University Press, Ithaca, pp. pp. 181-228.

Bradley, B. 2008, "A Paradox for Some Theories of Welfare", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* .

Bradley, Ben. 2009, *Well-Being and Death*, Oxford University Press.

Brandt, R. 1979, *A Theory of the Good and the Right,* Clarendon Press, Oxford.

Bricker, P. 1980, "Prudence", *Journal of Philosophy,* vol. 77, pp. 381-400.

Brink, D.O. 1989, *Moral Realism and the Foundations of Ethics,* Cambridge University Press, Cambridge.

Brink, D.O. 1984, "Moral Realism and the Sceptical Arguments from Disagreement and Queerness", *Australasian Journal of Philosophy,* vol. 62, pp. 111-125.

Bykvist, K. 2003, "The Moral Relevance of Past Preferences" in *Time and Ethics: Essays at the Intersection*, ed. H. Dyke, Kluwer Academic Publishers, Dordrecht.

Bykvist, K. 2007, "Comments on Dennis McKerlie's 'Rational Choice, Changes in Values Over Time, and Well-Being'", *Utilitas: A Journal of Utilitarian Studies,* vol. 19, no. 1, pp. 73-77.

Bykvist, K. 2006, "Prudence for Changing Selves", *Utilitas: A Journal of Utilitarian Studies,* vol. 18, no. 3, pp. 264-283.

Carson, T.L. 1981, "The 'Ubermensch' and Nietzsche's Theory of Value", *International Studies in Philosophy,* vol. 13, pp. 9-30.

Carson, T.L. 2000, *Value and the Good Life,* Univ Notre Dame Pr, Notre Dame.

Copp, D. 2004, "Three Grades of Normativity" in *Naturalism and Normativity*, ed. P. Shaber, Ontos-Verlag, Frankfurt.

Crisp, R. Dec. 9, 2008-last update*, Well-Being* [Homepage of Stanford Encyclopedia of Philosophy], [Online]. Available: http://plato.stanford.edu/entries/well-being/Apr. 25, 2009] .

Cummins, R. 1998, 'Reflection on Reflective Equilibrium' in Rethinking Intuition, ed. M. DePaul and W. Ramsey, Roman & Littlefield, Oxford.

Daniels, N. 1979, "Wide Reflective Equilibrium and Theory Acceptance in Ethics", *Journal of Philosophy,* vol. 76, pp. 256-282.

Darwall, S. 1983, *Impartial Reason,* CORNELL UNIV PR, ITHACA.

Darwall, S. 2002, *Welfare and Rational Care,* Princeton Univ Pr, Princeton.

DePaul, M.R. 2002, "A Half Dozen Puzzles Regarding Intrinsic Attitudinal Hedonism", *Philosophy and Phenomenological Research,* vol. 65, no. 3, pp. 629-635.

Feldman, F. 2009, *What is This Thing Called Happiness?* forthcoming.

Feldman, F. 2006, Review of Darwall, "Welfare and Rational Care", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* vol. 130, no. 3, pp. 585-601.

Feldman, F. 2004, *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism,* Clarendon Press, Oxford.

Feldman, F. 2002, "The Good Life: A Defense of Attitudinal Hedonism", *Philosophy and Phenomenological Research,* vol. 65, no. 3, pp. 604-628.

Feldman, F. 1997, "On the Intrinsic Value of Pleasures", *Ethics: An International Journal of Social, Political, and Legal Philosophy,* vol. 107, no. 3, pp. 448-466.

Feldman, F. 1986, *Doing the Best we Can: an Essay in Informal Deontic Logic,* REIDEL, DORDRECHT.

Foot, P. 2001, *Natural Goodness,* Clarendon Press, Oxford.

Frankena, W., K. 1963, *Ethics,* PRENTICE HALL, ENGLEWOOD CLIFFS NY.

Gibbard, A. 1990, *Wise Choices, Apt Feelings: A Theory of Normative Judgment,* Harvard Univ Pr, Cambridge.

Griffin, J. 1986, *Well-being: Its Meaning, Measurement and Moral Significance,* Oxford University Press.

Hare, R.M. 1981, *Moral Thinking: Its Levels, Methods and Point,* Oxford University Press.

Hare, R.M. 1989, "Prudence and Past Preferences; Reply to Wlodizimierz Rabinowicz", *Theoria: A Swedish Journal of Philosophy,* vol. 55, pp. 152-158.

Hare, R.M. 1989, "Prudence and Past Preferences; Reply to Wlodizimierz Rabinowicz", *Theoria: A Swedish Journal of Philosophy,* vol. 55, pp. 152-158.

Harman, G. 1977, *The Nature of Morality: An Introduction to Ethics*, Oxford University Press.

Haybron, D.M. 2007, "Well-Being and Virtue", *Journal of Ethics and Social Philosophy: Journal of Moral, Political and Legal Philosophy,* vol. 2, no. 2, pp. 1-24.

Haybron, D.M. " Happiness, the Self, and Human Flourishing", forthcoming, *Utilitas*.

Heathwood, C. 2006, "Desire Satisfactionism and Hedonism", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* vol. 128, no. 3, pp. 539-563.

Heathwood, C. 2005, "The Problem of Defective Desires", *Australasian Journal of Philosophy,* vol. 83, no. 4, pp. 487-504.

Heathwood, C. "Subjecive Desire Satisfactionism", *manuscript*.

Hooker, B. 1996, "Does Moral Virtue Constitute a Benefit to the Agent?" in *How Should One Live?: Essays on the Virtues* Clarendon Press, New York.

Huemer, M. 2005, *Ethical Intuitionism,* Palgrave Mcmillan.

Hurka, T. 1987, "'Good' and 'Good For'", *Mind: A Quarterly Review of Philosophy,* vol. 96, pp. 71-73.

Hurka, T. 2003, "Desert: Individualistic and Holistic" in *Desert and Justice* Clarendon Press, Oxford.

Hurka, T. 2002, "Capability, Functioning, and Perfectionism" in *Eudaimonia and Well-Being: Ancient and Modern Conceptions* Academic Print & Pub, Edmonton.

Hurka, T. 2001, "The Common Structure of Virtue and Desert", *Ethics: An International Journal of Social, Political, and Legal Philosophy,* vol. 112, no. 1, pp. 6-31.

Hurka, T. 1993, *Perfectionism,* Oxford University Press, New York.

Jackson, F. 1998, *From Metaphysics to Ethics: A Defense of Conceptual Analysis,* Clarendon Press, Oxford.

Joyce, R. 2006, *The Evolution of Morality,* MIT Press, Cambridge.

Joyce, R. 2001, *The Myth of Morality,* Cambridge Univ Pr, Cambridge.

Kagan, S. 1998, *Normative Ethics,* Westview Pr, Boulder.

Kagan, S. 1994, "Me and My Life", *Proceedings of the Aristotelian Society,* vol. 94, pp. 309-324.

Keller, S. forthcoming, "Welfare as Success", *Nous,* .

Keller, S. 2004, "Welfare and the Achievement of Goals", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* vol. 121, no. 1, pp. 27-41.

Kitcher, P. 1999, "Essence and Perfection", *Ethics: An International Journal of Social, Political, and Legal Philosophy,* vol. 110, no. 1, pp. 59-83.

Kornblith, H. 2002, *Knowledge and Its Place in Nature,* Clarendon Press, Oxford.

Korsgaard, C.M. 1986, "Skepticism about Practical Reason", *Journal of Philosophy,* vol. 83, pp. 5-25.

Kraut, R. 2008, "What Is Good and Why: The Ethics of Well-Being", *Ethics: An International Journal of Social, Political, and Legal Philosophy,* vol. 118, no. 3, pp. 557-562.

Kraut, R. 1991, *Aristotle on the Human Good,* Princeton Univ Pr, Princeton.

Kraut, R.H. 1994, "Desire and the Human Good", *Proceedings and Addresses of the American Philosophical Association,* vol. 68, no. 2, pp. 39-54.

Lewis, D. 1998, *Papers in Philosophical Logic,* Cambridge Univ Pr, New York.

Lewis, D. 1983, *Philosophical Papers, V.1,* OXFORD UNIV PR, NY.

Lewis, D.K. 1973, *Counterfactuals,* BLACKWELL, OXFORD,.

Mackie, J.L. 1977, *Ethics: Inventing Right and Wrong,* Penguin Books.

McKerlie, D. 2007a, "Comments on Krister Bykvist: 'Prudence for Changing Selves'", *Utilitas: A Journal of Utilitarian Studies,* vol. 19, no. 1, pp. 47-50.

McKerlie, D. 2007b, "Rational Choice, Changes in Values over Time, and Well-Being", *Utilitas: A Journal of Utilitarian Studies,* vol. 19, no. 1, pp. 51-72.

Mill, J.S. 2001, *Utilitarianism,* Hackett Publishing Company, Indianapolis, IN.

Mill, J.S. 1873, *Autobiography*.

Moore, G.E. 1903 *Principia Ethica,* Cambridge University Press (2nd Ed., 1993 Cambridge University Press, New York)

Nagel, T. 1970, *The Possibility of Altruism,* Clarendon Press, OXFORD,.

Nozick, R. 1974, *Anarchy, State, and Utopia,* BASIC BOOKS, NY.

Parfit, D. 1984, "What Makes a Person's Life Go Best?" in *Reasons and Persons* Oxford University Press, New York, pp. 493.

Parfit, D. & Broome, J. 1997, "Reasons and Motivation", *Aristotelian Society: Supplementary Volume,* vol. Supp, no. 71, pp. 98-146.

Rabinowicz, W. 1989, "Hare on Prudence", *Theoria: A Swedish Journal of Philosophy,* vol. 55, pp. 145-151.

Raibley, 2007, dissertation, UMass Amherst.

Railton, P. 2003, *Facts, Values, and Norms: Essays toward a Morality of Consequence,* Cambridge Univ Pr, Cambridge.

Railton, P. *Alienation, Consequentialism, and the Demands of Morality*.

Rawls, J. 1999, *A Theory of Justice: Revised Edition,* Harvard Univ Pr, Cambridge, MA.

Rosati, C. 1996, "Internalism and the Good for a Person", *Ethics: An International Journal of Social, Political, and Legal Philosophy,* vol. 106, no. 2, pp. 297-326.

Rosen, Gideon. 'Metaphysical Dependence: Grounding and Reduction', *ms*.

Ross, W.D. 1930, *The Right and the Good,* Oxford Univeristy Press.

Sarch, Alexander. 2009, ‘Bealer and the Autonomy of Philosophy’, forthcoming in *Synthese*.

Sayre-McCord, G. 1986, "The Many Moral Realisms", *Southern Journal of Philosophy,* vol. SUPP 24, pp. 1-22.

Scanlon, T.M. 1998, *What We Owe to Each Other,* Harvard University Press.

Schaffer, Jonathan. ‘Monism: The Priority of the Whole’, forthcoming in *The Philosophical Review.*

Setiya, K. 2007, *Reasons without Rationalism,* Princeton University Press, Princeton, NJ.

Sidgwick, H. 1874 (2001), *The Methods of Ethics,* BookSurge Publishing.

Skow, B. "Preferentism and the Paradox of Desire", *manuscript,* http://web.mit.edu/bskow/www/research/paradoxofdesire.pdf

Smith, M. 1996, "The Argument for Internalism: Reply to Miller", *Analysis,* vol. 56, no. 3, pp. 175-184.

Smith, M. 1995, "Internal Reasons", *Philosophy and Phenomenological Research,* vol. 55, no. 1, pp. 109-131.

Smith, M. 1994, *The Moral Problem,* Blackwell, Cambridge.

Smith, M. "The Humean Theory of Motivation", *Mind: A Quarterly Review of Philosophy,* vol. 96, no. 381.

Sobel, D. 1998, "Summer on Welfare", *Dialogue: Canadian Philosophical Review,* vol. 37, no. 3, pp. 571-577.

Sobel, D. 1997, "On the Subjectivity of Welfare", *Ethics: An International Journal of Social, Political, and Legal Philosophy,* vol. 107, no. 3, pp. 501-508.

Stevenson, C.L. 1937, “The Emotive Meaning of Ethical Terms”, in Stevenson, C. L., *Facts and Values*, Yale University Press, 1963.

Sumner, L.W. 1996, *Welfare, Happiness, and Ethics,* Clarendon Press, New York.

Sumner, L.W. 1995, "The Subjectivity of Welfare", *Ethics: An International Journal of Social, Political, and Legal Philosophy,* vol. 105, no. 4, pp. 764-790.

Trogdon, Kelly. 2009, ‘Monism and Intrinsicality’, *Austalasian Journal of Philosophy*, vol. 87, pp. 127-148.

Velleman, J.D. 1998, "Is Motivation Internal to Value?" in *Preferences* de Gruyter, Hawthorne.

Velleman, J.D. 1991, "Well-Being and Time", *Pacific Philosophical Quarterly,* , pp. 48-77.

Williams, B. 1982, "Internal and External Reasons" in *Moral Luck* CAMBRIDGE UNIV PR, NY.

Zimmerman, M. Feb. 7, 2007-last update*, Intrinsic vs. Extrinsic Value* [Homepage of Standford Encyclopedia of Philosophy], [Online]. Available: http://plato.stanford.edu/entries/value-intrinsic-extrinsic/ [2009, April, 25] .