Well-Being and Risk

Gergely Bognár

DOCTORAL DISSERTATION
MAY 2004

THIS DISSERTATION IS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY AT THE DEPARTMENT OF POLITICAL SCIENCE, CENTRAL EUROPEAN UNIVERSITY, BUDAPEST, HUNGARY.

SUPERVISOR:
PROF. JÁNOS KIS
DEPARTMENT OF POLITICAL SCIENCE AND
DEPARTMENT OF PHILOSOPHY
CENTRAL EUROPEAN UNIVERSITY
BUDAPEST, HUNGARY

Abstract

This dissertation develops a theory of well-being. It begins by identifying two deeply held but conflicting intuitions about welfare judgments which, respectively, underlie subjective and objective theories of human welfare. It argues that in order to develop a theory of well-being which is faithful to both of the intuitions, we have to reject the classification based on the distinction between subjective and objective theories.

The dissertation also examines the formal basis of welfare judgments. It argues that utility theory has an indispensable role in a theory of well-being for the making of such judgments, and it explores the relation of issues in utility theory and theories of welfare.

Several chapters are devoted to hedonism. I show that hedonism has both objective and subjective versions, and I raise several objections to it in both of its forms. I also examine the role utility theory can play in an hedonist theory of well-being. Finally, I survey a recent attempt to defend hedonism, and I argue that the commitment of this attempt to subjectivism causes problems.

Then I turn to desire or preference satisfaction theories of well-being. I make the case that such a theory must be formulated in terms of preference rather than desire, and it must be able to identify a subset of a person's preferences which are relevant to that person's well-being. The theory which arises from this discussion is the ideal advisor theory, on which something promotes a person's well-being if and only if the person would prefer that thing were she fully informed and ideally rational. After defending this theory against some recent counterarguments, I develop my own argument against it. My argument points out that the ideal advisor theory faces difficulties if the fully informed and ideally rational person's preferences are preferences over risky prospects, because the theory cannot distinguish between reasonable and unreasonable risks. But I conclude that instead of rejecting the theory, we should revise it. I argue that the revision of the theory that I propose is faithful to both of the intuitions underlying subjective and objective theories of well-being.

Finally, I set out the revision in more detail, and I defend it from some of the objections to which other versions of the ideal advisor theory are vulnerable. I also examine how welfare judgments can be made on this revised theory.

Acknowledgments

This dissertation was written at four places in three countries on two continents. I am grateful to several organizations for grants and scholarships: to the Soros Foundation for a scholarship for undertaking doctoral studies at CEU in Budapest; to the Department of Political Science at CEU for Joint Student-Faculty Research Grants in 2000 and 2002; to the Open Society Institute (Hungary) and the Foreign and Commonwealth Office (UK) for a Chevening Scholarship at Corpus Christi College, University of Oxford, in 2000–01; and to IDP Education Australia and the Department of Education, Training, and Youth Affairs (Australia) for an Australia—Europe Scholarship at the Philosophy Program at the Research School of Social Sciences (RSSS) of the Australian National University, in Canberra in 2002–03. Bibliographical research was also undertaken at the British Library of Political & Economic Science, London School of Economics, London, where I held a European Union Social Science Information Research Facility (EUSSIRF) research grant in 2002. The final revisions were undertaken when I was a research associate at the Department of Philosophy, Flinders University of South Australia, Adelaide, Australia.

There are also many individuals whose help I must acknowledge. In Budapest, I learnt the most from my supervisor, János Kis. I thank him for our many discussions in which he always expressed the problems I was writing about in enormously helpful, clear ways, pointing out inconsistencies, possible objections, and encouraging me to develop my ideas. I also thank Loránd Ambrus-Lakatos, Iván Csaba, Ferenc Huoranszki, Tamás Meszerics, Imre Orthmayr, Balázs Szepesi, and Balázs Váradi for helpful discussions.

In Oxford, I learnt the most from John Broome, my acting supervisor during my year there. Besides János Kis, he taught me the most about how to do philosophy. I thank him for our many discussions in which he meticulously went through my papers, challenging me with objections, forcing me to be more careful with my arguments, and giving useful suggestions for improvement. I am also grateful to James Griffin who gave highly useful comments on an early version of one of my chapters in Oxford.

In Canberra, I had the opportunity to discuss my ideas with many of the staff, students, and visitors of RSSS. I especially thank Michael Smith and Geoffrey Brennan for their insightful comments and advice. I am also grateful to Wlodek Rabinowicz, David Sobel, and Kim Sterelny for useful discussions. I also learnt a lot from a great number of discussions with Campbell Brown. Indeed, the greatest compliment I've ever got as a philosopher was from a fellow student at RSSS, referring to the "Broome-Brown-Bognar" way of doing philosophy.

Finally, I would like to thank Ami, Éva, and Kriszta for administrative help throughout the years of writing this work.

Contents

Chapter 1

Introduction

1.1 Conceptions of Welfare

Welfare, or well-being, is a fundamental concept of moral philosophy.[1] On many ethical theories, actions are evaluated in terms of the goodness of the outcomes they bring about for the individuals who are affected. These theories accept the priority of the good over the right—they give an account of what the good is, and they define the right as the promotion of the good. Many of these theories identify the good with what is *good for* persons. These are welfarist theories. Usually, even those theories that do not accept welfarism incorporate some account of well-being: welfare is one of the values that is to be promoted. Furthermore, even those ethical theories which reject the priority of the good over the right often find it inevitable to appeal to welfare. The concept of well-being is indispensable in ethical theory.

Well-being, however, has not always received the attention it deserves in philosophy. In fact, philosophers drew back from discussing it throughout most of the last century. This was not unrelated to the modern philosophical consensus that there is no one best way of living. Philosophers felt it was not their job to propose theories of what makes the life of a person good for that particular person. Recently, however, the concept of welfare has become again a central topic in ethics. This is indicated by the proliferation of works discussing the merits and demerits of its particular conceptions.[2] Philosophers found that constructing arguments about the concept can shed light to a number of conceptual issues, and they recognized that political philosophy and ethics cannot be done without an account of well-being.

There are different questions that can be raised about the *concept* and *conceptions* of welfare. Questions pertaining to the concept involve the following: Which creatures can welfare be ascribed to? Only to humans, to all sentient beings, or perhaps to lower animals, plans and whole ecologies too? Can it be ascribed to these

[1] Philosophers usually like to say well-being; economists and other social scientists prefer welfare. I will use both, interchangeably.

[2] Here's a sample: Arneson (1999), Bernstein (1998), Bond (1983), Brandt (1979), Brink (1989:217–36), Darwall (2002), Feldman (2002*b*), Finnis (1980:59–99), Goldsworthy (1992), Griffin (1986, 1996, 2000), Kagan (1992), Kraut (1979, 1994), Parfit (1984:493–502), Qizilbash (1998), Railton (1986*a*), Raz (1986:288–320), Scanlon (1993, 1998:108–43), Sumner (1992*a*, 1995, 1996, 2000), and Wiggins (1987), among others. The focus on the concept of welfare is now not unique to analytical philosophy. For an example of its discussion within the continental tradition, see Seel (1997).

welfare subjects only while they are alive, or is it possible to promote their welfare after their death? Do collective entities, like families, nations, or species have their own welfare? In contrast, questions about different conceptions of welfare are concerned with the adequacy of some particular view on what welfare consists in, often bracketing problems of the concept. In what follows, I do the same: I will be interested in particular conceptions of individual human welfare. That is, I only discuss views on what makes the lives of persons go well. I use the term "theory of welfare" to refer only to such conceptions.[3]

The task of a theory of welfare is not to specify the things that make a person's life good for that person. We already have some idea about what these things are. They include nutrition, health, sanitation, shelter, rest and security; literacy, certain intellectual and physical capacities, self-respect, aspiration and enjoyment; autonomy, liberty, personal relations and accomplishment, among others. Rather, the task of such a theory is to tell us *in virtue of what* these things are good for a person. The theory explains why all these things contribute to one's welfare.[4] In the contemporary literature, there are three well-known groups of views on this: *hedonism*, *desire satisfaction accounts*, and *objective theories*.[5]

Hedonism holds that welfare consists in having experiences that result in, or are accompanied by, some valuable mental state, like pleasure or happiness. Desire satisfaction theories hold that what is good for a person is what would best fulfill her desires, or some specified set of her desires. And objective theories hold that certain things are good for persons, irrespective of whether those persons want to realize or avoid these, or the experiences having these would cause. These views tell us what is good for a person as far as her own life is concerned. In this sense, what is good for a person may conflict with what morality requires, or what makes the lives of other persons good for them. Assessment of a person's welfare is thus assessment from the person's own standpoint: how well life is going for the person whose life it is.

[3]It is important to distinguish the question of how well a life goes for the person whose life it is from the question of what makes for a *good life* for that person. This is because a good life for a person is not necessarily a life with the highest attainable "amount" of welfare: it is not necessarily a life that goes well. Besides well-being, there are other values that make for a good life. I am not interested here in the question of what these other values are. How well off a person is can be assessed independently of how good, overall, her life is.

[4]When I say "things," I mean to include objects, experiences, states of affairs, etc. I will be more precise in later chapters. For instance, for hedonists only experiences are the sort of things that can promote welfare; other kinds of theories are more inclusive. The examples I gave are on the list given by Qizilbash (1998:67). His list is an extension of Griffin's (1996) list of "prudential values."

[5]This triple distinction comes from Dworkin (1981), Parfit (1984), and Griffin (1986). Hedonist theories are sometimes also called *mental state accounts*; in desire satisfaction theories, "desire" is often replaced by preference; and objective theories are sometimes called *objective list* or *substantive goods* theories.

This classification of theories of welfare is quite common in the literature. There is, however, another way of classifying such theories which is more useful for my purposes. According to this taxonomy, all theories of welfare can be grouped into *subjective* and *objective* accounts. Subjective theories of welfare require some connection between the person's attitudes and the sources of her welfare, while objective theories do not. Thus, on a subjective theory, something is good for you in virtue of your having some sort of pro-attitude towards it. On an objective theory, something is good for you in virtue of something else. The advantage of the subjectivity-objectivity divide is that it seems to give an exhaustive and mutually exclusive classification of theories of welfare.[6]

In terms of this classification, the various versions of desire satisfaction theories are all examples of subjective accounts of welfare. They are subjective because what is good for you is determined by what you desire or would desire in some appropriate conditions, and desiring is a pro-attitude. The case of hedonism is a bit trickier: whether it belongs to the subjective or objective group is determined by how one construes the mental state that is valuable on an hedonist theory. Thus, if an hedonist holds that welfare consists in pleasure, she may think that pleasure is a sort of pro-attitude; then, on this classification, her theory is subjective. But she may think that pleasure is merely a feeling, in which case her theory is in the objective group: she will think that this feeling is good for a person irrespective of the attitude the person has towards the feeling. Thus, hedonism turns out to have both subjective and objective versions.[7] Finally, objective theories are just that—objective. For something to promote the well-being of a person, they do not require that the person has any sort of pro-attitude towards that thing.

Nevertheless, I will define the distinction between subjective and objective theories of welfare in a slightly different, albeit similar, way. Instead of drawing the distinction in terms of pro-attitudes, I will draw it in terms of *preference*. Thus, on a subjective view, what is good for a person is determined by what the person prefers in the appropriate conditions specified by the theory; on an objective view, what is good for a person is independent of that person's preferences.

I have three reasons for drawing the distinction in terms of preference. First, perhaps the most influential group of theories, and one I will discuss at length, is the desire satisfaction theory of welfare. Many modern versions of this kind of view are put in terms of preference: they hold that what is good for a person is not what a person desires (or would desire in appropriate conditions), but what the

[6]For this way of drawing the distinction, see Sumner (1995, 1996:26–41). Dworkin (2000:216–8) draws it in terms of endorsement.

[7]The difference between these two kinds of hedonism will be explained in more detail in Section 4.2. In the literature, hedonism is almost always presented as a subjective theory. The one exception I am aware of is Scanlon (1993:189), who treats it as an objective (or, as he says, substantive goods) theory.

person prefers (or would prefer in appropriate conditions). Even though desire and preference are often treated as if they were the same thing, they are not. It is more precise to characterize these views in terms of preference rather than desire.

Second, a theory of welfare must be put forward in comparative terms. It must tell us not only in virtue of what some things are good for us, but also in virtue of what some things are better for us than other things. Otherwise we could hardly ever answer the question of what to do to promote our well-being. This is another reason why a desire satisfaction theory of welfare is couched in unsuitable terms, and it must be reformulated as a preference satisfaction theory: preference is a comparative notion, desire is not. Classical hedonists, for instance, were aware of the comparative nature of welfare judgments: that's why they said welfare consists in the net balance of pleasure and pain.

Third, I draw the distinction in terms of preference to sidestep the differences between subjective theories of welfare employing different pro-attitudes. This way, I do not have to address these differences, and I can argue for or against several versions of a theory at once. Since preferring is a disposition to choose, no matter in terms of what pro-attitude you construct your theory of welfare, that pro-attitude will operate through preferences: it is "manifested" in the person's preferences. All we have to require is that the person prefers for the appropriate *reason*: for example, if welfare consists in pleasure, the reason for which a person concerned with promoting her welfare prefers one alternative to another is that it brings her more pleasure.

Therefore, on my way of constructing the subjectivity-objectivity divide, subjective theories are those which maintain that what the person prefers in appropriate conditions determines what is good for her. Objectivists deny this. They hold that welfare is independent of preference. According to subjectivists, preference determines welfare; according to objectivists, welfare determines preference—in the sense that what is good for you specifies what you ought to prefer, rather than what you prefer determining what is good for you.

1.2 Two Powerful Intuitions

Subjective theories of well-being are intuitively attractive. Indeed, it has been argued that no objective theory can be adequate. According to L. Wayne Sumner,

> no theory about the nature of welfare can be faithful to our ordinary concept unless it preserves its subject-relative or perspectival character. ... Welfare is subject-relative because it is subjective. (1995:774 and 1996:42–3)

Nevertheless, there is a similarly compelling argument to the effect that our preferences cannot be sufficient for specifying what is good for us. Indeed, it is a depressing fact about our lives that we often prefer what does not turn out

to promote our welfare. In some cases, our preferences are mistaken: we lack information or reason erroneously when forming them. In other cases, even if we have all the relevant information and reason appropriately, we may still fail to prefer what is good for us. There may be things that contribute to our well-being, regardless of whether we prefer them or not.

This last claim is more controversial. The subjectivist holds that even if preferring it is not a sufficient condition for something to be good for us, it must be a necessary one. She concedes that sometimes people can be mistaken about what is good for them, but, she holds, were their mistakes corrected, they would prefer what in fact promotes their welfare. Objectivists, however, reject this claim. They point to cases where individuals engage in unworthy pursuits, but it seems we cannot show they would give up these pursuits even if they were in more favorable conditions for forming their preferences—if, for instance, they possessed all the information relevant to their situation and made no mistakes of reasoning.

Both parties express a strong intuition about the way the sources of our welfare are related to us. Roughly put, they are these:

The subjectivist intuition. For some thing, x, to promote your welfare, it is necessary that you prefer (or would prefer in some specified conditions) x.

The objectivist intuition. For some thing, x, to promote your welfare, it is necessary that this thing, x, is worthwhile independently of your preferences.

Currently, there seems to be an impasse between subjectivism and objectivism about well-being in philosophy. Both kinds of view incorporate a plausible intuition which we are reluctant to give up. Objectivism had better be true, otherwise there is no basis to our judgments about what promotes welfare other than preferences. Subjectivism had better be true, because otherwise we cannot anchor what is good for a person in that person's preferences. A theory of well-being that does not include both factuality and endorsement in some form seems implausible.

Both intuitions express deeply held convictions about the nature of human welfare. Hence, instead of rejecting one of the intuitions—and opting for either a subjective or an objective theory of welfare, as it seems to happen in the literature most of the time—I think we should stick to both. But what can we do then?

One thing we can do is to argue that classifying theories of welfare in terms of whether they are objective or subjective may be exhaustive, but not exclusive: all theories are either or both. Then we can work out a theory that is faithful to both intuitions. Thus, perhaps one way out of the deadlock is to develop what I call an *hybrid* account. Such a view would hold that

> nothing can intrinsically enhance an individual's well-being unless it is both truly worthwhile and also affirmed or endorsed by that very individual. (Arneson, 1999:114)

On an hybrid view, for some thing, x, to promote your welfare it is *both necessary* that you prefer x and that x is worthwhile independently of your preferences, but neither is sufficient in itself. This way, both intuitions are incorporated.

This may strike one as an attractive view, but it is not very plausible. In order to see why, suppose I am concerned with promoting you welfare, and I can give you some thing, x. Suppose also that there is a way to determine one of the necessary conditions about the purported source, x, of your welfare—that it is "truly worthwhile." Now all I have to do is to determine whether you prefer x to make sure that it would indeed promote your welfare. But what sort of preference should I look for? Is it necessary that you prefer x *ex ante*, before I give it to you, or can your preference be *ex post*? If it can be *ex post*, then often it is not true before giving you x that it is the sort of thing that promotes your welfare. But this implication is implausible. On the other hand, if you must already prefer x before getting it, then many things that you would prefer only *ex post* are ruled out. For instance, you might never have thought about them; or you might not have had enough information to form a preference; or you might have tried to form a preference, but they are the sort of thing whose contribution to well-being can be determined only retrospectively. Thus, the requirement that you prefer x *ex ante* is implausible too.

More generally, if x is "objectively" good for you by stipulation but you do not prefer it in the appropriate conditions, why would your welfare not be promoted by getting it? After all, it is truly worthwhile! Conversely, if you prefer something in the appropriate conditions, even though it is not truly worthwhile, why does it *necessarily* not promote your welfare? Why couldn't it be good for you merely in virtue of your preferring it?

Therefore, even though hybrid views appear to be able to incorporate both the subjectivist and the objectivist intuitions, they do it in an implausible way. Where can we go from here?

Another proposal may be what I call a *mixed* theory of welfare. Such a theory retains the idea that the subjectivist and objectivist intuitions provide conditions for some things to promote a person's well-being, but it denies that they provide *necessary* conditions. On such views, both intuitions provide *sufficient* conditions. The idea is familiar: in order for some purported source of your well-being to be good for you, it is sufficient that you prefer it in the appropriate conditions. In order for some other purported source of your well-being to be good for you, some other basis is sufficient rather than your preferences.

I think many philosophers are attracted to such a view, although few have explicitly embraced it. This is because having put this view on the table, now one has to explain when preferring is a sufficient condition, and when it is not a necessary condition, for something to promote a person's well-being. This, however, is bound to be controversial. Suppose we hold that health, achievement, and friendship are good for us independently of our preferences, but insofar as they are not

relevant in our choices, what is good for us is what we prefer. One can always ask why only these are good for us, and not, for example, wealth or knowledge. And one can always ask why health, achievement, or friendship are good for a person even if that person does not prefer them. A mixed theory appears to be merely an objective view in a new disguise, since it seems to arbitrarily specify certain things which are good for the person independently of the person's preferences.

The problem is that on all of these theories the judgments involved in determining what promotes welfare are *substantive judgments*—by which I mean that they are judgments about the *content* of preferences. On both the hybrid and the mixed views, one cannot avoid specifying certain goods that you ought to prefer when you are concerned with promoting your welfare. Such judgments are patently controversial, and any theory that allows for them is at least partly objective.

In order to avoid having to make such substantive judgments, subjectivists usually subscribe to a version of the preference satisfaction theory which holds that the preferences whose satisfaction promotes the person's welfare are formed in ideal conditions for preferring. This idea is also familiar, and very popular, in philosophy. I call theories incorporating this idea variants of the *idealization theory*. The advantage of such views is claimed to be that they make it possible to account for how persons can be mistaken in their judgments of what promotes their welfare without appealing to substantive judgments. On the best-known versions of the idealization theory, the ideal conditions are *epistemic* and *cognitive*: they require that the person is adequately informed and appropriately rational. From the many names for such views, I chose to call them versions of the *ideal advisor theory*. The idea is that metaphorically speaking, the adequately informed and appropriately rational preferences of the person are the preferences of her "ideal advisor." On such a view, in order to determine what promotes the well-being of a person, it is sufficient to ask what she would, if she was, counterfactually, adequately informed and appropriately rational, prefer herself to prefer.

There are many possible versions of the idealization theory besides the ideal advisor theory, depending on how they construe the ideal conditions for preferring. Moreover, there are many sub-versions of the ideal advisor theory, specified by how they interpret the epistemic and cognitive conditions. The details are obviously more complicated, but all idealization theories are still subjective, since they connect welfare to what the person would prefer in ideal conditions. What is good for the person is what she would ideally prefer herself to prefer, and determining what the person would ideally prefer herself to prefer does not require substantive judgments of the sort discussed above. On the ideal advisor theory, all that is needed for establishing the ideal preferences are rationality and information. It is claimed that this theory goes some way to meet the objectivist intuition, because at least it incorporates ways for criticizing the actual preferences of the person. But these ways involve only *formal* judgments, in the sense that only the basis and re-

lations of ideal preferences can be scrutinized, and no preference is to be criticized on the basis of its content.

But what if those ideal preferences cannot be established without something else besides rationality and information? What if there must be further constraints on ideal preferences? In that case, it is possible that some ideal preferences turn out to have a substantive basis—for the determination of what an adequately informed and appropriately rational ideal advisor would prefer her "counterpart" to prefer has to appeal to something else besides the norms of rationality and facts. It may have to appeal to the content of preferences.

My ultimate aim in this work is to explore this possibility. If I can show that the most plausible version of the ideal advisor theory has to appeal to the content of at least some of the preferences, we may have found a theory that is able to incorporate both the subjectivist and the objectivist intuitions. This seems to put one of the legs of the theory into the objectivist camp. If so, we have to choose between rejecting the theory or revising it. I see no reason to reject the theory other than an unwarranted aversion to the objectivist intuition on the part of subjectivists. According to my revised ideal advisor theory, welfare consists in the satisfaction of ideally rational and fully informed preferences with some constraints on the contents of at least some of those preferences. The theory is faithful to the subjectivist intuition, since it connects well-being to preference; and it is faithful to the objectivist intuition, since it also appeals to judgments about the content of preferences.

If it turns out possible to develop such a revised version of the ideal advisor theory, what will distinguish it from what I called the mixed view of well-being? After all, this theory also involves substantive judgments—judgments about the content of some preferences—and such judgments are controversial. The difference between the revised theory and a mixed view lies in the fact that the latter appeals to certain *goods* which promote the well-being of a person without the need for that person to prefer them; in contrast, my revised theory appeals to judgments about the reasonableness of *risks* which influence what preferences the person forms over goods. Thus, the two views differ in their construction of the substantive judgments.

Current discussions of well-being tend to center around the issue of which of the three major views of welfare is the best one to accept. I think this is a mistake. The question that ought to be asked instead is the relation of preference to well-being. Defenders of hedonism have part of the truth: many things that are good for us are good for us because they bring pleasure or happiness. But others are not. Also, many things are good for us because we desire them; but some things we do not desire are also good for us. In general, no subjective view can be entirely correct. But neither can any objective view contain the whole truth so long as it fails to give an account of why preferences should not matter. Suppose there is a list of the things that are ingredients of well-being. The list may even be relativized

to specific individuals. If the list is such that it is possible for any individual to fail to affirm all of its items, the account is implausible. There is no *prima facie* reason to think that this will not be the case with any such list.

One might think—and some philosophers do indeed seem to think this—that in order to incorporate the subjectivist and the objectivist intuitions, we should simply develop an objective theory which includes some "subjective" element. The idea is that our list of the goods which promote well-being could include such items as happiness or preference satisfaction. But I fail to see why such an enumeration of goods would be a *theory*. A theory has to be able to explain in virtue of what the goods it proposes promote welfare. A list of goods is not a theory. In addition, if the list includes both goods and attitudes towards goods, it is hard to see what the theory's reply may be to the question of in virtue of what items on the list are good for people. Therefore, I choose to proceed by discussing subjective theories and searching for a version that can incorporate the objectivist intuition, rather than discussing objective theories and searching for a version that can incorporate the subjectivist intuition. The latter enterprise seems to me rather unpromising.

In any case, the objectivity-subjectivity distinction is ultimately misleading, and the arguments based on it distract from the real issue. It is trivially true that our actual preferences may fail to correspond to what is good for us. But even if our preferences could be filtered some way to avoid mistakes and shortcomings, it is still possible that we would not prefer what makes us better off. So subjectivism fails, and objectivism fares no better. But we do not have to take sides. Sticking to the distinction has sidetracked philosophers. We should discard it, without discarding the intuitions that fuel different theories of well-being. What we should ask instead is this: what characterizes the preferences which matter for well-being, and when do we have to appeal, in addition to these characteristics, to their content?

1.3 Overview

Philosophers don't have a privilege of thinking about conceptions of welfare. Many of the social sciences are also concerned with what welfare consists in and how to promote it individually and collectively. The obvious examples are welfare economics, social choice theory, and some branches of political science. Thus, I hope that some of what I say may be interesting to economists and political scientists as well. Once we realize that welfare is a significant concept in these disciplines too, there is actually another advantage of cashing out theories of welfare in terms of preference. Since these social sciences are also concerned with preferences, and they have built formal methods to represent preferences—or welfare, insofar as preferences are relevant to welfare—some of their methods and results can be fruitfully employed in the context of a philosophical inquiry.

Thus, Chapter 2 is concerned with the history and present state of *utility theory*,

in the context of which economists have addressed problems of welfare. I discuss the relation of utility theory to conceptions of welfare, and explain why a theory of welfare needs utility theory. I also briefly present the history of utility theory in terms of the development of the interpretation of the concept of utility. I give a brief overview of modern utility theory, and close the chapter with some notes on the normative significance of preference and utility to ethics.

The subject of the following two chapters is *hedonism*. There are two reasons for devoting an extended discussion to this theory of welfare. The first is that hedonism has played a very important role in the history of moral philosophy and welfare economics, and it has shaped the way these disciplines are today. It has been especially influential in the histories of both the idea of welfare and utility theory. The second reason is that even though it is widely rejected today in both disciplines, echos of its influence linger on; hedonist ideas more or less implicitly still surface today in theories of welfare and interpretations of utility theory. One reason for this is doubtless intellectual carelessness; another is that hedonism as a theory of welfare is surprisingly difficult to defeat.

Therefore, Chapter 3 looks at the present state of the debate surrounding hedonism as a theory of welfare. It argues that Robert Nozick's *experience machine*, the well-known thought experiment taken by so many philosophers to provide a conclusive argument against hedonism, can be easily sidestepped by defenders of the theory. But it will also show how an analysis of the thought experiment raises doubts about the plausibility of any version of hedonism nevertheless. There are some features of an hedonist theory that need to be defended, and it is hard to see how hedonists could construct their defense. Sadly, this is not a conclusive counterargument: perhaps they will be able to make a comeback. In any case, until they do, we have no reason to accept hedonism as a theory of welfare.

On the other hand, showing that hedonism is incompatible with modern utility theory—and the way the concept of utility is employed in the modern social sciences and the more formal approaches to ethics—is a much easier task. I undertake it in Chapter 4. My discussion will take place in the context of utilitarianism. This should not be understood as embracing a utilitarian moral theory. Rather, utilitarianism provides a useful framework here in which I can address certain issues of theories of welfare and utility theory. Thus, I argue in this chapter that hedonist utilitarianism, on both of the objective and the subjective versions of hedonism I distinguish, is incompatible with modern utility theory—that is, the way utility has been constructed in welfare economics and increasingly in ethics for more than half a century now. I also present a couple of attempts to give utility an hedonist interpretation, and show why they are misconceived.

Chapter 5 reviews the most interesting recent theory of welfare in my opinion, L. Wayne Sumner's theory. Sumner defends a thoroughgoingly subjective conception of welfare. I will present some points in his theory where his commitment to

subjectivism causes problems. But these problems raise questions which are relevant beyond the internal consistency of his theory. Thus, the lessons to be learnt from the failure of Sumner's project have much wider implications: they highlight the problematic points of subjectivism about welfare in general.

The task of the next chapter is to pave the way to the most plausible and most influential kind of theory of welfare: the versions of idealization theory that hold that welfare consists in the satisfaction of the preferences the person would have were she adequately informed and appropriately rational—the ideal advisor theory. Chapter 6 surveys preference satisfaction theories of welfare in general and presents problems for some of the versions which have been proposed in the literature—specifically problems for those theories which do not employ idealization in determining the set of preferences whose satisfaction constitutes well-being. These problems provide an indirect argument in favor of moving to some version of the idealization theory. In addition, I attempt to refute more specific arguments against the conjecture that if one wants to be a subjectivist about welfare, then one has to accept an ideal advisor theory. My overall aim in this chapter is to narrow the set of plausible preference satisfaction theories to the ideal advisor view.

Chapter 7, then, turns to the ideal advisor theory itself. First, I present some of its well-known versions. Then I try and defend the theory from the various objections which have been lately raised against it. It is interesting to note that even though this kind of theory continues to be very popular, it has received various criticisms recently—objections which purport to show that the very idea of epistemic and cognitive idealization is conceptually incoherent. I defend the theory because it seems to me that these objections are either unfair, or their proponents help themselves to an awful lot of unsubstantiated background assumptions.

Having discarded the criticisms, in Chapter 8 I come to the main argument of this work—my own objection to the ideal advisor theory. First, I will further narrow it to the one version I take to be the most plausible—which I label with the abbreviation "IRP." On this version, the ideal advisor is *fully* informed and *ideally* rational. After giving an interpretation of this characterization, I present my objection. In brief outline, it is this. A theory that defines welfare in terms of the satisfaction of the preferences a person would have were she fully informed and ideally rational is either underdetermined—in the sense of being unable to specify what a fully informed and ideally rational person would prefer—or it needs to involve *substantive judgments* in the sense I introduced the notion on page 7: judgments appealing to the content of preferences. But I do not think this is a reason to discard the theory. Rather, it is an opportunity to revise it, hence to develop a theory of welfare which combines the subjectivist and objectivist intuitions.

Chapter 9 begins by examining an objection against the ideal advisor theory: roughly, that it does not respect the autonomy of persons. This objection is especially powerful, since one of the *prima facie* considerations in favor of a subjective

theory of welfare is that it is able to make sense of the idea that autonomy is constitutive of well-being. I show how the ideal advisor theory can meet this objection by an appeal to a principle of preference autonomy interpreted in terms of ideally rational and fully informed preferences. It turns out, however, that this defense has the implication that the promotion of well-being can lead to widespread paternalism. I suggest, however, that if the ideal advisor theory is revised in the way I propose, the revised version can avoid this implication.

Chapter 10 sets out in detail my revision of the ideal advisor theory and contrasts it with the version proposed by John C. Harsanyi. Throughout the previous chapters, Harsanyi's work will have served as an inspiration and starting point for many of my arguments. Here I explain why I think my revised version of the theory is preferable to his. I also discuss the role of utility theory in the context of the theory which emerged from earlier chapters. Finally, I briefly conclude and present some questions that merit further research.

Chapter 2

A Brief History of Utility

2.1 Welfare Judgments

Consider the following quote from the philosopher and political scientist Russell Hardin:

> The most articulated and sophisticated effort to deal with the theory of human welfare and even of value theory in general has been the development of utility theory in economics from roughly the time of Bentham to the present. Any effort to understand human welfare should take account of the issues in that utility theory—or perhaps one should speak of many utility theories. (1988:169–70)

The notion of utility has been historically closely related to conceptions of welfare. This clearly implies that any inquiry into the various conceptions of welfare has to explore this relation. But contrast now what Milton Friedman and Leonard Savage, two economists from the founding movement of modern utility theory, have to say about the relation of utility and welfare:

> The ethical precept that society "should" promote the "welfare" of individuals is meaningless until "welfare" is given content. Any identification of "welfare" attained by individuals with "utility" ... is itself an ethical precept, to be justified on ethical grounds ... (1952:473)
>
> [and] it is entirely unnecessary to identify the quantity that individuals are to be interpreted as maximizing with a quantity that should be given special importance in public policy. (1948:283; also 1952:473)

Both quotes express important ideas. The first stresses that conceptions of welfare and theories of utility are in close theoretical proximity. The second stresses that nevertheless their relation is not unproblematical at all.

In order to clarify the connection between conceptions of welfare and utility theory, I begin by asking a more general question: Why do we need a theory of welfare?

Welfare is a fundamental concept of moral philosophy, and, as the name implies, of welfare economics as well. Many popular ethical views today are *consequentialist*: they accept the view that outcomes are the only ultimate standard in deciding what we ought to do. These views might, in addition, accept *welfarism*: the view that only well-being is ultimately valuable. Of course, many philosophers reject consequentialism or welfarism or both. But it doesn't follow that they do

not need an account of well-being. It does not follow because an ethical view that holds that consequences *never* matter, and how good these consequences are for the people affected *never* matters, looks like a crazy view. Whatever one thinks of the merits of consequentialism and welfarism, it cannot be denied that at least sometimes we ought to act to bring about the best consequences, and at least sometimes those consequences ought to be evaluated in terms of how good they are for the people affected—according to the extent they promote their welfare. That is to say, at least sometimes what we have the most reason to do is to promote well-being. Whatever one's ethical view otherwise is, one must have a view on what well-being is for such occasions.

This is equally true of welfare economics and the other normative branches of the social sciences. These disciplines are concerned with how to promote the welfare of the individuals in a society, and designing institutions which most efficiently promote welfare or achieve other desirable goals under conditions of scarcity of resources and other constraints. In order to do this, it is necessary that they say something about what welfare is.

If it is true that well-being is something that is to be "promoted," then we need to know what it is to promote it—that is, a conception of welfare must be *operationalizable*. A theory of welfare must allow for the making of judgments which are necessary to be able to evaluate states of affairs and outcomes in terms of how good they are for the people who are affected. In sum, a theory of welfare ought both to be able to tell us in virtue of what something is good for a person, and enable us to make at least some kinds of *welfare judgments*.

There are several kinds of judgments which might be necessary for a theory of well-being to be operational. One sort of judgment is concerned with how well off people are. For instance, if you are an hedonist, how well off people are is determined, according to your theory, by how much pleasure, happiness, or some other valuable mental state they have. If you are a desire or preference satisfaction theorist, you believe how well off people are is determined by the extent some or all of their desires or preferences are fulfilled. And if you are an objective theorist, you look at how many of the objective goods you believe contribute to well-being people actually possess. On any account of well-being, you must be able to give some sort of a *representation* for how well off people are. Hopefully, your representation can take a mathematical form: you should be able to assign real numbers to levels of welfare. Of course, it is a question for further inquiry how precise these representations can be.

Other sorts of judgments are concerned with *comparisons*. Suppose we know that your level of well-being can be represented with the number 5. But to know that you have 5 "units of welfare" is not knowing much: we also need to know how those 5 units compare to the units of other people, or to your level of well-being at other times. Thus, a theory of welfare must tell us whether judgments

of *intrapersonal* comparisons of welfare are possible, and if they are, how they can be carried out. Similarly, it must tell us whether judgments of *interpersonal* comparisons are possible, and how they can be carried out if they are.

Furthermore, if any of these judgments is possible, then what are the exact forms they can take? Can we make intrapersonal or interpersonal comparisons of welfare *levels*? Or can we perhaps make these judgments more precise than "A is better off than B," or "A and B are equally well off"? Can we compare not only levels, but also *units* of welfare, such that we can determine how much better off one person is than another (or how much better off one person is at one time than at another time)? In addition, can we perhaps make comparisons of well-being with respect to some determined zero level? Whether any of these judgments is possible is settled by the mathematical form we are able to give to our representation of well-being.[1]

Consequently, a complete theory of well-being accomplishes two tasks: it specifies in virtue of what goods promote welfare, and it also specifies how to represent in an operational way the extent to which those goods promote welfare. This latter part of the theory determines what sort of judgments, and how precise judgments, we can make about how well off people are, and how much something may contribute to their welfare.[2]

Historically, the latter part of the theory has been dealt with in the framework of *utility theory*. When classical hedonists, like Bentham, used the term "utility," they meant the *quantity* that represents welfare. Thus, classical utilitarians proposed a complete theory of welfare: they held that well-being consists in pleasure, and utility is that which represents pleasure in an operationalizable way. In the framework of their utility theory, they could give an account of the problems of welfare judgments—problems of measurement and comparability. This is what the quote from Hardin with which I opened this chapter calls our attention to: no theory of welfare is complete unless it explores the problems of welfare judgments in the context of utility theory.

This point might go unnoticed, for utility is an ambiguous term. Very often, it is used in the sense of "value." This use is legitimate if and only if it literally means the *value of* something, for instance, welfare or preference satisfaction. In

[1] For these distinctions, see List (2003).

[2] A further question can be raised about the *temporal unit* of welfare measurement. Suppose we want to establish how well off a person is, or how well off she is in comparison to other persons. Do we then assign a value to how well off she is for her whole life, for this very moment, or something in between? This is an important question which might lead to controversial problems in metaphysics; for instance, to the problem whether persons persist through time, or whether they are merely a collection of persons at different time slices. (On this problem, see Parfit (1984).) I will bracket these problems entirely; for the sake of simplicity, I assume that levels of welfare can be assigned to any moment in a person's life. The exploration of these problems must await another occasion.

its legitimate sense, utility is a mathematical concept, as utility theory is first and foremost a mathematical theory of measurement.[3] Thus, utility theory is a theory for representing or measuring something: it explores the mathematical structure of representations by studying the properties of *utility functions*.

The relation of theories of well-being to utility theory is therefore this. Any complete theory of well-being must tell us how to make welfare judgments. Formally, these judgments can be made by representing welfare by utility functions, and the mathematical properties of these functions determine what sort of judgments are possible. Perhaps well-being is a value that cannot be represented by utility functions at all; perhaps it can be represented by a *class* of utility functions.[4] Or perhaps, for each person, there is one unique utility function that represents that person's welfare. Which of these is the case depends on our theory of what well-being consists in; and that theory is a part of ethics. This is what Friedman and Savage emphasize in the quotation on page 13. In sum, whereas the part of a theory of welfare which tells us what well-being consists in determines what sort of representation for well-being is possible, the part of the theory that addresses the problems of representation determines what kinds of welfare judgments can be made.

2.2 Utility from Pleasure to Preference

Ethics and economics are both concerned with welfare. Historically, these disciplines also developed hand in hand. Adam Smith, Jeremy Bentham, or John Stuart Mill contributed to both disciplines, and it is only in more modern times, with the specialization of the sciences, that ethics and economics grew apart.

Before they parted ways, both disciplines had been markedly influenced by early utilitarianism. It is thus not surprising that the development of economists' view of welfare parallels the development of utility theory, and developments in philosophy in general. Classical utilitarians were *hedonists*: they held that welfare consists in pleasure and the absence of pain. The founders of modern economics—Jevons and Marshall—overtook and, to some extent, distorted the conception of welfare of their forerunners, the utilitarian welfare economists. They adopted the view that welfare consists in pleasure, and equated pleasure with utility—the key term of their economic thinking, based on marginal utility analysis.

[3]There is a similar equivocation in the notion of preference. On the one hand, preference is a mathematical concept: the name of a *relation*. But it is also used in the sense of *disposition to choose*. The conflation occurs since preference (in the sense of a relation) is used to represent preference (in the sense of a disposition).

[4]A class of utility functions is defined by the transformations under which the informational content of the functions is invariant. These functions are "unique up to" some form of mathematical transformation.

There is some controversy in the literature over what classical utilitarians on the one hand, and Jevons and Marshall on the other, meant by "utility." For instance, John Broome (1991*a*:19–21) argues that by utility classical utilitarians meant *usefulness*: the tendency of an object to produce pleasure or benefit. This is the sense in which Jevons and Marshall overtook the concept, and it shifted to designate benefit or pleasure only later. I cannot undertake to untangle the interpretative issue here. What is important is that the shift in meaning occurred, not whether it did with Jevons and Marshall or sometime after them.[5] The shift in meaning is manifested by the many proposals to substitute utility with a less ambiguous term. For instance, Irving Fisher (1918) suggested "wantability" instead of utility. Meanwhile, philosophers have grown more and more critical of the conception of welfare behind classical utilitarianism, but their contributions did not penetrate economics.[6]

Why is the notion of utility so crucial for an inquiry into conceptions of welfare? Utility is a technical concept, an index that tells us something about a person's preferences. But it is natural to think, and it has been assumed, that it tells us more. It is natural to think, and it has been assumed, that it tells us something about a person's well-being.

You need such a formal index because no tangible commodity (e.g., money), social relation, or personal characteristic (knowledge, virtue, etc.) seems to be a good indicator of well-being. Economists have long recognized that money or income cannot fully represent welfare. They noticed that money and income are valued differently at different levels of wealth; that is, they have diminishing marginal utility for people.[7] In this context, two questions became focal in utility theory: whether utility can be measured and whether it can be compared across persons. Bentham (1789) introduced interpersonal comparisons by assumption and worked out an elaborate system of factors that determine one's level of utility (pleasure)—intensity, duration, certainty, etc. Economists, implicitly, followed

[5]Nevertheless, for the record, Jevons does devote Chapter 2 of his 1871 to pleasures and pains, and Chapter 3 to utility, where he says: utility "is a convenient name for the aggregate of the favourable balance of feeling produced—the sum of pleasure created and the pain prevented." In a fascinating article, Ross M. Robertson (1951) traces Jevons's concept of utility to Bentham and to an obscure economist, Richard Jennings, who based his theory of value on psychology and *physiology*. (See also Jevons's quotes from Jennings in Chapter 3 of his 1871.) Marshall uses all the terms "happiness," "enjoyment," and "pleasure" for well-being in Chapter 6 of Book 3 ("Value and Utility") of his 1890. He also speaks interchangeably of the utility of a commodity and the utility of the consumer.

[6]An early example is Harrod (1936:145–7).

[7]For an early conceptualization of the phenomenon of diminishing marginal utility, see the solution of Bernoulli (1738) to the so-called St. Petersburg paradox. It was, however, Jevons, Menger, and Walras—independently of one another—who introduced the concept of marginal utility into economics.

in his footsteps in assuming that utility represents some sort of *psychological state*. Jevons, Menger, and Walras had qualms about the measurability of utility and rejected the possibility of interpersonal comparisons. But they continued to assume that the problems stem from the shortcomings of the methods at the disposal of the social sciences of their day, and not from an unwarranted substantive assumption.

What has come to be known as the "measurability of utility controversy" was the debate between those economists who believed that there is a psychological state—pleasure or satisfaction—which, in principle, can be represented by *one unique* utility function for each individual. They hoped that this unique utility function could be unambiguously determined, perhaps by introspection, perhaps by some other method. If such a unique utility function could be determined for each consumer, economics could be made a precise, quantitative science.[8] This position is sometimes called *cardinalism*. The version of this view which holds that the unique ("cardinal") utility functions can be established by introspection is sometimes called *introspective cardinalism*.

Pareto (1927) was one of the first economists to doubt that such utility functions were possible to set up: he was celebrated for having shown the immeasurability of utility. Although Pareto still defines *ophélimité* in terms of pleasure,[9] his analysis is in terms of tastes—as we would say today, preferences. He started from indifference curves as given, and built up his economic theory based on them—as opposed to, for instance, Edgeworth, who assumed the existence of utility and derived indifference curves from it (1927:176). Pareto's novel idea was that the indices assigned to indifference curves represent the ordering of the person's preferences, but we can never know which indices truly measure the person's welfare. That is, many utility functions represent preferences, but we do not know which one truly or uniquely represents the person's pleasure. As opposed to earlier attempts, which hoped to get access to the true representation of welfare, the Paretian turn was showing that we do not need such access in order to carry out economic investigations.

Nevertheless, there were still attempts to come up with an empirical method for establishing unique cardinal utility functions. Wicksteed, Wicksell, Edgeworth, and Pigou continued to assume that introspective cardinal utility was empirically measurable. But this gradually became less and less plausible.[10]

The Jevons-Menger-Walras-Marshall tradition more or less explicitly assumed that utility functions represent some sort of psychological state of the decision maker—some sort of feeling towards, attitude to, or pleasure derived from, the

[8]For the debate on the ambitions of welfare economics in the early part of the 20th century, see Hicks (1939) and Little (1950:84–128).

[9]"Ophélimité" is his term for utility. For instance, §32: "For an individual, the *ophélimité* of a certain quantity of a thing … is the pleasure which this quantity affords him" (1927:168).

[10]For treatments on the development of the formal aspects of utility theory, see Little (1950:6–37), Stigler (1950*a*,*b*), and Ellingsen (1994).

outcomes of alternatives. If some of these sorts of psychological data could be accessed, one could build a *cardinal* utility function, which might be interpreted as reflecting the intensity of one's attitude towards alternatives. As a further step, many economists presumed that utility was interpersonally comparable. Modern economics has rejected this latter assumption since the so-called "ordinal revolution" of the 1930s.[11]

Classical utility theory has been supplanted by modern utility theory, starting from the work of John von Neumann and Oscar Morgenstern (1944). Different versions of this theory have been developed, and these are prevalent in economics and philosophy today. Most of these share the basic idea behind the Neumann-Morgenstern version. In what follows, I give a brief, nontechnical overview of this basic idea.[12]

The Neumann-Morgenstern utility theory (or NM utility theory, as it is commonly abbreviated) has two components: *axiomatic utility theory* for decisions under certainty, and *axiomatic expected utility theory* for decisions under risk and uncertainty. In modern utility theory, "utility" stands for a representation of a person's preferences: it is the value the function representing the person's preference ordering can take. Thus, a "utility function" represents the person's preference ordering.

In axiomatic utility theory, utility functions are *ordinal*; in axiomatic expected utility theory, these functions are *cardinal* (I will further clarify these terms below). On both versions, you only need to observe choice behavior and infer preferences in order to construct a utility function—there is no need to know "quantities of pleasure" or "quantities of satisfaction." That is, utility has a *formal* interpretation, without dubious substantive undertones.[13]

In order to construct a utility function for your preference ordering, your preferences need to conform to certain formal requirements. These requirements differ in different systems of axiomatization. But all of these share the basic idea: given that your preferences conform to a certain set of requirements, they can be assigned values (real numbers) such that their magnitude represents your preference ordering. Here is an example. Suppose there are three alternatives, x, y, and z; and that you prefer x to y to z. Any utility function will represent your preference ordering

[11]The most famous figure behind this was Lionel Robbins. For his description of the intellectual development behind the revolution, see Robbins (1938).

[12]Presenting different versions of modern utility theory is beyond the purview of this work. Fortunately, all I need to use from the theory is the basic idea presented here. For introductions, see Luce and Raiffa (1957:12–38, 275–326) and Kreps (1988). For a survey, see Schoemaker (1982), Fishburn (1968), and the references therein. For subjective utility theory, see Savage (1954), or Fishburn (1981).

[13]See Ellsberg (1954). Neumann and Morgenstern "defined numerical utility as being that thing for which a calculus of mathematical expectations is legitimate" (Neumann and Morgenstern, 1944:28). That is, utility has only operational, and no substantive, content.

if and only if it assigns utility values for x, y, and z such a way that their magnitude corresponds to your ordering:

$$x \succcurlyeq y \succcurlyeq z \Leftrightarrow u(x) \geq u(y) \geq u(z).^{14}$$

In axiomatic utility theory, a preference ordering is represented by an *ordinal* utility function. As long as the magnitudes of the values (utilities) of this function conform to your preference ordering, any set of real numbers will represent your preference ordering. For instance, both $u(x) = 1; u(y) = 0.5; u(z) = 0$ and $u(x) = 11; u(y) = 4; u(z) = 1$ are adequate representations of your preferences $x \succ y \succ z$. The differences between the utility values do not matter. A rational person will choose in a manner to maximize ordinal utility.

Since infinitely many ordinal utility functions can represent a person's preferences, there are only limited ways of making welfare judgments (assuming, for the moment, that utility measures well-being) on this theory. In this respect, modern axiomatic utility theory breaks with the classical, "cardinalist" tradition, and the reason the theory employs such utility functions to represent preferences over alternatives under certainty is not unrelated to the arguments of the "ordinal revolutionaries" of 1930s economics.

Modern axiomatic utility theory is used for decisions under certainty. I called x, y, and z *alternatives*. Your choice between them leads directly to an outcome: if you choose x, you will get x. Your preferences, therefore, are directly over outcomes. In axiomatic expected utility theory, however, you have to make your choice under conditions of *risk* or *uncertainty*. This means that choosing an alternative will *not* lead directly to an outcome, but to any one of several outcomes with certain probabilities. Your preferences now are over *prospects*.

Here is how we proceed in expected utility theory. Suppose that x stands for \$2, y stands for \$1, and z stands for \$0. You prefer x to y to z. Select $u(x) = 1$, and $u(z) = 0$. What is the value of $u(y)$?

In order to construct a utility function, I ask you to state a preference between getting y for certain and a gamble (or lottery) in which you get x with probability p and z with probability $(1 - p)$. What I am interested in is the value of p at which you are indifferent between the "certain prospect" and the "lottery prospect":

$$y \sim (px + (1 - p)z).$$

Suppose $p = 0.5$. Then your utility function has the values, $u(x) = 1, u(y) = 0.5, u(z) = 0$. This *expected utility function* is *cardinal* in the mathematical sense.

[14] "$\succcurlyeq$" designates the *weak preference relation*: a person weakly prefers x to y if she either strictly prefers x to y or she is indifferent between them. Thus, $x \succcurlyeq y$ can be interpreted as "y is not preferred to x," and strict preference ($\succ$) and indifference ($\sim$) can be defined in terms of the weak preference relation. I use examples of strict preference below for ease of exposition.

Cardinality is the following property of the function: the class of transformations under which the informational content of the function is invariant is limited to the form $au + b$, where $a > 0$. This is equivalent to saying that an expected utility function is unique up to positive affine transformations.[15] Thus, your preferences are just as well represented, for instance, by the function with the values $u(x) = 2, u(y) = 1, u(z) = 0$ or $u(x) = 8, u(y) = 3, u(z) = -2$.

Of course, in the interesting choice situations the probabilities with which the outcomes may obtain are already given.[16] Consider an example of a *basic risk paradigm*, depicted on Figure 2.1 on the following page. Suppose you have to choose between two options at node Y. You can either "move down," or "move across." In the former case, your choice leads to an outcome with the monetary value \$1. In the latter case, your choice leads to either one of two outcomes according to some variable outside of your control. This is captured by the idea of some state of nature obtaining, or "Nature making a move" at node N. Nature can move "down," which leads to an outcome with the value \$0, or move "across," which leads to an outcome with the value \$2. As it happens in this case, the probabilities of Nature moving down and across are equal. Thus, the expected values of moving down and across at node Y are equal. Expected values are calculated by weighing the value of the outcomes with the probabilities with which they may obtain. In the former case, the expected value is $1 \times 1 = 1$, in the latter case $\frac{1}{2} \times 0 + \frac{1}{2} \times 2 = 1$. Moving down at node Y is choosing the "certain prospect," and moving across is choosing the "lottery prospect."

The expected utility representation of your preferences is determined by your *risk-attitude* towards these prospects. Note that your risk-attitude is *exogenous* in the theory: thus, if you prefer the certain prospect, you are *risk-averse* towards these prospects (your utility function is thus *concave*); if you prefer to take the gamble by moving across, you are *risk-seeking* towards these prospects (your utility function is *convex*); and if you are indifferent between the prospects, you are *risk-neutral* towards them (and hence your utility function is *linear*). The theory does not tell you which of these you ought to be! Your utility function, however, reflects your attitude towards the riskiness of the possible outcomes; this is sometimes expressed as saying that expected utility functions work with *ex ante* utility.

[15] Ordinal utility functions are unique up to positive (or increasing) monotonic transformations: if $u < w$, then a transformation, ϕ, is positive monotonic if and only if $\phi(u) < \phi(w)$. A positive (or increasing) affine transformation thus allows a more limited class of utility functions.

[16] These probabilities may be *objective* or *a priori*, i.e., given exogenously. In this case, the choice is under conditions of risk. Or the probabilities may be *subjective*—they are not known, and the decision maker must form beliefs about them. In this case, the choice is under conditions of uncertainty. Note that the terms "objective" and "subjective" probability are also used in theories of the nature of probability (as opposed to theories of rational choice). I use them exclusively as they are employed in theories of rational choice. Moreover, all of my examples involve only objective probabilities in this sense.

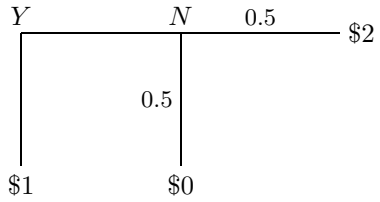

Figure 2.1

It is important to note that it is your weighing of the outcomes with the respective probability values and your risk-attitude towards the prospects which makes the assignment of an expected utility function possible. This function represents your preferences: it assigns a higher value to the prospect you prefer more. More precisely, the utility value of a prospect is the *expectation* of the *utilities* of the outcomes the prospect may lead to. This means that once we have calculated the expected utility values of the outcomes by the method given above (e.g., for utility values for $2, $1, $0), a rational person maximizes these expected utilities by maximizing their expectation—the risk-attitude towards them has already been taken into account when calculating the expected utilities of the outcomes.

It is also important to distinguish between the cardinality property of expected utility functions and classical cardinalism discussed above. The former is a mathematical concept and the latter is a substantive interpretation of what utility represents. Modern axiomatic expected utility representations do not allow for the kind of interpersonal comparisons classical cardinalists hoped were possible, since there are numerous possible expected utility functions representing each decision maker's preferences.

The core idea of modern utility theory is the tenet that a rational individual maximizes expected utility. More formally, the *expected utility hypothesis* is

$$\max \sum_x p(x)u(x),$$

which expresses that a rational individual chooses the probability distribution ("lottery") p over the set of possible outcomes ("prizes") x whose utility expectation is maximal. Different versions of modern utility theory may differ in the axioms they require the preference relation to satisfy, or the interpretation they give to the concept of probability, or whether they hold that the expected utility hypothesis is descriptive or normative or both. But the acceptance of the expected utility hypothesis is common to most of them.

In the remainder of this section, I present two controversies surrounding modern utility theory that are important for my purposes. The first concerns the (math-

ematical) cardinality of expected utility functions. The legitimate transformations of expected utility functions maintain the ratio between *pairs* of the *differences* between the utility values of the function. But does it follow that, for instance, if $x = \$2, y = \$1, z = \$0$, and your utility function has the values, $u(x) = 1, u(y) = 0.5, u(z) = 0$, then you value the difference between \$2 and \$1 just as much as you value the difference between \$1 and \$0? Do the magnitudes between utility values represent intensities of preference? The majority opinion is that they do not. However, there are philosophers who argue that expected utility functions should be interpreted this way. Since cardinal utilities are established by the person's reactions to risks, those who argue that cardinal utilities express intensities of preference must show how risk-attitudes and preference intensities are connected.[17]

The second controversy concerns cardinalism as a substantive view and the expected utility hypothesis. Traditionally, for the last half a century or so, the debate on the expected utility hypothesis has taken place between the "American school" and the "French school" of risk theory. While the former advocated the expected utility hypothesis, the latter has stressed that real-world decision makers do not seem to follow the norms of the Neumann and Morgenstern theory, or, in general, modern axiomatic expected utility theory; hence modern utility theory is deficient for both descriptive and normative purposes. The most vocal representative of the French school is Maurice Allais, who is a *neo-cardinalist*: he says he believes that

> cardinal utility in Jevons' sense exists. Some, including myself, even believe that it can be defined *independently of any random choice* by reference to the *intensity of preferences* ... (1984*b*:28)
>
> [and] the intensity of our preferences is *an indisputable datum of our introspection* and for a long time in other fields the psychophysiologists have been able to measure the intensity of our sensation by appropriate indices. (1984*b*:54, all emphases are his)[18]

It is important to emphasize what the disagreement between modern cardinalism and modern expected axiomatic utility consists in. The parties to both sides agree that under certainty, ordinal utility functions can be established. Their disagreement focuses upon the problem of cardinal utility; but they both agree that cardinal utility functions are unique up to only positive affine transformations. In this sense, modern cardinalism is different from classical cardinalism: it holds that

[17]For the "majority view," see, e.g., Luce and Raiffa (1957:22); for the contrary view, see Harsanyi (1993) and Binmore (1994:277–8). I will return to this controversy in Section 10.2.

[18]See also Allais (1953, 1988, 1994) and Hagen (1984, 1994). Allais himself carried out such research with questionnaires (see his 1984*a*). Breault (1981) reports that around the late 1950s, "psychophysicists" were able to determine a decreasing cardinal marginal utility function for money: it is approximately $u = kD^{0.43}$, where D is dollar value and k is a constant; thus, to get twice the amount of utility, you have to get approximately five times the amount of money.

not only one "unique" cardinal utility function represents a person's preferences, but a class of such functions. Cardinalists, like Allais, however, also believe that cardinal utility represents a person's intensities of preference, and it can be established on the basis of introspection—the person can report degrees of her pleasure or psychological sensations for this purpose. This is to say that modern cardinalism works with *ex post* utility: cardinal utility values are established without reference to the riskiness of the choice, based instead on the person's valuations of the outcomes. This is what Allais expresses by saying that cardinal utility can be defined independently of random choice.

Modern axiomatic expected utility theory, on the other hand, denies that intensities of preference by introspection can be established. That is denying that degrees of pleasure or sensation have any role in utility theory. Instead, the theory holds that *ex ante* or expected utility can be established with the aid of the person's reactions to risk, but the cardinal utility functions of the theory do not correspond to any psychological notion. Those representatives of the theory who nonetheless see it fit to talk about "intensity of preference" make a substantive argument about the interpretation of cardinal utility values.

The advantage of the approach of the American school is that the expected utility hypothesis is elegant: decision making in risky situations is characterized by maximizing a measure, expected utility. In contrast, the French school stresses that people do not behave this way: they do not maximize a simple measure, but they also look at properties of the probability distribution—like the dispersion of probabilities, the variance and skewness of the probability values of the distribution in lotteries, and they simplify complex risky prospects by certain salient characteristics. For instance, they may set "aspiration levels" and judge lotteries based on their expectation of whether these levels can be met. Thus, on the one hand, empirical studies on risk behavior tend not to confirm the expected utility hypothesis.[19]

On the other hand, it is unclear what these results imply for axiomatic expected utility theory, especially in normative contexts. The fact that people are not perfectly rational should not come as a great surprise. In any case, it is not my task to adjudicate in the debate between neo-cardinalists and proponents of the expected utility hypothesis. I am interested in the relevance of utility theory to normative applications: whether the concept of utility has any role in a theory of well-being

[19]The most famous experiments were carried out by Kahneman and Tversky (1979). They found that people overweight certainty, they are risk-averse in the domain of gains and risk-seeking in the domain of losses even if the monetary payoffs are equivalent. Lopes (1986) finds that decision makers look at cumulative probabilities, the distribution of probabilities, and they assign aspiration levels. MacCrimmon and Wehrung (1986) report that risk propensity is different in different contexts: business executives are for instance more risk-averse in personal decisions than in business decisions, again with the same monetary values at stake. For a detailed and general survey of experimental results for modern utility theory, see Camerer (1995).

for the making of welfare judgments. As the first part of the quotation from Friedman and Savage on page 13 stresses, in order to show that utility theory is useful in a theory of welfare, we must argue that what utility represents is indeed ethically important. It is in this respect that neo-cardinalism is disputable.

2.3 Utility Theory and Ethics

For modern cardinalists, utility functions represent preferences, and preferences are determined by "subjective values," "psychological values," "satisfactions," "pleasure," "degrees of sensation," and the like; utility is the "*psychological concept* which has been banished from [economic theory] after Pareto's time."[20] It is natural to think, then, that the underlying conception of well-being for the neo-cardinalist project is some sort of *hedonism*. If introspective cardinal utility is to have any ethical relevance, it must be so because hedonism has ethical relevance. Thus, from the perspective of an inquiry into welfare, neo-cardinalism is relevant only if hedonism is the correct theory of well-being.[21]

Consequently, it is warranted to concentrate on the philosophical problem—the plausibility of hedonism—as the next step. Chapter 3 discusses this theory. It looks only at the contemporary debate surrounding the doctrine; exploring the various forms hedonism has historically taken is beyond the scope of this work. Since most contemporary philosophers reject hedonism as a theory of well-being, it suffices to examine the argument they give. Then, Chapter 4 asks whether an hedonist account of welfare, even if it was plausible, could be operationalized with the aid of modern axiomatic expected utility theory. It argues that it can not, and such projects are misconceived.

For modern axiomatic utility theory, a utility function is not more than a representation of a person's preference ordering, such that if the person, from x and y, prefers x to y, the function assigns a higher number to x. The function does not represent any feeling about, or pleasure derived from, x and y. On this interpretation of utility, it is a mistake to say that the person prefers some alternative, x, because it yields higher utility; x yields higher utility because the person prefers it, and not *vice versa*. That is, preference is logically prior to utility.

[20] Allais (1984*a*:48, his emphasis). He even calls units of cardinal introspective utility "jevons" (1994:3), as utility units are sometimes called "utils."

[21] I do *not* intend to suggest that Allais himself is an hedonist, let alone that any other modern cardinalist is one; as far as I am aware, they do not discuss ethical issues, and they see their project as a merely descriptive approach to economic theory. At the same time, for what other reason, apart from some implicit commitment to hedonism, would this psychological concept of utility be ultimately important? At one place where Allais does discuss normative issues (1988:69–70), he says that introspective cardinal utility and interpersonal comparisons of introspective cardinal utility are indispensable for any theory concerned with social choices, distributional issues, and government policies.

This, however, leads to the question raised in the second part of the quotation from Friedman and Savage on page 13—the question why utility should have any normative significance in welfare economics and ethics. Obviously, the answer must be that *preferences*—more precisely, the satisfaction of preferences—have normative significance in ethical, political, and economic decision making. Thus, we need the concept of utility because utility is the index of preference satisfaction, so it is indispensable for practical applications. The sophisticated mathematical constructions of utility theory are useful tools for determining how to promote the satisfaction of preferences. But utility theory is only relevant to ethics and welfare economics if the satisfaction of preferences has relevance to these disciplines.[22] What reasons are there for thinking that preferences have normative significance?

I can think of three views to defend the relevance of preferences. The first could be called the *consumer sovereignty* view. On this view, it is in itself valuable that preferences are satisfied, and utility theory is relevant because it measures the satisfaction of preferences. The problem with this view is that it cuts short, instead of resolving, the problem. Individuals can and do have preferences which are objectionable—they are not seldom based on lack of information or shortcomings of reasoning, or they might be objectionable for other reasons, for instance, for reflecting antisocial feelings, envy, hatred, and the like. The consumer sovereignty view makes all these preferences normatively important. Perhaps the view can be made more attractive by appealing to individual *autonomy*: preferences are important because the person expresses her autonomy through them, thus, they must be respected. Of course, this version is only attractive if it manages to exclude not only preferences which are objectionable in the ways mentioned above, but also preferences which are non-autonomous in some other ways: for example, they are the result of subtle manipulation or brain-washing.

The other two views appeal to well-being. According to the first one, preferences are important because they are reliable indicators of what is good for the person: preferences reliably "track" welfare. Hence, preferences ought to be respected because individuals are the most reliable judges of what promotes their welfare. This is the *best judge principle*: the person is the best judge of what is good for her.

Whether or not the best judge principle is correct is ultimately an empirical question. A defense of this position must show that a person tends not to misjudge what is good for her and, in any case, she is a better judge of her well-being than

[22]This is not to say that modern utility theory is not useful in *positive* applications of economics and the social sciences in general, but these are not my concern here. In any case, economists often point out that at the price of some technical inconvenience, the concept of utility could be discarded from these applications. One attempt to do this is *revealed preference theory*; see Samuelson (1938*a*,*b*), Richter (1966), Kreps (1988:11–5), and Sen (1971, 1973, 1993, 1994). It is, however, highly doubtful that normative applications could do with revealed preferences only.

anyone else—in the context of social policy, she is a better judge than government officials. This is evidently a controversial issue. The view must also specify what the conditions are in which a person can be considered to be a reliable judge of her well-being.[23]

Finally, perhaps the most straightforward way to argue for the relevance of preferences is to argue that welfare consists in the satisfaction of preferences. But because there are objectionable preferences and preferences might not be formed in appropriate ways, this view is likely to take the form of an idealized preference satisfaction view: it specifies the conditions under which the preferences of the person determine what is good for that person. What is good for the person is the satisfaction of her preferences formed in ideal conditions—or, equivalently, the satisfaction of those of her preferences which withstand the test of idealization.

The most popular theory like that is the ideal advisor theory. The version I prefer holds that welfare consists in the satisfaction of the preferences the person would have were she fully informed and ideally rational. Now, it turns out not only to be a very attractive and influential account of well-being, but one that can be operationalized in terms of utility theory to enable us to make welfare judgments. The theory can use utility functions to represent fully informed and ideally rational preferences. It is a complete theory of welfare in the sense that it tells us in virtue of what something promotes a person's well-being, and it can also be operationalized. Indeed, in a recent survey on the relation of utility theory to ethics, it is proposed as the theory that can establish the relevance of utility to well-being, and underpin the importance of a formal approach to ethics.[24]

The case for the relevance of utility theory to well-being is most likely to appeal either or both to the idea that welfare consists in the satisfaction of preferences, or to the idea that preferences must be respected because they express the person's autonomy. In this work, I examine both possibilities. Chapters 7 and 8 discuss the ideal advisor theory as a theory of well-being. Chapter 9 addresses the relation of preference and autonomy, and Chapter 10 returns to the problem of welfare judgments.

[23]Goodin (1990) discusses several interpretations of the principle.

[24]See Mongin and d'Aspremont (1999). John C. Harsanyi (1953, 1955) already proposed this theory for acceptance in welfare economics half a century ago.

Chapter 3

The Experience Machine Revisited

3.1 The Thought Experiment

Historically, hedonism has often been held as a comprehensive theory of value. Hedonists could explain the whole realm of human valuing by resorting to some version of their theory. They could explain action by *psychological hedonism*, the view that the source of human motivation is striving for pleasure and avoiding pain. They could use *ethical hedonism*, the view that the rightness of actions and the goodness of outcomes are a function of how much pleasure they bring into the world, to give an account of morality. And they all held a distinctive mental state theory of well-being: they believed it consists in pleasure or happiness.

Few philosophers today accept such a comprehensive hedonist theory. Psychological hedonism is widely rejected, though ethical hedonism still has some defenders.[1,2] Hedonism as a theory of well-being, however, is alive. But it is alive and not well.

This is because many philosophers today believe that the following thought experiment—the experience machine—by Robert Nozick refutes hedonism as a theory of well-being:

> Suppose there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences? (Nozick, 1974:42)

The intuitive answer most of us would give is *no*. Nozick explains our reluctance to give our lives to the experience machine by pointing out that how our lives "feel from the inside" is not all that we value. He gives three reasons for this. First, we also want to *do* certain things, and not just feel as if we were doing them. Very

[1]It is commonly accepted that Bishop Butler (1726) refuted psychological hedonism. His argument is that any pleasant experience presupposes a desire for something other than pleasure, that is, a desire for some aspect of the experience that gives one pleasure; but the latter cannot be then derived from the desire for pleasure. For example, the desire for eating (a pleasant experience) presupposes a desire for food—eating is pleasant only if you desire food, but then you already desire something else than pleasure. The success of this argument is questioned by Sober (1992). See also Section 3.3.

[2]One contemporary proponent of ethical hedonism is Fred Feldman (1988, 1996, 1997*a*).

often, what we value is the activity itself, and not just the resultant feeling of experience. When you write a poem, you find the writing itself valuable, and not—or not only—the way that activity feels. Second, we also want to *be* certain ways, having particular character traits and exercising our capabilities in certain ways. We value generosity and honesty, for example, and we strive to become generous and honest—and we do not just want to feel like we were generous or honest. Finally, we want to be in touch with reality. We want our lives to be in accord with the way things actually are, and we want to have genuine relations with other people. We do not care only about the experiential quality of our lives, and the feelings relations with other people are accompanied by. We want our experiences to be genuine. In the words of James Griffin:

> I prefer, in important areas of my life, bitter truth to comfortable delusion. Even if I were surrounded by consummate actors able to give me sweet simulacra of love and affection, I should prefer the relatively bitter diet of their authentic relations. (Griffin, 1986:9)

Notice, firstly, that the target of the thought experiment is not psychological hedonism. We already know that this view cannot be true, if, upon reflection, we agree that we would not plug in. In any case, we don't need such an elaborate thought experiment to discard that doctrine. Secondly, the experience machine does not provide an argument against ethical hedonism. Even though it would be quite extraordinary, it is not impossible that although we have reasons to value, when we deliberate about what would be best *for us*, not only how experiences feel from the inside, it is nonetheless all we have reason to take into account when we deliberate about what would be morally right. The target of the experience machine is hedonism as a theory of well-being.

Furthermore, note that the first two reasons Nozick gives to explain our reluctance to plug into the machine collapse into the third one. When you value an activity, you want that activity to actually take place, and you value it as it actually takes place. When you want to be a certain way, you want your being that way real. In short, the point the experience machine makes is that our valuation goes beyond our inner world. Your well-being is determined by more than how your life feels from the inside. Many philosophers agree on this point. They agree that the thought experiment shows that hedonism cannot incorporate our preference to be in touch with reality, hence it cannot be a plausible theory of well-being.[3]

[3]These philosophers include Nozick (1974:42–5), Finnis (1980:95–7), Griffin (1986:9–10), Attfield (1987:33), Thomson (1987:40–4), Brink (1989:223–4), Kymlicka (1990:13–4), Louden (1992:48), and MacNiven (1993:4), among others. Nozick returns to, and gives a more thorough analysis of the thought experiment in Nozick (1989:104–8). For an even more sustained analysis, see Finnis (1983:37–42) or Sumner (1996:94–8). The thought experiment is anticipated by Smart (1973:18–23).

Hedonists, however, do have a counterargument. They can explain the reluctance of most of us to plug into the experience machine by pointing out that valuing authenticity, being in touch with reality, or lack of delusion is not incompatible with their view of well-being. It would not be worse to plug in, though it probably would not be the right thing to do given that we care about other values besides well-being:

> There are things we desire other than our own well-being, and by plugging into the machine we would forgo some of them forever. ... Most of us would choose not to plug into the machine for this reason, *which is independent of any concern for our own well-being*. (Goldsworthy, 1992:18, his emphasis)

The hedonist concedes that our intuition not to plug in is right, but she gives a different reason for it. She claims that the proponents of the experience machine have mislocated the source of the intuition. There are many things we value, and the reason we value authenticity is not that it literally makes our lives better:

> If we value things besides our mental states, they can produce rival values to our own well-being. The values can come into conflict. Thus, it is clear that the mere existence of other values does not show that they must necessarily contribute to our well-being. There is a large gap between the premiss that we value more than our mental states and the conclusion that there is more to our personal well-being than our mental states. Seeing our commitments and values as rival values to our well-being allows us to properly understand cases in which it might seem that more contributes to our well-being than our mental states. (Kawall, 1999:385)[4]

It seems to me that this counterargument is not implausible at all. We do care about more than our own well-being. Hedonists say that when we are reluctant to plug into the experience machine, our reason is not that we think it would be worse for us to do so—our reason is given by some other consideration. The experience machine could seem so convincing to so many philosophers only because they have not separated reasons given by our well-being from reasons given by other values.

Moreover, it is not an easy task to adjudicate in this debate, since, as it stands, the thought experiment is not an argument, but a mere appeal to our intuitions. Its proponents have seemed to think its appeal is powerful enough to make developing it into an argument superfluous. The hedonist counterargument shows that it needs to be developed if it hopes to be successful. In this chapter, I attempt to do just that. I argue that the thought experiment, even if not convincing in its present form, gestures at an untenable background assumption of hedonist theories of well-being. Once that assumption is uncovered, it can be seen that the hedonist counterargument does not salvage hedonism.

[4]In addition to Kawall (1999) and Goldsworthy (1992:14–20), Haslett (1990:91–3), Bernstein (1998), and Silverstein (2000) make similar points.

3.2 Principles of Indifference

Let me first ask what the comparison we are asked to make is when we are deliberating whether to plug into the machine. When we say *no* to plugging in, our reason is that we prefer our experiences to be authentic—we want real experiences, relationships, and genuine efforts to realize our own ends, given our real circumstances, instead of virtual substitutes. But that *no* must immediately be qualified. Most of us probably would choose to plug in for short experiences, to virtually orbit Earth, to go diving at the Great Barrier Reef, and so on. Moreover, some of us might accept the offer of the "superduper neuropsychologists" if their alternative was a profoundly better life. That is, for example, if the alternative was a lifetime of misery, some of us would surely want to plug in for a pleasant, enjoyable life. Therefore it matters what the comparison we are asked to make is. Since the ultimate question is whether one has a reason to prefer an authentic life over a lifetime spent in the experience machine, I take it that the relevant choice is between the life you have now and its *identically pleasant* counterpart spent in the machine.[5]

So here is the relevant scenario. You can choose between plugging into the experience machine or not. Your life in the machine will be identical to the life you would otherwise have with respect to its hedonic mental states (suppose the operators of the machine can establish what your life would be like if you did not plug in). That is, the experiential quality of your life is going to be the same either way. Moreover, if you choose to plug in, you will not know that you are in the machine (if you did, that would change the experiential quality of your life). Which alternative should you prefer?

Non-hedonists claim that *one* reason to prefer the authentic life over the virtual one is that it is better for the person whose life it is. Hedonists, in contrast, reject this; they hold that if it was the case that no other reasons were relevant, you should be indifferent between the two lives. They concede there are relevant other reasons to refuse to plug in, but the reason that plugging in would be worse for you is not a good reason. After all, the two alternatives are the same "from the inside."[6]

[5] I use the term "pleasure" as a placeholder for any mental state hedonists usually argue constitutes well-being. If they think, for instance, that only happiness is ultimately valuable, they can just substitute the terms. I sometimes also say "hedonic mental states" to emphasize that the argument pertains to all common versions of hedonism about well-being.

[6] Kim Sterelny pointed out to me that the thought experiment is a nonstarter if hedonists become externalists about mental content in general, and about pleasure in particular. In this case, experiences are individuated partly by reference to external states of affairs, hence you are not going to have the same experience when you are in the machine and when you are outside. Another modification of hedonism which avoids experience machine objections is Fred Feldman's "veridical intrinsic attitudinal hedonism" (2002*b*:614–6). On this view, "takings of pleasures enhance the value of a life more when they are takings of pleasures in *true* states of affairs" (2002*b*:616, his emphasis). While the former view denies that veridical and illusionary experiences give rise to the same mental states,

So the ultimate question is neither whether delusion can ever be better for us than reality, nor whether we have reason to prefer reality to delusion on grounds other than that it is better. The parties to the debate can agree on these points. The ultimate question is whether, when the time spent in the machine and the time spent outside would be identical with respect to the hedonic mental states they involve, you have a reason to be indifferent between them. The debate between the hedonists and their opponents, then, comes down to the plausibility of a claim an hedonist is committed to make: in the absence of other reasons, you should be indifferent between refusing and accepting to plug in if the lives inside and outside of the machine are experientially identical. Ultimately, non-hedonists say there is a reason not to plug in—that it would be worse for you; hedonists, in contrast, claim there is a reason to be indifferent between plugging in or not.

Note that even though the original thought experiment involved a comparison of lives, this hedonist claim must apply to all experiences, not only whole lives. In fact, focusing on lives just muddies the waters. It is unfortunate that the thought experiment was couched in terms of such evaluations in the first place. The experience machine gestures towards an implicit claim of hedonists, one whose scope is wider than the proponents of the thought experiment may have thought. If the accompanying hedonic mental states of any two experiences—no matter how isolated and momentary, or over-arching and extended—have equal value, then, barring other reasons, hedonists think you have a reason to be indifferent between them.[7] So it is more useful to focus on the evaluation of any experience, instead of restricting our attention to the evaluation of whole lives.

In accordance with this, suppose now that you had chosen to plug into the machine and the machine has given you identical experiences to those your real life would have, had you not chosen to plug in. But one day, the machine breaks down, and you come out. You learn the truth: for the last few years, you have been

this view denies that veridical and illusionary experiences are equally valuable. (Feldman in effect denies that veridical intrinsic attitudinal hedonism is committed to any principle of indifference; see below. See also DePaul (2002) and Feldman (2002*a*).) Fascinating as they are, I will ignore these possibilities here. My reason is that both work by profoundly changing hedonism: the first changes what many of us think hedonism must hold about mental states, the second changes what many of us think hedonism must hold about value. Thus many of us would doubt that these modified views would be intelligible as *hedonist* views.

[7]By "having equal value," I mean that the experiences are just as pleasant, or have the same "amount" of hedonic mental states. That is, hedonists are committed to the view that if the quantities of pleasure two experiences are accompanied by are equal, then you have reason to be indifferent between these. Some hedonists, however, may want to distinguish between pleasures not only by quantity, but by quality as well. Their indifference claim, therefore, requires that the experiences are qualitatively on a par. (Such a view would have to give us trade-off levels between "higher" and "lower" pleasures.) In any case, even though I concentrate here on quantitative hedonism for the sake of simplicity, my arguments would work against qualitative hedonism as well—because that view is also committed to some version of the indifference claim.

floating in a tank, having been given the experiences that you would have had, had you not decided at some point to plug in. Has your life been worse in the machine than what it could have been otherwise? Since our intuition says we should not plug in, it seems we would regret having made the decision to plug in when the machine breaks down and we come out.

Seeing your regret, an hedonist would tell you the following: "I agree that you have reason to regret the decision you made to plug in. But having spent all these years in the tank has not, literally, been worse for you. After all, you had just as much well-being as you would have had otherwise. If only your well-being mattered, you should be indifferent between being in the machine and living outside. So your regret is based on something else, not on your faring worse than you would have."

A non-hedonist, on the other hand, would tell you, "It is not true that you cannot regret your decision in terms of your well-being. While I agree that you may have grounds for regret other than your well-being, it is perfectly possible that you have regrets on *that* ground. If you can have regrets on that ground, then you might not have fared as well being in the tank than you would have, had you spent all these years outside the machine."

The proponents of Nozick's thought experiment reject the hedonist's argument. They hold that there is nothing implausible in valuing authenticity, or being in touch with reality, or lack of delusion, on the ground that they make our lives, literally, better. But if they can make our lives better, when you spend all these years in the tank you fare more badly than you could have done. L. Wayne Sumner puts this argument the following way:

> If what you have accepted as an important constituent of your well-being—your achievements, say, or the feelings of others about you—turns out to have been an elaborate deception, you are likely to feel hurt and betrayed. How else to explain this, except to say that, in this area at least, what mattered to you was not merely how things seemed but how they actually were? Your reaction to the deception certainly looks, and feels, like a reassessment, in the light of your own priorities, of how well your life has been going. (Sumner, 1996:97–8)

Let me now ask whether you could be justified in making this sort of reassessment. As I argued, the thought experiment points to an important background assumption behind hedonist theories of well-being. Call this the *hedonist principle of indifference*:

(H) If the satisfaction of your preference for x and the satisfaction of your preference for y are equally pleasant, then, other things being equal, you have reason to be indifferent between x and y.

By "other things being equal," I mean that the only consideration that is relevant to you when you form your preference between x and y is which promotes your well-being more.

In order to test the plausibility of the hedonist principle of indifference, let me ask more generally when you have reason to be indifferent between two alternatives or goods. It turns out to be useful to see what economists say about this matter. One thing they do say is that people are typically indifferent between two bundles of goods if those goods are *perfect substitutes* for one another. Technically, if two goods are perfect substitutes, then their marginal rate of substitution is the same along any indifference curve of the consumer. Roughly, this means that people will not care about the ratio of the two goods that make up their bundle—they treat both goods as equally useful. For instance, butter and margarine may be perfect substitutes. For many people, it does not make a difference whether they spread butter or margarine on their toast in the morning. But whether two goods are perfect substitutes also depends on the context. Thus butter and margarine are perhaps not perfect substitutes—or so I am told—when you want to make a cake.

This suggests that whether two goods are perfect substitutes for a person depends on their *causal properties*. Thus we can turn the observation of economists into a normative principle. Call it the *causal principle of indifference*:

(C) If x and y are perfect substitutes, then, other things being equal, you have reason to be indifferent between them.

The *ceteris paribus* clause is needed again to remind us that we are interested only in indifference with respect to well-being—goods may be perfect substitutes with respect to their contribution to our well-being, but not with respect to their contribution to some other value.

(C) needs to be amended by an explanation of what makes two goods perfect substitutes. In general, it seems that if x and y are perfect substitutes, then x and y must have some relevant causal properties in common. But which causal properties are relevant? One possibility is given by (S_1):

(S_1) The satisfaction of your preference for x and the satisfaction of your preference for y are equally pleasant.

Of course, if we substitute (S_1) for the antecedent of (C), we just get (H). In this case, two goods or experiences are perfect substitutes if and only if the satisfaction of your preference for either results in the same amount of hedonic mental states—if and only if they are equally pleasant. (Note that in the debate surrounding the experience machine it is not doubted that pleasure, in general, contributes to well-being; the issue is whether it is only pleasure that does). Is this a plausible principle of indifference?

Consider the following example. You are forced to take either of two pills—a red or a blue one. You can choose freely the one you will take. You also know that taking either pill will result in equally pleasant feelings. If you take the blue one, you will feel mild contentment for a few hours, then the effects wear off. But if you choose the red pill, besides feeling mild contentment for a few hours, you will instantly get hooked—the red pill is addictive. Should you be indifferent between taking the pills? On the hedonist principle of indifference, identical experiential qualities justify indifference: equally pleasant options are perfect substitutes. On an hedonist theory, therefore, you should be indifferent between taking the red pill and the blue pill. But it is hardly controversial to maintain that there is reason to prefer the blue pill. It is better to prefer the blue pill to avoid addiction, because addiction is likely to induce undesirable changes in your subsequent preferences.

By construction, the only difference between taking the red pill and taking the blue pill is that the former causes addiction: it makes a difference to your subsequent preferences. An hedonist may object that this feature disqualifies the two experiences from being perfect substitutes. But then what she in fact does is conceding that (S_1) is not sufficient. She concedes that taking the red pill has a further causal property that is relevant. If she is willing to grant that, however, it becomes clear why it was unfortunate to set up the experience machine thought experiment in terms of whole lives. The only reason why hedonists can say that (other things being equal) you should be indifferent between plugging in and staying outside of the machine is that if the question is whether you would plug in for a whole life, there are no further preferences to which your experience in the machine could make a difference. But if the experience in the machine is not encompassing, knowing that you have been in the machine will make a difference to your subsequent preferences, and the hedonist must accept that such differences are relevant to the question whether you should be indifferent between the machine and reality. That is, choosing to plug in or to stay out—but having identical experiential qualities in either case—will make a difference to subsequent preferences. The reassessment of your life in light of the truth—recall the quote from Sumner—*is* a reassessment in terms of your well-being, since now you realize you should not have been indifferent: had you been in touch with reality, your preferences would have taken a different causal route.

Incidentally, note that nothing hinges on the example of addiction. Any example in which one alternative influences subsequent preferences in ways the other alternative does not would do, for it seems the hedonist must take these changes into account. Note also that I have been supposing that you eventually come to *know* that you have been under deception. This suggests a way out for the hedonist: if you never learn about the deception, there will be no changes in subsequent preferences.

But this is beside the point. After you plugged into the machine, you will not

know that you have plugged in. You cannot tell whether the experiences you have are real or only illusions. You probably don't even ask that question in the machine. But if counterfactually you would have different preferences if only you knew you were being deceived, it must be the case for the hedonist that those counterfactual preferences would not make a difference to your well-being—otherwise the plausibility of hedonism depends *solely* on your having false beliefs about your experiences.

In other words, in order to test the hedonist principle of indifference, we suppose that the hedonic mental states you have while you are in the machine are identical to those you would have if you were living outside of the machine. We keep these mental states fixed. We then ask what would happen if you learned the truth. It is possible that you would re-evaluate your time in the machine, and you might think you would have had different preferences, had you known the truth, and you might now come to have different preferences. If so, there is a reason you should not have been indifferent between plugging in and staying out. Next, we suppose you never learn the truth. But it probably would still be the case that counterfactually, had you known the truth, your preferences would have been different. If so, there was a reason not to be indifferent at the time of plugging in, which you did not know. But your not knowing it cannot make a difference to whether there was such a reason.

Consequently, the definition of perfect substitutability must be extended. An hedonist may want to try something like (S_2):

(S_2) Choosing either x or y does not make a difference to the satisfaction of those subsequent preferences whose satisfaction results in hedonic mental states.

On this reading, the antecedent of (C) is substituted with the conjunction of (S_1) and (S_2). What the new principle of indifference says is that you have reason to be indifferent between two experiences if they are equally pleasant, and if the experiences they may give rise to through altering your subsequent preferences are also equally pleasant. Note that an hedonist opting for this view should also accept that there are two kinds of preferences: *hedonic preferences*, whose satisfaction results in some hedonic mental state, and *non-hedonic preferences*, whose satisfaction does not. The distinction is needed because if hedonists do not allow for the possibility of non-hedonic preferences, they are committed to a crude form of psychological hedonism—and I take it that they do not want their view to depend on such a discredited doctrine. Thus, the view under consideration now is, roughly, that as long as the choice between two equally pleasant experiences will not make a difference in the pleasures they will later lead to, you are justified to be indifferent between them.

Nevertheless, this new definition of perfect substitutability will not do for the hedonist's purposes. Even if choosing either of two goods or experiences will lead

to the same "amount" of pleasure from now on, there may be differences in non-hedonic preferences and their satisfaction may very well make a difference to how well one's life goes. So the distinction causes problems. Recall the choice between taking the red pill or the blue pill. Suppose first that you have taken the blue pill—you were prudent enough not to believe the hedonist when she told you that you should be indifferent between the pills. But actually you quite liked the effects of the pill, so you start taking it regularly anyway. After all, it has no harmful side-effects and does not interfere with other pursuits in your life. As it happens, you take the pill only when, had you taken the red pill and become addicted, your addiction would kick in. Second, suppose you took the red pill and you are addicted now. But you prefer not to have the addiction, although you still prefer taking the pills. So it is not that you disapprove of the feelings that the pill causes, and it is not that you are depressed because you are addicted. In fact, you realize you would be taking the exact same number of blue pills anyway. It is just that you prefer not to be the sort of person who has addictions, or prefer to have the possibility of not taking the pills, even if you would never use that possibility. On the hedonist's account, there is no difference between the two lives, since each contains the same amount of pleasant experiences. Once again you should be indifferent between taking the red pill or the blue pill when you originally choose between them. But although you will have the same amount of pleasure either way, you will have different non-hedonic preferences.

And these non-hedonic preferences *do* make a difference to how well your life goes. If you chose the red pill, there will be a causal effect on your further preferences—the addiction. It seems to make a difference to your life: it seems to make your life worse for you than the life of taking blue pills without being addicted, given that you prefer not to be addicted. For the hedonist, that preference does not matter. In general, however, our hedonic and non-hedonic preferences are not separable: they interplay to form our subsequent preferences and influence their satisfaction. You typically have to take into account non-hedonic preferences when you form your preference for what would be good for you; you have no reason to be indifferent between two equally pleasant experiences which will make a difference to subsequent non-hedonic preferences, even if they do not make a difference to later hedonic preferences. Consequentially, hedonists cannot accept the conjunction of (S_1) and (S_2) to fill in for the antecedent of (C).

Can hedonists widen their principle even further? It might be thought that they could propose (S_3):

(S_3) The satisfaction of your preference for x and the satisfaction of your preference for y do not make a difference to any subsequent preference.

The antecedent of (C) is now filled in with the conjunction of (S_1) and (S_3). On this version, two goods or experiences are perfect substitutes if and only if they

are equally pleasant and they make no difference whatsoever to your subsequent preferences.

Consider again plugging into the experience machine for life. By my construction of the thought experiment, your life in the machine will be just as pleasant (or, for that matter, miserable), as your life would be, had you chosen to stay outside. Moreover, since you have plugged in for life, you do not know you are in the machine, and you never come out of it; hence the choice, by construction, makes no difference to your subsequent preferences. Thus, according to the hedonist, in such a scenario you have reason to be indifferent between plugging in and staying out. Is hedonism with this new principle of indifference plausible?

Whatever the answer may be, the scope of this new hedonist principle of indifference is very limited. In any less stringent scenarios involving changes in subsequent hedonic or non-hedonic preferences, the hedonist principle of indifference based on (S_1) and (S_3) will not be relevant. If the comparison is not between whole lives, but any less comprehensive experiences, the principle, in most cases, will not apply, because normally all of our choices make a difference to our subsequent preferences. But if they do, this version of the hedonist principle of indifference is usually irrelevant in our deliberation about what would promote our well-being.

As a matter of fact, however, even this principle of indifference under such a stringent scenario turns out to be implausible. In order to see this, consider the following real-life example.

In 1903, shortly after X-rays were discovered by Roentgen, a French physicist, René Blondlot (1849–1930), a professor of physics at the University of Nancy, reported that he had discovered a new type of radiation, which he called N-radiation after his home city. The discovery was made while he was trying to polarize X-rays. Many others have claimed to detect the radiation, and a number of scientific papers appeared on the subject. But many laboratories could not replicate the results.

Nature magazine finally decided to send Robert W. Wood, an American physicist of Johns Hopkins University—one scientist who could not detect Blondlot's radiation—to witness the experiments that purportedly prove the existence of N-rays. While observing the experiments in Blondlot's darkened laboratory, Wood surreptitiously removed a crucial component from the N-ray detection device—an aluminum prism. Blondlot and his assistant did not see this, but they claimed to have detected the rays in the experiment nonetheless. They deceived themselves due to their wishful thinking. Wood published his finding in *Nature*, and N-rays subsequently disappeared from science. Nevertheless, Blondlot insisted that he made a genuine discovery, and continued his research for many years.[8]

Let me construct an example from Blondlot's story. Suppose there are two

[8]For the story, an extensive bibliography and reprints of some of the original documents, see Ashmore (1993).

physicists—Blondlot* and Roentgen*—whose lives are identical down to the last detail, except for one minute difference. After much hard work, they both discover a type of radiation in identical conditions and through almost the same experiments. The only difference in their experiments is in the apparatus they use, and the only difference in their apparatuses is that Blondlot*'s uses an aluminum prism and Roentgen*'s uses a glass prism as a crucial component to detect radiation. In fact, this is the only difference in their lives. Blondlot* discovers N*-rays, and Roentgen* discovers X*-rays. There is one crucial difference between these two sorts of radiation: while X*-rays are a genuine natural phenomenon, N*-rays are not. But this fact does not come to light until well after their deaths. In their lifetimes, both physicists are equally famous and respected, and their lives are equally pleasant—since they are identical before and after the discovery. The question is: which life would you choose?

If the lives of Blondlot* and Roentgen* are perfect substitutes, an hedonist will think that you have reason to be indifferent between them. She thinks if she was to live either of these lives, she should not care which one she does live. But that cannot be right. If we asked her whether she would prefer to make a genuine or a bogus discovery, she surely would say that she prefers to make a genuine one. For if she did not, it seems that she either does not understand our question, or there is something wrong with her understanding of the aim of scientific pursuits and the point of discoveries—and how they can contribute to the value of one's life. That difference cannot be accounted for by hedonism. For the real Blondlot, it must have mattered whether the N-rays he claimed to have discovered were a real or a pseudo-phenomenon; his reason for trying to make a discovery wasn't that he wanted to be pleased to have made a discovery. Thus, he himself would have been likely to disagree that he should be indifferent between these possibilities, and the basis of his disagreement was not some other value: it is how well these lives would go.

What goes wrong with hedonism? I suggest the problem is that hedonism does not take seriously what it is for a preference to be satisfied. Consider again my earlier distinction between hedonic and non-hedonic preferences. The satisfaction of an hedonic preference is the obtaining of some hedonic mental state: its satisfaction is pleasant. But not all preferences are like that. The satisfaction of many preferences results in some mental state, but these are not the mental states required by the hedonist—their satisfaction is not having pleasure, as it were. And many other preferences are satisfied without resulting in any mental state at all. I surely prefer that many people will read this work, many of whom I don't know; and I will never know that they will have read it. Hedonists deny that the satisfaction of such a preference can contribute to my well-being. On this, some non-hedonists agree with them. But hedonists also deny that the satisfaction of any other non-hedonic preference can contribute to our well-being; they even deny that the *having* of non-

hedonic preferences can make a difference to well-being. For hedonists, only the satisfaction of hedonic preferences matters, everything else is irrelevant.

3.3 Reallocating the Burden of Proof

In the debate on the experience machine, hedonists have taken the lead. They can easily counter Nozick's thought experiment by pointing out that the source of the intuition that drives it is mislocated. What I have tried to do is to put the ball back into the hedonist's court by arguing that hedonists are committed to an indifference principle, and that none of the versions of this principle are plausible.

One way hedonists might be tempted to counter this argument is by falling back on a more sophisticated form of psychological hedonism. For instance, Silverstein (2000) argues that hedonists can defend their theory against the experience machine by incorporating further components into their account of well-being. They can do this, first, by giving an account of the formation of desires along the lines of Railton (1989) and Brandt (1979), who suggest that the ultimate source of all of our desires is pleasure. On their story, we initially come to have our desires because we expect their satisfaction to result in pleasure, and later, through a process of positive reinforcement by the pleasant experiences accompanying the satisfaction of these desires, we come to hold them for their own sake. (And these desires may generate other, possibly non-hedonic, desires.) That is to say that our reasons for our desires, even though originally they spring from our striving for pleasure, eventually extend beyond the hedonic mental states they realize.

A further component is given by the consideration, long noted by hedonists, that pleasure can easily elude us if we aim for it directly. A better strategy is to have a diverse set of aims and pursuits in one's life, to work towards a plurality of goals in order to most effectively secure pleasure (or happiness) for ourselves. For this, we need to care about more than how our lives feel from the inside. Being in touch with reality, for one thing, saves us from a lot of pain and unhappiness in our lives. We value authenticity because deception tends to cause misery. Given these two components, hedonists can argue that well-being consists in pleasure, but it does not follow that we should care only about pleasure; on the contrary, the story on desire formation and the need for the indirect pursuit of hedonist ends explain why we are better off not to do so.

Let us grant the hedonist these components. After all, the first is an empirical account of our psychology, which may or may not be true; and the second seems plausible enough. The new hedonist counterargument is then this: given the story of how we come to have our desires, and given that an indirect strategy for the pursuit of our happiness is the most efficient, we can explain our intuition not to plug into the machine. We note that we are conditioned to strive for happiness in an indirect way, and through a complex process of desire-formation that extends

beyond valuing the experiential quality of our lives only. When we are faced with a thought experiment such as the experience machine, however, these desires and intuitions are not reliable because they are formed and reinforced in conditions that are assumed away in the thought experiment. Since in real life we do not have foolproof experience generators, and since we are more efficient in accomplishing our hedonic aims in pursuing them indirectly, our intuition not to plug in does not show that hedonism is implausible. The experience machine breaks down.

As opposed to the thought experiment as it is usually presented, this story has the advantage that it actually explains our intuition. As I remarked earlier, it is extraordinary that so many philosophers have taken the experience machine to provide a conclusive refutation of hedonism as a theory of well-being, while not bothering with giving an explanation for why we share that intuition. Appealing to intuitions without giving a reason why we have them is hardly useful in settling controversies. Hedonists could make an easy comeback by providing alternative accounts of why we have the intuition not to plug in.

Nevertheless, I am at a loss to see why such a story on the formation of desires would save hedonism. Even if the story was true, why would it support hedonist theories of well-being? Suppose initially I desired achievement because it brought pleasure; now I have an intrinsic desire for achievement, and I have this intrinsic desire even if, in some cases, achievement does not bring pleasure. Does it follow that in those cases achievement is not good for me? If, according to hedonists using this story, it is not good for me, then they have merely restated their position. If it is good for me, then it is unclear why it would be good for me only *in virtue of* the causal background of my desire for it. The general problem is that it seems to be an implausible idea that the value of the satisfaction of our desires is determined by their causal history.

What I have tried to show is that the hedonist counterargument to Nozick's thought experiment is not the end of the debate. Hedonists are committed to a principle of indifference. All versions of this principle seem to be implausible. Given this, it is not sufficient for hedonists to give an account of our intuition. They have to defend some version of their principle. Until some such defense is forthcoming, we have a reason to reject hedonism as a theory of well-being.

Chapter 4

Hedonism and Utility

4.1 Interpretations of Utility

James Griffin opens his influential book on well-being with the words, "How are we to understand 'well-being'? As 'utility,' say the utilitarians…" (1986:7). But not all utilitarians say this. For instance, Sidgwick didn't. And even those who do, do not *identify* well-being with utility; they hold that utility is the measure or *representation* (in the mathematical sense) of well-being. Therefore they need an account of well-being to tell us what utility is the representation of.

Utilitarians usually subscribe to either of two conceptions of well-being. On the one hand, many utilitarians are *hedonists*: they accept the view that identifies well-being with happiness, and analyzes happiness in terms of pleasure. Typically, however, modern hedonists want to include, not only pleasure and happiness, but many other valuable mental states in their account of welfare. On the other hand, there are utilitarians who are not hedonists. Most modern utilitarians accept the *preference satisfaction* account of well-being: they hold that welfare consists in the fulfillment of preferences.

Most utilitarians, from both of these camps, do indeed interpret utility as representing well-being. They hold that to promote welfare is to maximize aggregate utility—and they further disagree about the formula aggregation should take. When utilitarians who accept a preference satisfaction view of well-being talk about utility, they typically refer to the concept as it is used in modern decision theory. They use the Neumann and Morgenstern (1944) utility theory or Bayesian decision theory as the starting point for their arguments for interpersonal comparisons and aggregation of well-being.[1] The question then arises whether hedonist utilitarians can do the same: can modern hedonists interpret utility as a representation of pleasure, happiness, or valuable mental states in general in the vein of modern utility theory?[2]

I am interested in the possibility of this sort of hedonism. One of its basic tenets would be the interpretation of utility as the representation of pleasure, happiness, or other valuable mental states. On this theory, utility is an index of how much pleasure different alternatives result in. If your preferences rest upon correct assessments of how much pleasure some course of action will bring about, when you

[1] See, for instance, Harsanyi (1953, 1955, 1977*b*, 1978, 1985, 1995); also Mongin (2001).

[2] For the distinction between these two kinds of utilitarianism, see, e.g., Sen (1980).

maximize utility, you maximize pleasure. In other words, when you are concerned with promoting your welfare, you will prefer the option with more expected pleasure; and that is just to say, you will prefer the option with higher expected utility.

Is this a tenable interpretation of utility?

4.2 The Concept of Pleasure

Before I discuss this question, I need to review the main rival conceptions of pleasure, in order to ensure that my argument in Section 4.3 is general enough.

According to hedonism, well-being consists in pleasure. On some of its versions, what is ultimately valuable is *happiness*, defined in terms of pleasure.[3] On some modern versions of hedonism, well-being consists in other valuable mental states as well. However, in order to keep the exposition simple, I will here concentrate on the concept of pleasure. It is not unreasonable to think that what I have to say about pleasure pertains to other valuable mental states too.

Jeremy Bentham held that pleasure is an homogeneous mental state, common to all experiences we find pleasant. This distinct mental state has its own quantity and duration.[4] Let me call this conception of pleasure the *monistic conception*, and the type of hedonism based on it *monistic hedonism*. Its proponents have several monistic interpretations of pleasure to choose from. They could simply say that pleasure is a distinct feeling. Or they could say with G. E. Moore that certain experiences cause a special sensation, which we identify with pleasure; this sensation or feeling is a natural kind that does not lend itself to further analysis, but one that is available for introspection. This account is the *causal theory of pleasure*.[5] Or they could hold that pleasure is not something "further to" experiences, but one of their intrinsic aspects. This is the doctrine of the *hedonic tone theory*: all and every experience can be rated along a single dimension of pleasant and painful. A pleasure is any mental event which has a "positive tone" to it; respectively, pain is a mental event with a negative hedonic tone to it. In Karl Duncker's metaphor: "hedonic tone" pervades an experience much like the way price relates

[3] One theory which denies that happiness can be defined in terms of pleasure is discussed in Chapter 5.

[4] More precisely, Bentham held that pleasure has intensity, duration, certainty, propinquity, purity, fecundity, and extent—factors which determine its value. In the following, I will not enter into a discussion of the various possible interpretations of pleasure as the concept is used by Bentham (1789), Mill (1861) and Sidgwick (1907). Although there are important differences between their views, they all used the concept in a comprehensive way to include related notions; therefore different interpretations are possible and have been given. An excellent brief overview is to be found in Sumner (1996:83–92).

[5] See Moore (1903:12–3, 64–6); for interpretations, see, for example, White (1958:116–47) and Baldwin (1993). Moore's view later in his life changed, so this is only his 1903 version. For criticisms of the causal theory of pleasure in general, see Feldman (1988:86–8).

to an economic commodity, which the commodity assumes in the context of exchange. It is a contextual—relative to the person whose experience and pleasure is in question—and accidental property. Moreover, just as prices vary and can be compared, pleasures and pains can vary and be compared.[6] A contemporary proponent of *a* monistic interpretation of pleasure is David Brink: "the one and only intrinsic good is pleasure, which is understood as a simple, qualitative mental state" (1989:221).

What these views share is that they all hold that pleasure is a distinct mental state. For some experience to be pleasant, it needs to produce, or to be accompanied by, this mental state. Whether an experience is pleasant for someone is independent of this person's attitude towards the mental state the experience gives rise to.

The monistic conception of pleasure has been thought problematic by many philosophers. Why should we, they ask, think that all pleasures are one sort of thing, when, phenomenologically, they seem to be so distinct? The pleasures of, for example, watching a peaceful sunset, reading, and playing do not seem to have much in common:

> Compare the pleasures of satisfying an intense thirst or lust, listening to music, solving an intellectual problem, reading a tragedy, and knowing that one's child is happy. These various experiences do not contain any distinctive common quality. (Parfit, 1984:493)[7]

Different pleasant experiences vary *qua* experiences, but why would anything else than "being pleasant" be needed? The answer is that there must be something *in virtue of* what some experiences can all be called pleasant. If there is no such thing, these experiences may be incommensurate. An hedonist needs to be able to explain in virtue of what all pleasant things can be identified and ranked, and why pleasant experiences are nonetheless phenomenologically and qualitatively distinct.

Because of this difficulty, the monistic conception of pleasure is almost universally discarded today. It is supplanted by the *attitudinal* conception of pleasure, which stems from the views of Sidgwick (1907), Brandt (1959, 1979), and Ryle (1954, 1949:103–6). Ryle's suggestion is that pleasure should primarily be understood not as a feeling, but as a *dispositional* mental state: something is pleasant not because it causes, or is accompanied by, pleasure—but because the person is reluctant to give up whatever the activity or source of her pleasure. According to the attitudinal view, what unites and identifies various pleasant experiences is that the subject who experiences them has some sort of pro-attitude towards the feeling

[6] Duncker (1940:400). C. D. Broad makes the analogy between hedonic tone and shades of color to illustrate the same point (1934:232). This account was advanced by Broad (1934), Schlick (1939), and Duncker (1940). For criticism, see Brandt (1959:305–6) and Perry (1967:193–4).

[7] See also Griffin (1986:8).

they give rise to. Different attitudinal accounts work with differing pro-attitudes. They include "desirable" (Sidgwick, 1907), "wanting to prolong" (Brandt, 1979), "liking," etc.[8] What is common in these views is that pleasure is a compound:

Pleasure = feeling F + attitude A towards F.

F can range over any feeling. This entails that a painful feeling towards which one has a favorable attitude counts as a pleasure (this is how one can sometimes find ordinarily painful experiences pleasant). The relevant attitude is specified by the version of the theory. Note also that on this view, both F and A have *intensity*. Attitudinal hedonists usually mean intensity of the attitude by "intensity of pleasure," though this is sometimes unclear from their writings.

The attitudinal conception of pleasure is also motivated by the following distinction. Many pleasant experiences are "ostensibly locatable bodily pleasures."[9] In many cases, however, when one has pleasure, one is pleased *that* some event, state of affairs, etc., is the case, where the state of affairs, event, etc., is described by a proposition—that is, the clause opened with the "that" is filled in with a proposition. This sense of pleasure is usually called *enjoyment* in many contemporary discussions. The reason for this is that arguably the meaning of pleasure has become too close to designate only sensory pleasures. Enjoyment is employed to avoid the misconstruction of hedonism as an exclusively sensualist doctrine—which is nowadays an implicit connotation of monistic hedonism.[10]

Attitudinal hedonism replaces the mental state of pleasure with the mental state of enjoyment in order to make clear that the basic hedonic concept refers to an at-

[8]For a reply to Ryle, see Gallie (1954). For further criticism of the Rylean theory and a case for an "episode" view of pleasure (i.e., that pleasure is more akin to such episodic mental events as feelings and less akin to dispositions), see Penelhum (1957), Quinn (1968), and Momeyer (1975). A brief overview of the variances within the attitudinal view is to be found in Feldman (1997*a*:449–54), where he calls it the "Sidgwickian view." I take the term "attitudinal" from Feldman (2002*b*), which contains an elaborate defense of this type of view. For general analyses of the concept of pleasure as a mental state, see Wright (1963), Perry (1967), McCloskey (1971), and Momeyer (1975).

[9]Gallie (1954:148).

[10]For example, according to Richard Brandt, pleasure "makes people think only of wine, women, and song" (1959:304). Enjoyment, however, is closer in meaning to what was originally meant by pleasure. For an influential discussion of the enjoyment view, see Griffin (1986:18–20); for a review of Griffin's formulation of the enjoyment view, see Sumner (2000). In passing, note that it is logically possible that the concept of (non-referential) pleasure is unrelated to the concept of (propositional) enjoyment. But normally, we use locutions like "I feel pleasure" and "I enjoy myself" or "I enjoy ϕ-ing" and "I am pleased by ϕ-ing" interchangeably. I assume that hedonists would be reluctant to treat these separately. See, for example, Goldstein (1985): he thinks that "being pleased about" refers to states of affairs, while enjoyment is always connected to (our own) actions, and it always contains an element of pleasure. Thus "I doubt if enjoyment is even a *species* of pleasure; I suspect that any pleasure can be spoken of in terms of 'enjoying' a thing. During any period in which a person experiences pleasure there is some activity or state that he can be said to be 'enjoying' " (1985:53, his emphasis).

titude toward the experience of some state of affairs, and not to some occurrent feeling. The more flexible notion of enjoyment can better express the various attitudes we have towards pleasant experiences. On final analysis, what matters for all hedonists is how the world "feels" from the inside, but how the world feels from the subject's perspective depends on the sort of attitude the subject has towards the experiencing of states of affairs.

Enjoyment-based attitudinal hedonism holds that well-being is constituted by a compound of an attitude towards the feeling the obtaining of a state of affairs gives rise to—indirectly, towards the state of affairs itself. The intensity of the feeling is not related to the intensity of the attitude towards it. This account allows more robust reference to states of affairs, connected to us *via* our pro-attitudes, but, at the same time, it includes only those attitudes that are connected to our subjective inner experience. This is to say that those states of affairs that are allowed to count as determinants of our well-being must enter our experience; we must take some "liking," "enjoying," "welcoming," etc., in their respect. Thus, attitudinal hedonism identifies pleasure (as enjoyment) with an attitude, and recognizes that our attitudes towards pleasure and pain do not change in a linear fashion with the intensity of the feeling-component of pleasure and pain. Well-being, on this interpretation, is determined not by *this* intensity, but by the intensity of the attitude, revealed through preferences. Attitudinal hedonism, therefore, prescribes the promotion of preferences when we are concerned with the promotion of welfare, with the proviso that the preferences in question must be about, or their satisfaction must result in, some pleasant feeling for us. On the other hand, monistic hedonism identifies well-being with the feeling or sensation of pleasure itself. The more intense these feelings are, the better, irrespective of people's attitudes and the intensity of these attitudes towards the corresponding experiences. What matters for this view is the intensity of the feeling itself.

Now I would like to weave some of the points I have made so far together. In Section 1.1, I distinguished between subjective and objective accounts of well-being. On this distinction, attitudinal hedonism is a subjective view. On the other hand, monistic hedonism does not preserve the connection between the sources of one's well-being and one's preferences, since it does not require that the person prefers these sources. On this account, one's attitudes towards some experience, expressed in her preferences, is not constitutive of it being pleasant. The view prescribes the promotion of pleasure, but how you view some pleasant experience is irrelevant. Therefore, monistic hedonism is an objective account of well-being.

In Section 1.2, I also argued that an account which cannot incorporate the subjectivist intuition is not a plausible account of welfare. That is, even if monistic hedonism gave a plausible interpretation of pleasure, it would be subject to the qualms one can raise about objective views. If it gave a plausible interpretation, it would be a good candidate for one of those things that are sources of welfare, but

it could still not be identified with welfare. On the other hand, even though attitudinal hedonism preserves the subjective element in its account of what promotes a person's well-being, it is still subject to the argument developed in Chapter 3.

4.3 Hedonism and Risky Prospects

Hedonism as a theory of well-being can take many different forms. There are few components which are common to all of these forms. However, the distinction between the monistic and attitudinal conceptions of pleasure gives an exhaustive and mutually exclusive classification that can be used in assessing the theory. Moreover, one component I assume is common to all versions is that an hedonist theory of well-being is able to distinguish between different *intensities* of pleasure and enjoyment—hence between the degrees to which different experiences are good for the person whose experiences they are. Indeed, by giving a straightforward way for discriminating between the values of different experiences, hedonism has an attractive feature.

Recall my discussion of the central ideas of modern utility theory from Section 2.2. In modern axiomatic and modern axiomatic expected utility theory, a utility function is a representation of a preference ordering, such that if, from x and y, x is preferred to y, then the function assigns a higher number to x. On this interpretation of utility, it is a mistake to say that you prefer some option, x, because it yields higher utility; x yields higher utility because you prefer it, and not the other way round. But utilitarians who give an hedonic interpretation to utility may claim that they do not make a mistake when they say you ought to prefer the option with higher utility. This is because they advance a substantive interpretation of utility as a representation of valuable mental states: x has higher utility for you if and only if it results in "more" (or more intense) of a valuable mental state for you. If x gives rise to more of a valuable mental state, you ought to prefer it; if you do not, you prefer what is worse for you. Is their claim compatible with modern utility theory?

It is not. In order to show this, I want now to launch an argument against the hedonist interpretation of utility, in both of its monistic and attitudinal forms. Consider the following example. Suppose you are offered a choice between, on the one hand, some experience that has 99 "units" of hedonist well-being for sure (that is, this is how pleasant this experience is going to be), and, on the other hand, a gamble of either 0 or 200 units of well-being with equal probabilities (that is, having either of two experiences, one which is not pleasant and another which is very pleasant). Mathematically, the gamble has a higher expectation of well-being from the two prospects. But which one do you prefer? If you prefer the certain prospect, you will end up with an experience of 99 units of hedonist well-being; if you prefer the lottery prospect, you may end up with an experience of 0 units of

hedonist well-being, or one with 200.[11]

Actually, you may prefer the 99 units for sure over the gamble. You are not irrational to have this order of preference: you may be *risk-averse* towards these prospects. Of course, if you have some other risk-attitude, you may have the opposite preference in the example.

One question I am *not* asking now is whether, if you are risk-averse, the 99 units is better for you in the example. You may think that since the expectation of well-being is higher in the gamble, it is better for you. Or you may think that since your attitude towards well-being should not be irrelevant, the sure outcome is better for you.

Nevertheless, this much is clear. Utility represents your preferences. In the example, depending on your attitude towards risk, you can, not irrationally, prefer either the certain prospect or the lottery prospect (or you can be indifferent). Your preference can lean either way, or none. Whichever prospect you prefer, however, has higher utility for you—by definition. But, by hypothesis, we know that the numbers represent well-being. And we know that the lottery prospect has a higher expectation of well-being. The hedonists I have been talking about want to identify well-being with valuable mental states. At the same time, they want utility to represent the "amount" of these valuable mental states, such that when you maximize utility, you maximize hedonist well-being. But utility and well-being can come apart.

Suppose our hedonist accepts the monistic conception of pleasure, or valuable mental states. Then the example goes like this: the choice is between 99 units of a valuable mental state for sure, and 0 or 200 units of this mental state at equal odds. On this view, since the mathematical expectation of the lottery prospect is higher, it is better. This is because monistic hedonism is an objective view of well-being, and your preferences over prospects are irrelevant. Thus, if you are risk-averse, the prospect that is better is different from the prospect with higher expected utility.

The above argument applies to the attitudinal conception, too. In this case, the value of a mental state is constituted by some pro-attitude towards that mental state. But, once again, that value cannot be represented by utility. In terms of my example, the values 0, 99, and 200, represent the values of the alternatives. On this view, you again know that the lottery prospect is better, but, since the view is silent on attitudes towards risk, it does not tell you which prospect is better for you.

Note that on this view, your preferences *constitute* what mental states are valuable to you. But the attitude that you need to have towards some mental state in order for it to be valuable is not to be confused with your risk-attitude towards those valuable mental states. What the former attitude constitutes is not utility, but

[11]This example comes from Broome (1991*a*:24), but he uses it in the context of the preference satisfaction account of well-being. In that context, see also Ellsberg (1954).

well-being. What the latter attitude constitutes is not well-being, but utility. The former attitude is already "incorporated" in the numbers, 0, 99, and 200. We still know nothing about your preferences between the prospects of 99 units of well-being for sure and the gamble between 0 and 200 units at equal odds. In order to know that, we have to know your risk-attitude. Your preferences between risky prospects are influenced by your risk-attitude. Therefore, utility represents these preferences, not well-being. The expectation of the amount of valuable mental states is not representable by expected utility.

Perhaps an hedonist could say that she can take attitudes towards risk into account this way. An attitude towards risk is a mental state, and it can be enlisted by the hedonist as a valuable one. The reply is that this is beside the point. Suppose I am a natural born gambler: I find risk-taking pleasant, whatever we take pleasure to be. And suppose that I am, nonetheless, risk-averse in my choices, because I find risk-seeking imprudent. Now when I choose in a conservative manner, the prospect I choose has higher utility for me. So even if risk-attitudes can be incorporated into hedonism in this fashion, utility and well-being can still come apart.

Or perhaps an hedonist could attempt to take risk-attitudes into account another way. She could say that prospects are complex objects, and the risk-attitude towards them is a constitutive part of their pleasantness. Of course, this hedonist is an attitudinal hedonist—there's no room for any attitude in the monistic conception of pleasure.

But I am not sure how a prospect can be pleasant. If I choose the gamble of 0 and 200 units of well-being at equal odds over the 99 units for sure, and you ask the reason for my choice, it would be strange for me to say that I found the *prospect* more pleasant. I should say instead that I find receiving 200 more pleasant than receiving 99, and I find receiving 99 more pleasant than receiving 0, *and*, given the odds, I am willing to take the risk. Or I could say that I find gambling more pleasant—but the attitude towards risk is not the same thing as the pleasure of gambling.

Historically, utilitarian philosophers and economists were hedonists about utility before Neumann and Morgenstern. Modern utility theory, however, is incompatible with an hedonist conception of utility. Nevertheless, utility as a representation of well-being is still sometimes interpreted as a representation of pleasure. In the remainder of this chapter, I discuss two such interpretations.

4.4 The Happiness View and the Compromise Model

Most utilitarians accept either the hedonist or the preference satisfaction account of well-being. In "Two Concepts of Utility" (1982), Richard B. Brandt discusses these two views in the context of interpretations of the concept of utility.

What Brandt calls the "desire theory" is the view that well-being is preference

(or desire) satisfaction. On this account, the more of a person's desires are satisfied, the more utility this person has, and the more utility a person has, she is on a higher level of well-being. But, according to Brandt, there are problems with this view: there are desires whose satisfaction seems not to contribute to our well-being. Moreover, some of our desires are mistaken: we think their fulfillment would make us better off, while in fact they won't. In addition, on this account, desires which one mistakenly believes have been satisfied can contribute to one's well-being. There are also other well-known controversies about this view: for instance, whether the fulfillment of a person's desires after she dies can make that person better off.

Proponents of the desire theory, therefore, may want to put some restrictions on the desires which matter for welfare. They might want to include in their view, for example, only desires that are somehow corrected, improved, or ideal. But, Brandt claims, this is not their most pressing problem. Desire theorists need to tell us at what temporal point the satisfaction of desires of the individual can be taken to promote her welfare. Since desires and the intensity of desires change, we cannot select this "date" in any non-arbitrary way. Furthermore, whatever temporal location we pick, we have to decide whether the fulfillment of desires which are no longer held with respect to that time, or whose intensity have changed, or which will only be formed in the future, should be included in determining what is good for the person. But on what grounds could we answer these questions? The answers all seem arbitrary. The desire theory is underdetermined.

In contrast, what Brandt calls the "happiness view" can cope with these problems. On this view, utility is *pleasant experience*, where

> the degree of pleasantness of an experience is fixed by the magnitude of the wanting to continue it for itself, which the experience *causes at the time*. So, on this view, an experience E at t_1 is more pleasant than E' at t_2 if and only if at t_1 E is making the person want more intensely to continue E beyond t_1 than E' at t_2 makes him want to continue the quality of E' beyond t_2. (1982:165, emphasis in original)

Brandt evidently works with an attitudinal version of hedonism. The preferences that determine what is good for the person are preferences for the continuation of pleasant experiences. If you want to promote your well-being, you have to establish which one of two experiences, x or y, will be more pleasant (will be more wanted to continue for its own sake); that is, what difference it makes to your future experiences whether x or y is chosen. For every future moment, we can thus plot your level of utility or happiness.

According to Brandt, this procedure is superior to the one available to the desire theorist. The main difference between the two sorts of view is that the desire view assigns utility values to an experience E on the ground that it is wanted, while the happiness view assigns utility values on the ground that the *continuation* of E is

what is wanted (1982:174). I think what he means is this. The problems of the desire view stem from the fact that desiring is dispositional—a desire is not an occurrent mental state you perpetually have until it is fulfilled. This is the reason for the difficulties with "dating" a desire in order to assign a utility value to its fulfillment. Pleasure, however, is occurrent: it is constituted by the attitude you have towards a *presently occurring* mental state.

When we assign utility numbers on the happiness view, we assign them to moments of pleasant experience, and we assign them according to the intensity of pleasantness—the magnitude of utility numbers corresponding to the intensity of (the attitude component of) pleasure (1982:174). In order to determine well-being over an extended period of time, we simply aggregate the utility values over time. "Whatever the practical difficulties in measurement," says Brandt, "this conception is clear" (1982:166).

The procedure goes like this. Of two alternatives, A and B, we can establish which one will promote well-being more if we measure the level of happiness each brings for every future moment. Suppose we measure level of happiness (utility) on the Y-axis, and we measure time on the X-axis:

> Let us represent these results by a broken curve, plotting the moments at which he is happier if A is done above the X-axis, the distance above the axis fixed by how much happier he is than he would have been had B been done; and similarly plotting points below the X-axis representing how much happier he is at some moments if B is done than he would have been had A been done. This operation will give us curve-segments probably both above and below the X-axis. Let us then compute the area under these curves... (1982:166)

Brandt wants to make use of the differences between the utility levels associated with the alternatives A and B, and he explicitly mentions the Neumann-Morgenstern procedure as a way to establish the respective utility values.[12] This procedure defines utility as an *ordinal* representation of a preference ordering. The utility values assigned are arbitrary, as long as their magnitude reflects the order of preference. But that is just to say that the utility differences will be arbitrary, at least under conditions of certainty. On the other hand, it is true that the cardinality property of expected utility functions ensures that the ratio between *pairs* of the *differences* between utility values of the function will be maintained. But these numbers do not represent intensity of preference, level of happiness, or level of pleasure.[13] In any case, the procedure establishes a *cardinal* utility function in ac-

[12] E.g., 1982:169. It is actually mentioned in the context of the desire theory. But his remarks suggest that he would use this procedure in the happiness theory as well.

[13] Compare a remark by Ellsberg: "much confusion probably stems from the fact that they [i.e., von Neumann and Morgenstern] are prone to write in large, clear type about comparing differences in preferences and to discard such notions in fine print at the bottom of the page" (1954:551).

cordance with the risk-attitude of the individual. If this procedure is used, one will "lose" the intensity of pleasure—the resultant utility values do not indicate anything about how much the continuation of pleasant experiences is wanted. These data, however, are crucial for Brandt's procedure. Without them, the happiness view is vacuous.

In other words, the happiness view works with *ex post* utilities. If Brandt indeed wants to utilize modern axiomatic and axiomatic expected utility theory to represent happiness, he is confusing *ex ante* and *ex post* utilities.[14] On the other hand, if he does not want to use modern utility theory for this purpose, he ignores the role of risk-attitudes in the determination of preferences over risky prospects.

Can an hedonist avoid this problem? Consider the following example. Mary goes on a safari in Africa. While she is in the bush, unbeknown to her, the stock market crashes in New York. In fact, she loses her entire fortune. Is she made worse off by the crash, before she learns the news? Many would say she is. Preference satisfaction accounts of well-being imply that she is. Hedonism, however, entails that she is not, for the news of the crash did not yet enter her experience. This seems implausible for many.

Now suppose Mary is killed on this safari by a tiger, sometime after the crash, without ever having learnt that she lost her entire fortune. Did the stock market crash make her worse off in this final phase of her life? The intuition of many people could lean the other way now. After all, the crash made no difference to her life whatsoever, since she had died before it could have any effect. Hedonism reflects this intuition, but preference satisfaction views give the counterintuitive answer this time.

I introduce with this example D. W. Haslett's "compromise model" of utility as a representation of well-being.[15] This model is intended to incorporate elements of both the preference satisfaction and the hedonist conceptions of well-being. I will argue that Haslett's account fails because it treats the notion of utility illegitimately. His attempt to "delineate utility" to make it correspond to welfare is misconceived.

The compromise model defines utility in terms of the satisfaction of "preferences for experiences."[16] Your well-being is determined by what you prefer, if, and only if, your preferences are ranked according to the expected value of the experiences which are their object, and they conform to a number of requirements. First, these objects of preferences must be *particular* experiences (tokens) and not

[14]These terms are explained on page 21 and page 24, respectively.

[15]Haslett (1990). The example is discussed on 1990:68, 70, 76, 78–79, etc. Of course, there are no tigers in Africa. But who says philosophers should be familiar with basic facts of zoology?

[16]Where the experiences do not have to be pleasures. All valuable mental states are to be included. Incidentally, Haslett claims the compromise model has a nice pedigree: he thinks it was held by Mill (1861), Sidgwick (1907), and Hare (1981), among others.

experience-types. This is the "experience requirement."[17] Second, the preferences must be *fully informed* such that any preference of x over y must be formed by perfect knowledge of x and y, and nothing else. Your personal history, conceptual apparatus, brute tastes, and the like are excluded; what decides the ranking is the desirability of the experiences themselves. For any particular experience, as opposed to experience types, the basis for preferring is "inherent in the very experiences themselves" (1990:74, italicized in original).

I think what Haslett means is this. Suppose you and I were fully informed. It follows, by hypothesis, that we would have the same preference ordering between experience tokens, but not necessarily between experience types. For instance, fully informed, we would both prefer eating *this* piece of cake to reading this book now, which we would prefer to eating *that* piece of cake. Neither of us have preferences over the experience-types of "eating a piece of cake" and "reading a book."

But there is a problem now. If fully informed agents do not have preferences over experience-types, then what is the "inherent basis for preferring" in experience tokens? If the only valuable mental state was, say, pleasure, then presumably a more pleasant token would be preferable. But Haslett admits many valuable mental states (1990:68, 72). So what makes one token of a type inherently preferable to another token of a different type? How can these tokens be compared, unless there is a hierarchy of different sorts of experiences? But that would just mean that some experience-types are preferable to others. Since this is excluded by hypothesis, what makes one particular experience inherently preferable to another particular experience is somewhat mysterious.

Haslett wants to define utility in a way that makes it correspond to well-being. What we are looking for is a one-to-one relation of well-being to utility, but this relation cannot be identity. Such a view could not explain how Mary was made worse off by the stock market crash on her safari, before she had learnt the news. So Haslett proposes, instead, to conceive of the relation in causal terms. What contributes to your welfare is whatever increases your utility. That is,

> something (an event, act, object, or state of affairs) is in one's personal welfare to the extent, and only to the extent, that it results in one's utility (or prevents one's disutility). ... [Thus, something] is *likely* to be in one's personal welfare to the extent, and only to the extent, that it is *likely* to result in one's utility (or prevent one's disutility). ... Moreover, what we mean by "likely to result in one's utility (or prevent one's disutility)" can, with this model, be made more precise by substituting for it the technical concept of "expected utility." (1990:75, emphases in original)

Mary thus was made worse off by the stock market crash because of its likely effects on her future experiences; it changed her "expected utility." When she

[17] The requirement will be discussed in detail on page 78 in Section 6.3.

dies, however, without having learnt of the crash, the crash has no likely effects on her future experiences, since she will not have future experiences. Her "expected utility" remains unaffected by the crash.

Now utility is a representation of preferences. As a representation, it cannot be caused. Contrast the notion of temperature. Temperature is a representation of warmth. If there is a heat wave, it does not cause the temperature to be higher. It causes the weather to be warmer, and that is represented by a higher temperature value. Likewise, events, acts, objects, state of affairs, etc., cannot directly cause higher utility. They can cause *something* utility is representation of. This something, in modern utility theory, is preference satisfaction. But Haslett is only interested in preferences whose objects are experiences. Only these are represented by utility. So what is caused must be the satisfaction of these. The only difference between these preferences and other preferences is that the fulfillment of these preferences is the obtaining of some mental state. So what is caused must be this mental state. But then what utility represents is this mental state—otherwise it is not possible to "delineate" the concept of utility the way it is necessary for it to represent well-being. Haslett, therefore, implicitly assumes that utility represents mental states.

Consider next what Haslett says about the possibility of interpersonal comparisons of utility on the compromise model:

> With the compromise model, but not with the preference model, the only kind of preference-satisfactions that count are those in the form of particular experiences. Only by having the experiences themselves can preferences for experiences be satisfied. Thus, with the compromise model, a comparison of preference-satisfactions comes down to a comparison of experiences. *And we already know how to make sense of comparing the experience of one person with that of another*. (1990:89, emphasis in original)

What Haslett has in mind is something akin to Mill's "competent judges" test (1861). We duplicate or represent the relevant experiences to ourselves, or ask someone who has had them, and see which one is preferred. Since we now know, in principle, how to do this test, we can say that interpersonal comparisons on the compromise view are at least meaningful. But, as I said earlier, Haslett does not allow the comparison of preferences for experiences to be based on experience types, and it is unclear how the basis for preferring could be inherent in experience tokens. So the competent judges must make their judgments based on something else.

In order to construct utility functions, the objects of the preferences do not have to be experiences in modern utility theory. Not for Haslett, though. To use his own examples (1990:81), since it makes no difference to my life whether planets in other galaxies are beautiful rather than ugly, or whether it is a sunny day on the other side of the world rather than a cloudy day, my preferences for these cannot be

assigned utility values. But I can very well tell you whether I prefer planets in other galaxies to be beautiful rather than ugly (I do), or whether I prefer a sunny day on the other side of the world rather than a cloudy day (I am indifferent). There are bases for preferences other than their experiential quality, and it is not meaningless to have preferences over objects that do not enter our experience. And, in order to construct a utility function in utility theory, all you need is this sort of data; I already provided them in saying what I prefer from these non-experiential objects. Were you interested in constructing an expected utility function, you would have to ask about my attitude towards these in risky situations. This is possible and perfectly legitimate whatever the objects of my preferences are. It is true, however, that these functions are no basis for the kind of interpersonal comparisons Haslett wants.

So Haslett must mean something else. He says preferences for experiences are only satisfied by having these experiences. And he wants to compare, not your preferences between these experiences, whether you have them or not, but something about the experiences themselves. The only thing I can think of is that he wants to compare the intensities, *vis-à-vis* one another, of these experiences. But he cannot do that. This sort of information is excluded from utility theory, even if we could indeed make sense of comparing intensities of experiences across different persons.

Hedonists must attach importance to the intensity of experiences. At least their theory would be much less attractive if it could not distinguish between severe suffering and a mild itch. But if they try using modern utility theory, they disqualify data on intensities—they exclude what is essential for their theory.

Chapter 5

Authentic Happiness

5.1 Between Pleasure and Desire

In my opinion, L. Wayne Sumner's *Welfare, Happiness, and Ethics* (1996) is the most original recent contribution to the philosophical literature on well-being. Nevertheless, as far as I can see, it has not received the attention it deserves from philosophers.

This is a pity, for Sumner's work is interesting for at least four reasons. First, it contains a detailed examination of the merits and shortcomings of the main rival conceptions of welfare—hedonism, objective theories, and desire satisfaction views. Second, it develops a new theory of well-being in terms of *authentic happiness*. The building blocks of this theory are taken from the most plausible elements of hedonism and desire satisfaction views. As Sumner says, his theory is "something in between" (2000) these two kinds of views. Third, Sumner argues that welfare is subjective (see the quote from him on page 4), and, finally, he gives a defense of welfarism. Thence, by examining his contribution, I can further my discussion of hedonism and desire satisfaction theories of well-being, and by highlighting the problems of his theory, I can explain why I think the commitment to subjectivism about welfare is problematic. Thus, I will concentrate on the first three elements of his contribution. I will first discuss the shortcomings of hedonist and desire satisfaction views which prompt Sumner to develop his own view; then I present this view, followed by my critical remarks. My focus will be on showing that the problems for Sumner's authentic happiness view stem from his commitment to subjectivism.

Sumner's rejection of desire satisfaction theories of well-being starts from considering two features of wants and desires: that they are *intentional* and that they are *prospective*. Desires are intentional because a desire is always a desire *for* something—some object, activity, or state of affairs. However, all desires can be expressed as desires for states of affairs—the getting or having the object I desire, the doing of the activity I desire—and states of affairs can be described with propositions. Therefore, to desire a state of affairs is to desire the proposition describing that state of affairs to be true.

The intentionality of desire implies that a person's desire can be satisfied without the person ever getting to know that her desire is satisfied. Thus, on a desire satisfaction theory, the person can be made better or worse off by the satisfaction or dissatisfaction of her desire, without her knowing that her desire has been fulfilled

or frustrated. In order to avoid this implication, many philosophers are attracted to imposing the *experience requirement* on the desires whose satisfaction can contribute to a person's well-being: according to this requirement, only those desires are relevant to well-being whose satisfaction enters into, or makes a difference to, the experience of the person. Sumner believes that all desire satisfaction theories of well-being must incorporate this requirement.[1]

But if desire satisfaction theories incorporate the experience requirement, they run into difficulties with the other feature of desire. A desire is prospective, that is, it is always a desire for some future state of affairs. Thus, it is always possible that your *ex ante* expectation of how good it will be for you to satisfy some desire is misguided and, *ex post*, you find that the satisfaction of your desire has not made you better off. (Conversely, you may not desire some state of affairs, but nevertheless be pleasantly surprised or satisfied when it obtains.) In order to avoid this problem, many philosophers are attracted to the further requirement that the relevant desires be *informed*: formed in awareness of all relevant facts and without committing any cognitive error. Sumner, however, argues that this move does not save desire satisfaction theories. Even if your desire is informed in the sense required—and different versions of the desire satisfaction view construe the information requirement differently—it reflects your *present* belief about how good the satisfaction of your desire will be for you when it is satisfied. But there can always be a gap between your *ex ante* expectation and your *ex post* experience. In order to eliminate this possibility, the information requirement must be constructed in a way that it can exclude those desires whose satisfaction will be disappointing or unrewarding. You must know how good the satisfaction of your desire would be for you. But, Sumner claims, this

> would be tantamount to conceding that what matters, so far as our well-being is concerned, is *our* satisfaction and not merely the satisfaction of our desires. If an information requirement has any genuine work to do within a desire theory, therefore, it will be inconsistent with the basic rationale of the theory. (1996:132, his emphasis)[2]

In sum, desire satisfaction is neither logically necessary nor sufficient for well-being. It is not sufficient because you may find that something you desire does not turn out to be good for you. Nor is it necessary because something you do not

[1]See Sumner (1996:128). The experience requirement was introduced by James Griffin (1986:13). Although he attributes it to Jonathan Glover (1977:63–4), I could not find the reason for that attribution. I further discuss the requirement on page 78 in Section 6.3.

[2]The "basic rationale" of a desire satisfaction theory is that it appeals to states of the world, and not only to states of the mind, in determining what makes your life go well. If an information requirement was added to a desire satisfaction theory, what would matter is not what antecedent desires you have, but what your experience is like when these desires are satisfied—when the desired states of affairs obtain. Thus, antecedent desires "drop out" from the picture. See also Sobel (1998).

desire may turn out to be good for you. If we incorporate some sort of information requirement to avoid these problems, our theory effectively ceases to be a desire satisfaction theory of well-being—it looks rather like an hedonist theory.

Consider hedonism, then. Classical hedonism identifies welfare with happiness, and gives an account of happiness in terms of pleasure. It interprets pleasure as a mental state, with the sensational model of physical pleasure and pain. But if pleasure is understood as a sensation, it is too narrow for the purposes of an hedonist theory. Relying heavily on results of research into the psychology of pleasure and pain, Sumner (1996:98–112) opts for an attitudinal conception instead. On this model, pleasure and pain are not pure sensations; they involve an attitudinal dimension—how pleasant or painful an experience is depends on the subject's reaction. In order to emphasize the shift, we ought to talk, instead of pleasure and pain, of enjoyment and suffering. The shift also incorporates the ideas that to be pleased or displeased is to be pleased or displeased *about* something, and that both can take various forms of attitude. None of these points, Sumner argues convincingly, have been adequately coalesced by classical hedonism. Enjoyment and suffering, furthermore, have the advantage of being "compounds," or composites of a state of mind and some state of the world. In Sumner's formulation:

> Construed extensionally … enjoyment and suffering are no longer merely mental states; instead, they are complexes consisting of mental states plus their objects … (1996:111).

The sensational model of pleasure and pain, held by classical hedonists, suffers from the fact that pleasure and pain do not seem to be homogeneous mental states. The attitudinal model of enjoyment and suffering, held by many modern hedonists, fails however to adequately distinguish between enjoyment (or suffering) from veridical and from illusionary experiences. (I have explored these arguments in Section 4.2 and Chapter 3, respectively.) Perhaps a better starting point is the concept of happiness.

Sumner distinguishes four different senses in which the concept of happiness is used (1996:143–7). If you are *happy about* or *with* something, then you have a favorable attitude towards something being the case. The attitude in question can range from mild contentment to euphoria, but it is also possible that your attitude is cognitive. You can be happy about something if you approve of it, even without some occurrent positive feeling. In contrast, you can also *feel happy*, in which case you do have some occurrent feeling, but your happiness does not have an intentional object. In this sense, happiness is a frame of mind, a general sense of optimism or joy. If you often experience phases of such happiness—if you often feel happy—then you have a *happy disposition* or *personality*. While all these instances of happiness are ingredients of a life that goes well, well-being is connected with the fourth sense of the concept: that of *having a happy life* or,

simply, *being happy*.

Having a happy life has to do with how you view the way your life is going, or the way it has been. It has both an affective and a cognitive component: you are happy in this sense if you affirm or endorse the conditions of your life, if you think your life is worth its while to live, or it has been worth its while to live, according to your own standards and expectations. This evaluation may be global, embracing your life as a whole, or it may concern some particular aspect of it: your career, personal life, and so on. Having a happy life does not require that you have a happy disposition, nor that you often feel elevated or happy in the occurrent sense. Neither does it require that you are happy about things beyond those that touch directly upon your life. You can be happy even if you think that the world, or some aspect of it which is more or less independent of you, is going badly—and *vice versa*.[3]

Hedonism defines well-being as pleasure, and some hedonist theories reduce happiness to pleasure. Sumner's theory gives up the sensational (or monistic) conception of pleasure in favor of enjoyment (a type of attitudinal conception). If well-being consists in happiness, the obvious strategy would be to reduce happiness to enjoyment. But Sumner thinks this is not possible. The first sense of happiness, being happy about something, is clearly not identical to enjoyment: you can be happy about many things that you do not enjoy. Neither can happiness in the occurrent sense (feeling happy) be identified with enjoyment: while enjoyment and suffering are intentional, the occurrent sense of happiness is non-referential. As happiness in the sense of having a happy disposition reduces to the occurrent sense of happiness, enjoyment cannot be identified with it either.

That leaves happiness in the sense in which it refers to a global evaluation of some or all aspects of your life as the only candidate. But Sumner denies that it can be reduced to enjoyment in a direct way:

> Enjoyment and suffering are still too episodic, too tied to experiences of specific activities or conditions, to be identifiable with happiness and unhappiness … Like pleasures and pains, enjoyments and sufferings are typical sources of happiness and unhappiness. But they are not the only such sources … Hedonism, even the improved version which takes enjoyment and suffering as its central notions, … confuses an important source of happiness with its nature. (1996:148)

Denying that happiness is reducible to enjoyment or pleasure, as far as I can judge, is a novel idea in the hedonist literature.[4] If Sumner can give an alternative analysis, he opens up the logical space for an original theory.

[3] See also Sumner (1992*b*).

[4] But see also Nozick (1989).

5.2 Sumner's View

In order to connect welfare and happiness, Sumner must find a strategy different from reducing happiness to enjoyment. His strategy is to propose a *non-reductive* account of happiness. Noticing that desire satisfaction views suffer from the problem that there is a logical gap between the satisfaction of your desires and *your* satisfaction, he proposes that a non-reductive conception of happiness can be constructed in terms of this sense of satisfaction, captured by the notion of *personal* or *life satisfaction*.

The concept of life satisfaction provides the basis of a theory that is "in between" hedonism and desire satisfaction views. Life satisfaction concerns how you experience the conditions and circumstances of your life, but it appeals to more than your subjective experience—since it also appeals to these very conditions and circumstances. Hedonism provides the component that well-being must be a matter of experience; desire satisfaction views provide the component that the veridical conditions of your life also matter. These components are amalgamated into the non-reductive conception of happiness Sumner is after.

The concept of life satisfaction is also used in quality of life and social indicators research.[5] These research directions measure people's subjective evaluations of the external circumstances of their lives with the aid of surveys and questionnaires. In Section 2.1, I explained that a theory of welfare must not only be able to specify in virtue of what something is good for us, but it must also be operationalizable. Happiness as life satisfaction already has an operational counterpart in the social sciences; moreover, this counterpart also provides the missing components for happiness as life satisfaction to count as a theory of well-being.

Your happiness or life satisfaction is determined by your subjective evaluations. Your evaluation, however, can be unreliable in several ways. In order to accept it as a reliable indication of how well your life is going, it must be *relevant*, *sincere*, and *considered*. It must be relevant in the sense that you must understand that you are being asked about your welfare—and not, for example, about whether your life is successful in terms of some ethical standard, or whether it conforms to some aesthetic ideal, or whether it is a life that is appropriate to the kind of life it is in some perfectionist sense. Of course, insofar as you make some ethical, aesthetic, or perfectionist ideal central to your life, it will be central to your happiness; but your happiness is not evaluated in terms of these other standards. Even though they can influence your happiness, they are conceptually different.

Your assessment of your happiness can be taken to be authoritative only if there are no grounds to doubt that it is sincere. If there is reason to believe that you

[5] For surveys of these research areas, see, for example, Neufville (1975:40–56), Carley (1981:1–21), and Zapf (2000).

understate or overstate your satisfaction with your life, your assessment cannot be taken at face value. Similarly, if there is reason to believe that your assessment is not considered—that it is influenced by passing moods or by your not having given enough attention to the subject—there is reason to question the authenticity of your evaluation.

Even if all of these conditions are met, your evaluation may still be *underinformed*. Suppose you lived, for years, in a relationship that you thought was faithful and dedicated by both parties. Now you realize that your partner was faithless on many occasions and their dedication and commitment were all false. What shall we say about your happiness during this period? That you were happy in this relationship? Or that you were not happy, since you did not have all the information which is relevant to the assessment of your situation?

In order to avoid the problem of lack of information, we may be tempted to require that you can only be happy if you have all the relevant information, or at least enough information to justify your evaluation. But Sumner rejects both of these alternatives. As far as your happiness in this relationship is concerned, facts do not, retrospectively, change how happy you were. So more information only matters if we are interested in whether your life was *worse* in the period you were with the deceptive partner. In this sort of evaluation, you assess the *importance* of your happiness (in that period) to your well-being. Since Sumner thinks that well-being is thoroughly subjective, he thinks that both a full-information and a justifiability requirement would unduly restrict your authority over your well-being. Instead, when more information is relevant is a question left to your own jurisprudence. Being more informed is relevant to well-being, but the extent to which it influences how well your life goes is up to your own assessment. For instance, if you think that your partner's deception blighted and betrayed your relationship, you may judge that although you were happy, your life did not go well. Or you may accept the facts now without thinking that your life was worse because of them—after all, you were happy for years, and there is no sense in denying the importance of happiness past. In summary, more information is relevant if and only if it influences your evaluation (1996:157–61).

Consequently, the relation of happiness as life satisfaction and well-being is not a one-to-one matter. Happiness can be identified with well-being if and only if it is *authentic*. Authenticity, on the one hand, requires that your happiness is based on an informed evaluation—informed by your own defensible and defeasible standards. In addition, authenticity also requires that your evaluation is *autonomous*.

Your assessment of how well your life is going, or how well it has gone, must reflect its autonomous endorsement. If your happiness is based on manipulation or socially conditioned preferences, it cannot be authentic. To take a well-known example, consider the subdued and battered housewife who adapts her expectations and satisfactions to her situation and opportunities. Intuitively, it seems that even

though she sincerely says and feels that she is happy, her life is not going well for her.

This problem has motivated a number of philosophers to develop so-called *hybrid views* of welfare. These views take endorsement as a necessary condition of well-being: nothing can contribute to your well-being unless it is endorsed by you. But these views also stipulate that what you endorse must be independently valuable in order for it to promote your well-being. As he thinks welfare is thoroughly subjective, Sumner rejects such accounts. For him, there is no external viewpoint from which it is possible to establish how well your life is going for you.

Once again, Sumner goes for the option of subjective evaluation of autonomy. Having been liberated from external conditioning influences, it is up to you to decide how you evaluate your welfare in the period when you were not autonomous. For instance, the subdued housewife may completely discount her years spent under manipulation; she may think her life did not go well, that those years were wasted. But she may equally think that that period was part of her life, and there is no point of denying that she was happy. So she may choose not to lower her welfare assessment. Between these two extremes, of course, she may weigh the importance of negative external influences on her autonomy for her well-being any way she sees fit. As long as her evaluation is defeasible—we have no reason to doubt its reliability—it must be taken as the authoritative determination of her well-being (1996:161–71).

In the literature, there are two influential accounts of autonomy. The first holds that a person is autonomous if and only if her desires and preferences are, or would be, affirmed by her in her higher-order preferences. This is known as the *hierarchical* account of autonomy. The second view holds that a person is autonomous if and only if the causal history of her desires and preferences—their formation—is free of manipulative influences. This account is known as the *historical* account of autonomy. Sumner stipulates that even though neither is completely satisfactory, both may be needed for explaining when a person can be considered autonomous. Therefore, the theory which emerges identifies well-being with authentic happiness. Happiness is authentic if and only if your evaluation of your happiness is both informed and autonomous in a defeasible manner. Your evaluation is autonomous if and only if it is based on desires or preferences which count as autonomous both on the hierarchical and the historical accounts of autonomy.

5.3 Problems with the Commitment to Subjectivism

Sumner's theory of authentic happiness is a thoroughly subjective conception of well-being. It is subjective both in the sense that on this theory, what promotes a person's well-being must be connected to the person's attitudes, and in the sense that the appropriate evaluation of how good a person's life is for that person must

be the person's own evaluation. There is no external evaluative standard (1996:159, 161, 164).

This commitment to subjectivism, however, causes several problems. Consider first the conditions involved in authenticity. The role and importance of information and autonomy for a person's well-being are left to the person's own jurisprudence. More information breaks the connection between happiness and welfare only if it changes the way you evaluate your circumstances; emancipation from a non-autonomous condition breaks the connection between happiness and welfare only if you judge that being autonomous makes a difference to how well your life goes. As long as your evaluation is *defeasible*, it must be taken as authoritative.[6]

But what are the conditions for defeasibility? Assume that your evaluation is relevant, sincere, and considered. It is also defeasible, unless we have reason to think that it would change if you had more information available. But whether more information would be relevant and would make a difference to your evaluation is ultimately a judgment left to you. This threatens to be circular, however, because whether or not more information would make a difference depends—well, it depends on whether or not it would make a difference. Assume I ask you a series of questions about how your evaluation of your life would change if conditions were different. If some of these conditions are indeed different, and your knowing that they are different would make a difference to your evaluation, your present evaluation is not defeasible. Now suppose some important conditions of your life are indeed different from what you believe, but even if you were aware that they are not the way you thought they were, your evaluation would not change. The disturbing possibility remains that your evaluation should change, but it would not. You sincerely insist that your life satisfaction would be unaffected.

Such cases are not uncommon at all. Think of a depressed person, who believes her life is a failure, that nobody likes her, that she cannot achieve her goals. You point out that she has impressive achievements, she is popular and has many true friends, that by all accounts she is a successful person. These facts won't change her evaluation, yet we know her life is not as bad as she thinks it is. An even more common case in point is a person who is a pessimist: she has no reason to have such a bleak outlook on life, but your attempts to change her evaluation will be unheeded. Yet you are not entirely unjustified to take her evaluation indefeasible. Or consider the well-known phenomenon that people may adopt their aspiration levels to their situation: they tend to choose goals for which they have the resources and lower their expectations in adverse circumstances. These will influence their evaluations, even though in some cases there is no reason for that.

[6]This is the term Sumner uses. That evaluations are defeasible means that "they are authoritative unless we have some reason to think that they do not reflect the individual's own deepest priorities" (1996:161).

In replying to these cases, Sumner cannot appeal to the distorting effects of depression or pessimism on these persons' evaluations, since then he would appeal to some non-subjective standard. Nevertheless, the evaluations of these persons are indefeasible even if they have all the relevant information because they don't seem to be able to "appreciate" or take into account the relevant facts. But Sumner's commitment to subjectivism entails that the evaluations of people with such deficiencies cannot be second-guessed.

Part of the problem is that although Sumner recognizes that a person's *report* of her evaluation of how satisfied she is with her life may be unreliable in several ways, he does not seem to take seriously the possibility that the *evaluation* itself may be suspect—even if the person's report is reliable.[7] There are no real constraints on the validity of evaluations, since the ultimate appeal can only be to the person's own priorities.

This suggests that perhaps the connection of life satisfaction to well-being is looser than Sumner thinks. A fancy example for illustrating this point is Blondlot's story of the N-rays he claimed to have discovered, but which were not a genuine natural phenomenon (see page 38). How could Blondlot evaluate his life in the period of his mistaken discovery in light of the fact that there are no N-rays? I presume that he was happy during that period—he found personal satisfaction in his work, and he believed success in his scientific career was an important ingredient of his well-being. He could say that the fact that his work was futile means that even though he was happy, his life did not go as well as he thought. It would have been better for him if his efforts had yielded results. Or he could say that the fact that his work was futile makes no difference whatsoever to his well-being; all that matters is that he was happy doing his research, that he found satisfaction in his work. By his own lights, his life was just as good as he thought it was.

In my view, whatever Blondlot himself thought about the relevance of the fact that N-rays were a pseudo-phenomenon should *not* influence the judgment about how well his life was going. The extent to which the truth actually changed his evaluation of his life satisfaction was certainly up to him—but the way it should have changed his evaluation of the relevance of his life satisfaction to his well-

[7] Psychological research has shown that there are many problems with the validity of such evaluations. For instance, the same event may increase or decrease reported satisfaction or happiness; subjects make implicit comparisons, and their reports are influenced by transient moods, as well as by features of the research design: the presence of a handicapped person can increase reported subjective well-being, as well as an interviewer of the opposite sex. For overviews, see Schwarz and Strack (1999), and Kahn and Juster (2002). In a famous experiment, Strack *et al.* (1988) surveyed college students about their dating life and happiness. They found that if the question about the students' general happiness preceded the question about the number of their dates in the previous month, the correlation between happiness and dating was very weak. But if the questions were asked in the reverse order, the correlation was significantly increased. It seems that the question about dating prompted the respondents to include different information in the second case.

being was not up to him. Blondlot actually continued to claim that N-rays were a genuine natural phenomenon long after they were discarded from science. For the sake of the argument, suppose that he also continued to be personally satisfied (which is actually not true: later in his life he experienced depression reportedly not unrelated to the fate of N-rays). But his project was a failure. On Sumner's view, for Blondlot's well-being only *his* satisfaction was relevant, and not the satisfaction of his desires. Since Blondlot made no discovery, his desire to discover a new form of radiation was not fulfilled, although he was personally satisfied. But his personal satisfaction had no basis. Since it had no basis, his life at that period was worse. Blondlot himself, given a choice between a genuine and a bogus discovery, would have surely chosen to make a genuine discovery. And he would have chosen it on the basis that it was better for him. If he would not have chosen this, then there must have been something he did not quite understand about his own desires and the importance of his scientific pursuits to his well-being.

But if this is so, it is ultimately not his own satisfaction with his life which determines how well his life goes. Your life may go well without your being satisfied with it. And your life may go badly with your being very satisfied with it. The possibility remains that what determines how well your life goes is not *your* satisfaction, but the satisfaction of your desires.

Your evaluation of your happiness is also indefeasible if there is reason to think that it is not autonomous. But the judgment about the role and the importance of happiness based on non-autonomous desires in your well-being is also up to you. Not surprisingly, a similar argument applies here. Suppose your desires are manipulated. The more thoroughgoing the manipulation has been, the less you are going to be able to recognize it—the mark of successful manipulation or indoctrination is that you do not think that your desires or preferences have been subtly altered by outside factors. If you come to expect very little from life, for instance, due to the way you were brought up, and you are told that were you to re-examine and change your preferences you could aspire to more than you are content with now, the more thoroughgoing your indoctrination has been, the more likely you are to reject that this is possible. You will insist that your assessment of your opportunities and the worth of your pursuits is correct. So, paradoxically, the more indoctrinated you are, the less likely it is that you are able to discount your non-autonomous preferences, and the more likely it is that you will insist that your happiness is authentic. This gets things the wrong way around, since it is hard to see how you could be convinced that your evaluation of your happiness is indefeasible, which it certainly seems to be, without, once again, appealing to some external, non-subjective standard. It appears that sometimes we cannot avoid appealing to such external standards to evaluate how well a person's life is going.

Notice that neither the hierarchical nor the historical accounts of autonomy would help Sumner here. Suppose your evaluation of your life satisfaction is the

causal result of manipulation. Your evaluation may still be defeasible if, having been freed from manipulative influences, you still embrace your evaluation in some of your higher-order evaluation. But whether *that* evaluation is defeasible remains an open question. It must be more informed than your original evaluation (if nothing more, you must know that your original evaluation was a result of a manipulative process). But then we are back to the previous problem: certain pieces of more relevant information sometimes should make a difference to evaluation, regardless of the person's propensity to allow them to make a difference.

In summary, Sumner holds that well-being is authentic happiness, but neither of the conditions he gives for authenticity can be entirely subjective. His commitment to subjectivism about welfare is deeply problematic.

There are two more points I would like to make. The first concerns the operationalizability of Sumner's theory. How would we evaluate social states and design policies for the promotion of well-being on this theory? A "Sumnerian social choice theorist" would evaluate social states with reference to authentic happiness, and when she is concerned with promoting well-being, she would propose the promotion of authentic happiness. But for evaluation and policy design, she would not use utility theory, since utility represents preferences and not authentic happiness. Presumably, she would use quality of life indices or social indicators of life satisfaction (1996:149–56).

These methods of establishing social welfare have a number of implicit presumptions. First, they presuppose that individual welfare is subjective. Second, they assume that it is related to happiness, and, third, that happiness is a matter of how satisfied people are with their lives. Finally, they assume that the evaluations people give are reliable indicators of how satisfied they are with their lives, and, ultimately, valid indicators of their welfare (1996:153).

By now, we have reason to raise doubts about all of these presumptions. Welfare cannot be entirely subjective, even if it is related to happiness the way Sumner thinks it is; and authentic happiness is not entirely a matter of how satisfied people are with their lives. Moreover, people are not necessarily best situated to assess their welfare: their evaluation may be misinformed or non-autonomous. That is, what a Sumnerian social choice theorist reads from her survey results cannot be social welfare—it may, at best, be a report on how happy people think they are.

This is not to say that it is not useful to know this. We may even have reasons to promote people's happiness, but when we attempt to do so, we are promoting happiness, and not welfare. Sumner's theory ultimately gives us little help when our goal is to promote social welfare. And even in the cases when our goal is to promote authentic happiness, we run into difficulties. When he discusses authenticity, all of Sumner's examples are retrospective. In light of new information or freed from manipulation, the individual re-assesses her happiness at an earlier point of her life. If she is *currently* lacking information or she is under manipulative in-

fluences, she cannot establish the link between her happiness and her well-being. We cannot take her current assessment as authoritative until she gets into better epistemic conditions. We could, of course, try and second-guess how she would, counterfactually, evaluate her happiness. But Sumner explicitly and repeatedly forbids this on the ground that this would be paternalistic.

When we make collective decisions, we are interested in the future. We try to find out which alternative we ought to choose. This exercise is prospective. Sumner's theory, however, is ill-suited for prospective evaluation: often, we would not be able to establish what sort of social welfare level we would end up at. Even if your happiness is authentic, how can we know that it will remain so given the choice of this or that alternative? The problem of the validity of individual evaluations of life satisfaction exacerbates as we try to predict authentic happiness in the future.

My final point is about what to do if we reject Sumner's theory. Where do we go from here? Sumner selected the most plausible elements of desire satisfaction and hedonist theories of well-being to build his theory. But this theory cannot meet several objections, and these objections target his commitment to subjectivism about welfare. Does this mean that we should proceed to an objective conception of well-being? I think not! In Section 5.1, I discussed Sumner's objections to hedonism and desire satisfaction theories. I agree with his rejection of hedonism. But he also rejects desire satisfaction theories of well-being by showing that the features of desires—that they are intentional and prospective—open up a logical gap between your *ex ante* expectation of the satisfaction of a desire and your *ex post* experience of the satisfaction of that desire. This argument, however, is based on the premise that desire satisfaction theories must incorporate the experience requirement. I suggest that this premise should be dropped. As I argue in Section 6.3, desire satisfaction theories do not need the experience requirement. If it is rejected, an information requirement can close the logical gap without becoming inconsistent with the "basic rationale" of these theories.

There are, however, many other issues that must be sorted out for desire or preference satisfaction theories of well-being. These are also discussed in Chapter 6.

Chapter 6

That Obscure Object of Desire

6.1 The Concept of Desire

According to the current orthodoxy in discussions in philosophy, all conceptions of well-being belong to one of three groups of theories: hedonism, desire or preference satisfaction accounts, and objective views. The subject of this chapter is desire and preference satisfaction theories.

Many different versions of this type of theory are possible. We can classify these versions by relying on three distinctions. The most familiar distinction is whether a desire or preference satisfaction theory holds that *all* of the desires or preferences of a person are relevant to her well-being, or it holds that desires or preferences must somehow be filtered—that only the satisfaction of a subset of the person's desires or preferences is relevant to that person's well-being. The former is often called the unrestricted or *actual* desire or preference satisfaction theory. ("Actual" is taken here in the modal, and not the temporal, sense.) There is wide agreement that this type of theory is implausible, because people can be mistaken about what is good for them. You desire the liquid in the glass in front of you, because you think it is gin and tonic, whereas it is petrol. Clearly, satisfying this desire will not make you better off (Williams, 1980:102). For this reason, most versions of the desire or preference satisfaction theory of well-being specify a subset of (actual or possible) desires or preferences, and argue that only the satisfaction of desires or preferences belonging to this subset constitutes what is good for the person. These versions are discussed in Section 6.2.

The second distinction is more infrequently mentioned in the literature, even though it gives another and not less important reason for specifying a subset of actual or possible desires or preferences which is relevant to well-being. It is based on the fact that many of our desires or preferences are not about what would be good for us. We all have many desires whose satisfaction does not make us better off or even makes us worse off. Otherwise, self-sacrifice or doing our moral duty would be impossible. And we all have many other *other-regarding* desires—desires which concern the welfare of other people. If you desire what is good for me, the satisfaction of your desire is irrelevant to *your* well-being.[1] A desire or preference satisfaction theory of well-being needs to be able to separate *self-regarding* desires

[1]Unless, and to the extent that, promoting my well-being promotes your well-being too, because, for instance, you care about me. But this is not particularly a conceptual problem. On self-sacrifice, see Overvold (1980).

or preferences from other kinds of desires or preferences. Self-regarding desires or preferences are those which concern only the well-being of the person, and concern others only to the extent that their faring well (or badly) influences the well-being of that person.

This distinction implies that an actual desire or preference satisfaction theory cannot be true, since many of our desires or preferences are not self-regarding. Thus, even the actual desire or preference satisfaction theory must be interpreted only on a subset of the person's desires or preferences—her self-regarding desires or preferences.[2] Other theories specify a subset of (actual or possible) self-regarding desires or preferences, and argue that only the satisfaction of desires or preferences belonging to this subset constitutes what is good for the person. Self-regarding desires and preferences are discussed in Section 6.3.

The third distinction I use is virtually unmentioned in the literature. I have so far used the clumsy expression, "desire or preference satisfaction theory of well-being." Desire and preference are almost always treated by philosophers as if they were the same thing. But they are not: desire and preference are distinct concepts. Whether it is cast in terms of desire or preference makes a difference to your theory of well-being. Since this distinction is the least familiar, I begin by arguing for it.

Consider the concept of desire first. One account of desire is that desires are mental states that are similar to feelings or sensations. On this view, roughly, you have a desire if and only if you feel attracted to the object of your desire. That is, desires have a certain phenomenological content. Call this the *strong phenomenological conception* of desire. The examples that may come to mind are urges like hunger, thirst and sexual desire. The problem with this view is that it cannot explain the fact that desires have, not only phenomenological, but propositional content as well. When you have a desire, you typically have more than a feeling: if you desire that p, the clause after the "that" is filled in with a sentence. But if desire is analogous to sensations or feelings, its propositional content cannot be given an account of. Thus, one must move to a *weak phenomenological conception* of desire: the view that desires are like sensations or feelings—in that they have phenomenological content—but they are also unlike sensations in that they also have propositional content. This view, of course, needs to explain how it is that desires are different from feelings in this respect.

Another problem is that desires are *conative* or *appetitive* mental states, while sensations and feelings are not. Hence a weak phenomenological conception must also give an account of the conative feature of desire. Even though in ordinary language desire is often used in the sense given by the phenomenological conception, the prevalent *philosophical* conception of desire today is different.

[2]Perhaps the egoist is someone whose subset of self-regarding desires or preferences is congruent with her set of desires or preferences.

The main problem with phenomenological conceptions is that many desires do not seem to have phenomenological content at all. Because of this difficulty, most philosophers reject that desires necessarily have phenomenological content. Rather, on the currently popular *dispositional conception*, desire is a mental state identified by its functional role in producing action. Desire is a dispositional state, or, more precisely, a dispositional mental state that grounds other dispositions. Thus, if you desire to ϕ, then you have the tendency to ϕ, you tend to be pleased about the result of ϕ-ing or disappointed if that result does not obtain, and so on. This analysis makes it possible to have a desire without an occurrent phenomenological state, even though from time to time desires may manifest themselves in such phenomenological states. That is, this conception does not deny that desires can have phenomenological content, but it does not entail that desires necessarily have such content. The dispositional conception is also able to give an account of the propositional content of desire—it is determined by the functional role desire plays. If a desire motivates you to act to bring about a state of affairs, you are motivated to make the proposition describing that state of affairs true.[3]

There are many controversies surrounding the concept of desire and its role in motivation and practical reasoning. These need not concern us here. I am interested in desire satisfaction theories of well-being, and have independent reasons to reject them. These theories typically use the dispositional conception of desire: they accept that desire is a disposition to act. At any one time, a person is motivated to act in different ways. Coupled up with her beliefs, her desires form different motivational states. A motivational state is a compound of a desire and relevant beliefs. On this picture, the person will act on her strongest motivational state—as it is often put, on the basis of her strongest motivational reason.

What is it for a motivational state to have strength? It must have something to do with the strength of the desire, since—at least on the currently most influential Humean theory of motivation—beliefs are motivationally inert. But what is it for a desire to have strength?

One alternative is to say that the strength of a desire is the intensity of the feeling its satisfaction gives rise to. But this is unacceptable, since the satisfaction of many desires do not result in any feeling. This would imply that such desires do not have any strength, and a desire without strength must be motivationally inert, which entails that these desires do not exist! Another alternative is to say that the

[3]See, for more detail, Smith (1987, 1994:92–129). For some implications of this conception of desire, see Smythe (1972). A fashionable way of defining desire nowadays is to contrast it with belief in terms of *direction of fit*. A belief is any mental state with a *mind-to-world* direction of fit: if the world does not fit the content of the belief, the belief must be changed. A desire, in contrast, has a *world-to-mind* direction of fit: if the world does not match the content of a desire, the world must be changed, since desires seek the realization of states of affairs. See, e.g, Smith (1988:250–1). For arguments against this distinction, see Sobel and Copp (2001).

strength of a desire has something to do with its phenomenological character. For instance, you can be mildly or very thirsty. But this proposal is also unacceptable, since on the prevalent, dispositional conception of desire, the strength of a desire is not its felt intensity, since desires do not necessarily have any felt quality. On this view, these desires would again be motivationally inert, hence nonexistent. This suggests that we have to look in the neighborhood of motivation for the strength of desire.

Perhaps the strength of desire is the "pull" of the motivation the desire constitutes. But we have to be careful with this metaphor; it is highly misleading. If "pull" is understood in some phenomenological sense, we are back to the previous, implausible proposals. The "pull" in question must thus be understood in some purely dispositional sense: the potency of some disposition to act. On this picture, the strength of a motivational state is identified by the potency of its tendency to produce action. Thus, strength of desire is the potency the desire contributes to the motivational state to produce action. But that potency cannot be identified by introspection; it must be specified by the relations of motivational states *vis-à-vis* one another. Thus, the notion of strength of desire cannot be made sense of unless it is *comparative*.

However, we do already have such a comparative concept: it is *preference*. I suspect that desire satisfaction theorists of well-being actually very often mean "preference," when they say "desire." Some philosophers are quite explicit about this. For example, James Griffin (1986:14–5) suggests that strength of desire is the rank of the desire in a preference ordering. But if what it takes a desire to have strength is to have a place in a preference ordering, then desire is a superfluous concept in these theories. It should be replaced with preference.

On reflection, it is not surprising that we should reformulate desire satisfaction theories of well-being in terms of preference. Desire is a monadic notion: it involves only one object, whereas a preference involves two objects, and tells us about the relative importance the person assigns to these objects—her priorities between them. Preferences enable us to construct utility functions, while desires do not. While perhaps we can infer desires from preferences (given the dispositional conception of desire), we definitely cannot infer preferences from desires.

Desires are non-compared, unordered, and there is no constraint on them: you can have conflicting desires, and you can desire contradictory states of affairs. Hence, desire satisfaction theories of welfare, given that people can have inconsistent desires, can yield paradoxical results in determining what is good for a person. Thus, there is no coherent desire satisfaction theory of well-being. But this does not entail that there is no plausible preference satisfaction theory of well-being. Preferences are compared, ordered, and there are coherence requirements imposed on them. Whether you cast your theory in terms of desires or preferences makes a big difference. In determining what is good for you, you need to weigh, balance,

and rank your desires, hence your theory must be a preference satisfaction theory. It must be comparative.[4]

"Preference" here is used in the sense familiar from economics: it is a disposition to choose. Preference is tied to behavior. On this approach, we have access to preferences through choices (and perhaps verbal reports of what the person would choose). Now some philosophers take a non-behaviorist approach to preference. They think preferences cannot be unequivocally inferred from choices, since it is possible to choose what you do not prefer—it is possible to make "counterpreferential choices." As far as I can see, they want to allow for these in order to maintain the possibility that a person can choose what she does not "really want." But the behaviorist approach is entirely compatible with this possibility. It makes no commitment to any psychological conception of why people choose as they do; neither does it make any commitment to what the *reasons* may be for their choices. The non-behaviorist approach has no advantage in these respects, but it does introduce a mysterious new entity into our stock of psychological concepts. As I see it, there is no advantage whatsoever to accepting a non-behaviorist approach.[5]

The behaviorist approach also has a tremendous advantage: it is connected to decision theory and utility theory. These theories impose coherence constraints on preferences. Thus, a preference satisfaction theory of well-being has some formal constraints to start with for determining which preferences are relevant to welfare. Of course, these constraints might be controversial, but some such constraints are necessary. As Allan Gibbard says:

> Decision theory makes claims about the abstract form of a rational complex of preferences, degrees of belief, and dispositions to action. It consists … in norms governing preferences, beliefs, and actions, but these norms leave the substantive questions unsettled. They don't say what intrinsic preferences to have—though they do tell us what abstract form our intrinsic preferences are to have. They don't say what to believe—though they do tell us what abstract form our degrees of belief are to have. (1998:246)

This is already a lot of material to build a preference satisfaction theory of well-being. Are there any other, more substantive constraints?

[4]On the failure of philosophers to think comparatively, see Broome (1999*b*). For philosophers who do think comparatively, see, for instance, Rawls (1971:437); Smart (1973:48); Hare (1952:186). For some arguments against comparative thinking, see Rohr (1978). For a clear comparison and analysis of the concepts of desire and preference, see Harsanyi (1992, 1997:135–6). Note also that my argument is not that desire has no place in value theory at all. Consider, for example, the dispositional theory of value of David Lewis (1989). On his view, values are to be analyzed in terms of counterfactual desires, but how we balance and instantiate values in our own lives is not. He says, "our present business is not with the balancing, but with the prior question of what values there are to balance" (1989:124).

[5]But see Sen (1977). For general analyses of the concept of preference, see the papers in Fehige and Wessels (1998).

6.2 A *Smorgasbord* of Desire Satisfaction Theories of Well-Being

It is quite uncontroversial that the actual desire and preference satisfaction theories of well-being are implausible. There are many reasons for this. First, a person's set of desires can be incoherent: she can desire contradictory states of affairs. This should prompt us to move to an actual preference satisfaction theory. Second, actual preferences may however be irrelevant to the person's well-being: this should prompt us to include only self-regarding preferences. Third, the satisfaction of actual self-regarding preferences may still fail to promote the person's well-being. For these reasons, different proposals of desire satisfaction theories of well-being—as they are often called in the literature—could be understood as proposals for specifying a subset of (actual or possible) self-regarding preferences. The proposals argue that only the satisfaction of the preferences in this subset constitutes what is good for the person.

What distinguishes these proposals, then, are the *constraints* which define the relevant subset. One way to classify these constraints is according to the domain of self-regarding preferences they apply to. The domain might be the self-regarding preferences *the person actually has*.[6] What these constraints do is to take the self-regarding preferences of the person, and throw some of them out. Only the resultant preferences are relevant to the person's well-being. Borrowing an expression from Robert Goodin (1986)—although he uses it in a somewhat different context—I call these *preference laundering constraints*. For example, such a theory might operate with a *causal history* constraint. On this view, only the satisfaction of those preferences promotes the well-being of the person whose formation has followed some appropriate causal route—like being uninfluenced by external manipulation.

Another set of constraints includes *counterfactual constraints*. These operate not on the set of the actual preferences of the person, but on her set of possible counterfactual, hypothetical, or ideal preferences. These constraints take some of the person's internal features and external circumstances and abstract away from them. In this idealization process, the person's preferences change as the person, counterfactually, becomes more ideal to form preferences. Metaphorically speaking, only the preferences of the "ideal counterpart" of the person are relevant to the well-being of the (non-idealized) person. These preferences might not be preferences that the person actually has. Rather, the idea is that the satisfaction of only those preferences is relevant to well-being which would be formed by the person under ideal conditions for preferring. Counterfactual constraints are used in what I call idealization theories of well-being (see page 7). The ideal advisor theory is one such theory. It holds that only the satisfaction of those preferences promotes the

[6] "Actual," once again, is used in the modal, and not the temporal sense. Also, henceforth in this section, for the sake of brevity, when I say "preferences," I mean self-regarding preferences. What self-regarding preferences are will be explained in detail in the next section.

person's well-being which the person would have were she adequately informed and appropriately rational. Another version of the idealization theory may hold that the relevant preferences are those which would be formed under Buddhist meditation.

Note that many constraints can have both preference laundering and counterfactual forms. A preference laundering form of the ideal advisor theory would hold that only the satisfaction of those preferences makes the person better off which are formed *when* the person is adequately informed and appropriately rational. Similarly, a twin of the "Buddhist meditation theory" would single out those preferences which *are* formed under meditation. But these versions are not very convincing, given that people are hardly ever adequately informed and appropriately rational, and most of their preferences are not formed under Buddhist meditation. Hence such theories would not be very useful. On the other hand, the causal history constraint may also have a counterfactual form. It would select those preferences the person would have, were her desire formation entirely free of external manipulative influences. Of course, such a theory has to give an account of what counts as manipulative influence, and why such influence can never contribute to the formation of preferences whose satisfaction makes the person better off.

There are many other possible constraints. This means that there is a plethora of possible preference satisfaction theories of well-being. It would be a hopeless undertaking to try to discuss them all. What I can do instead is to discuss only a few—those which have actually been proposed or discussed by philosophers. After all, there must be a reason why these, and not other, versions of the theory have been proposed; other versions are likely be the target of obvious objections. What I shall try to do is to build the case that no form of the preference laundering constraint is plausible; if you want to be a preference satisfaction theorist, you have to choose some counterfactual constraint to identify the preferences whose satisfaction is relevant to welfare. Given, however, the possibility of many unexplored preference laundering constraints, my arguments yield little more than an initial presumption against theories employing such constraints. The case for counterfactual constraints will be examined in the remaining sections, and there again, I take the discussions in the literature as evidence that only certain forms of counterfactual constraints would survive systematic scrutiny.

Consider first the theory which is often called the *basic desires theory*. On this view, there is a special class of desires and only their satisfaction promotes well-being. This class includes desires which are generally or even universally held by human beings as a matter of their psychology and biology. This sort of view can be found in some of the writings of John C. Harsanyi. He says,

> all human beings have *much the same* basic biological and psychological needs, and, therefore, have *much the same basic desires*. (1997:139)
>
> What makes all these things intrinsically valuable to us is the fact that

> they are the *objects of our basic desires*, which we largely share with other human beings, due to our *common human nature* and to our *common biological and psychological needs*. (1997:141)[7]

The basic desires theory employs some constraint which is specified by psychological and biological criteria for identifying basic desires, and argues that the satisfaction of these desires makes us better off.[8]

There is no doubt that the satisfaction of the desires which we have due to our biological and psychological make-up contributes to our well-being. Satisfying these desires may even have special urgency for social policy. But as a *general account* of what makes people better off, this view is hopelessly limited. We have many other desires—more precisely, self-regarding preferences—whose satisfaction is good for us, and these are neither determined by our biology and psychology, nor are they generally held. Your desire to read tedious philosophical works is surely neither general nor determined by your biology or psychology, but your reading them may be good for you (especially if you *do* prefer reading such works).

An apparently more plausible theory based on preference laundering is the *global success theory*. It is discussed (but ultimately rejected) by Derek Parfit (1984:494–9). On a success theory, what makes you better off is the satisfaction of your preferences about your own life. On the global success theory, moreover, only those preferences are relevant which are about a part of your life considered as a whole, or about your whole life. In contrast to a global preference, a local preference is any preference about some limited aspect of your life. Hence, on the global success theory, well-being consists in the satisfaction of global self-regarding preferences.

Parfit applies the local-global distinction to desires, and not preferences, but this does not make a difference, since he uses desire in its dispositional sense. He introduces the distinction because he worries that under certain further assumptions, the local success theory leads to counterintuitive results. Suppose you are addicted to a drug, and each morning you have a very strong desire to take this drug. Your desire does not have phenomenological content—it is not unpleasant to have this desire in any way. Moreover, since I give you the drug every morning, a very strong desire of yours is fulfilled each day; and the drug does not interfere with other pursuits in your life, and it causes neither pleasant nor unpleasant feelings. It is likely, however, that you would rather not be addicted to this drug.

[7] All emphases are his. Harsanyi is better known for his version of the ideal advisor theory (see page 91). But in light of some of the things he says about well-being, it seems that sometimes he has the basic desires theory in mind. I think he intends this theory not as a rival to his ideal advisor theory, but as a complement to it.

[8] A similar theory might identify basic desires not by psychological and biological criteria, but simply by their being generally or even universally held. But it's hard to see what argument could show that *only* generally or universally held desires promote well-being.

However, by satisfying your very strong desire each day, I do what is better for you than satisfying your desire not to be addicted—which is weaker either on its own, or it is weaker if we "add up" the strengths of the daily desires for the drug. Hence, on the local success theory, your well-being is promoted by making you an addict (1984:497).

In contrast, the global version of the success theory seems to yield the intuitively right answer in such a case: since it only counts the desires that you have about your life considered as a whole, or about some aspect of your life taken as a whole, being addicted does not make you better off. On the contrary, by making you an addict, your global desire not to become addicted is not satisfied. You end up being worse off.

The global success theory, however, runs into difficulties when it tries to explain why local desires or preferences do not matter for well-being. Passing an ice-cream parlor, you form a strong whim to have some ice-cream. Why satisfying that local preference is not good for you? It seems to me that narrowing the set of relevant preferences according to the scope of their objects is not a promising strategy, because it is off-target. Satisfying local preferences can make you better off. Furthermore, global self-regarding preferences may be formed and ingrained without sufficient information, for instance. People can have preferences over parts of their lives, even their whole lives, whose satisfaction does not make them better off—for they may be based on insufficient information and inadequate reflection.

A variation on the scope problem plagues other type of theories based on preference laundering constraints. Consider *life plan* theories. Roughly, these views connect well-being to the execution of a person's life plan, and particular preference satisfactions derive their value from the place they have in the hierarchy of preferences constituting that life plan. One variant of such theories is proposed by Joseph Raz (1986:288-320). On his view, well-being consists in the satisfaction of the person's biologically determined needs or desires, and the success in achieving her goals, where these goals are nested in a hierarchical structure with the person's "comprehensive goals" on top. The satisfactions of the person's particular preferences contribute to the person's well-being insofar as they contribute to the attainment of more comprehensive goals.[9]

Life plan theories have a scope problem due to their overly *intellectualist* nature. As a matter of fact, I doubt that most people execute life plans or entertain a complete and consistent hierarchical system of more or less comprehensive goals. I suspect such persons exist only in the fantasies of philosophers. Surely, people have global preferences, plans, and important goals; but they do not form the structured hierarchy some philosophers seem to suggest. This makes it very hard to

[9]Note that according to Raz, the person's goals derive from values, and not desires. But insofar as those goals determine her preferences, Raz's different position in value theory is irrelevant here.

evaluate their well-being on these views. But even if people had life plans and systems of comprehensive goals, these may have nothing to do with their well-being. You are not necessarily better off by executing a life plan of sacrifice, or realizing some aesthetic or perfectionist ideal; indeed, your very reason for choosing these plans or ideals may have nothing to do with your well-being. Life plan theories disregard the distinction between self-regarding desires and preferences on the one hand, and other-regarding desires and preferences, on the other.

I have surveyed a few preference satisfaction theories of well-being with preference laundering constraints to lend plausibility to the case that preference laundering constraints are inadequate, and preference satisfaction theories must employ counterfactual constraints. Of course, there are many other preference laundering constraints I cannot examine. But this last point—that preference laundering constraints may be unable to separate self-regarding preferences from other-regarding preferences—may strengthen my case for counterfactual constraints. Thus, I turn to the question of how to distinguish between self-regarding and other-regarding desires and preferences.

6.3 Self-Regarding Desires

By definition, a self-regarding preference is a preference that has to do with the person's well-being. A preference satisfaction theory of well-being must hold that only the satisfaction of those preferences contributes to our well-being which are self-regarding, since many of our preferences have nothing to do with what is good for us. If you prefer that one day the human race encounters an intelligent and benevolent race from another galaxy, your preference is not about what is good for you. If you prefer that people in poor countries have a better life, you don't prefer what is good for you.

Therefore, I have to give an analysis of "self-regarding." In order to do this, I have to revert back to desire-talk. This is because preference is a comparative notion: preferences form orderings. Given the dispositional conception of desire, the elements of preference orderings can often (but not always) be considered desires. That is, if you prefer an apple to a pear, then often it is the case that you both desire the apple and the pear, and you give priority to your desire for the apple. But then a self-regarding preference is nothing more but the comparison of self-regarding desires. More precisely, *self-regarding preferences are (orderings of) pairwise comparisons of self-regarding desires, and only such desires*—because many preferences register our priorities between what is good for us versus what is good for others, or desirable for some other reason. These preferences have no role in a theory of well-being, because they concern how we weigh our well-being with other values. No doubt such preferences are important, but they are important from the perspective of some other theory—perhaps our more general moral theory.

But what is a self-regarding desire? There are several ideas abroad about this question. One proposal is that a self-regarding desire is one that conforms to the familiar *experience requirement*. According to this requirement, in order for the satisfaction of a desire to contribute to a person's well-being, the person must in some way experience the satisfaction of the desire. But in what way? Perhaps it must enter the conscious experience of the person. This reading, however, is misleadingly hedonist. It is natural to understand it as requiring that the person takes some pleasure or enjoyment in the satisfaction of her desire. I suspect such a misreading lead Sumner to argue that if preference satisfaction theories of well-being incorporate the requirement, they become hedonist theories (see page 57). If only those desires are self-regarding whose satisfaction causes pleasure or enjoyment, it is hard to see why the desire, and not the pleasure or enjoyment its satisfaction results in, is relevant to well-being.

Perhaps "experiencing the satisfaction of desire" means that the satisfaction of the desire must make a difference to the person's life. On this reading, the experience requirement holds that the satisfaction of a person's self-regarding desire must have some *causal effect* on the way that person's life goes, including what subsequent desires or preferences she has. Thus, if some of your desire is satisfied, but its satisfaction never affects your life—maybe because you never learn that it has been satisfied—it does not make you better off. And if some of your desire is not satisfied, even though you mistakenly believe that it has been satisfied, this cannot make you worse off. (Even though your false belief that it has been satisfied may make a difference to your life.)

Now, there are widely differing intuitions among philosophers about cases like these. Some believe that causally inert satisfactions (or frustrations) make their respective desires irrelevant to a person's well-being. Others believe that desires with causally inert satisfactions (or frustrations) do make a difference to the person's well-being. Whichever intuition you have, however, the experience requirement is not the proper requirement for identifying self-regarding desires. If you have the former intuition, notice that the requirement admits into the determination of your welfare the causally potent satisfactions and frustrations of desires whose satisfaction has nothing to do with how well off you are. Suppose you desire that the quality of life of people in poor countries is improved. This desire is satisfied, and this has some causal impact on your life. But it does not mean that it is better for you that they are better off. You desired *their* life to go better, not yours. This distinction has nothing to do with the causal role of the satisfaction of this desire in your life. On the other hand, if you have the latter intuition—that desires with causally inert satisfactions (or frustrations) are relevant to your well-being—you have already given up the experience requirement. Whatever your intuition is, self-regarding desires need some other criterion.

Parfit (1984:494–5) proposes an *historical* criterion. He imagines two scenar-

ios: in both of these, he's an exile who cannot communicate with his children, but he desires that they succeed in their lives. In the first scenario, one of his children is killed by an avalanche. In the second scenario, in addition to his desire that the lives of his children go well, he had tried to give them a good start in life before he became an exile. Despite his efforts, the lives of his children go badly, and partly this is due to his mistakes as a parent. But he never learns of their fate.

Parfit thinks that the fact that his desire is not fulfilled is bad for him in the second case, but not in the first. The difference is that in the second case he bears part of the responsibility for his unsatisfied desire, since he had a hand in the subsequent failure of his children's lives. On this view, self-regarding desires are identified by the role that the person's efforts have played in their fulfillment or frustration. But clearly, this is not a good criterion either. Whether you have been in a position to facilitate the satisfaction of your desire has nothing to do with whether the desire is about what is good for you. If you desire to be rich, and you inherit a lot of money from a distant, unknown relative, it is good for you that your desire is satisfied, even though you have not made any effort for its satisfaction in any way.

A more promising alternative is to say that self-regarding desires are the desires the person holds "for her own sake." But this a bit mysterious. I think what it means is this. In order to distinguish self-regarding desires from other desires, we have to appeal to the *reasons* for which a person holds the desires she does. Quite simply, a self-regarding desire is a desire that you hold for the reason that the satisfaction of this desire would promote your well-being.

Consider Parfit's famous example of meeting a stranger on the train (1984:494). The stranger is ill, but after their conversation Parfit forms the desire that the stranger be cured. He never meets the stranger again, who, unknown to him, is subsequently indeed cured. The satisfaction of Parfit's desire does not make Parfit's life better—but not because it never enters his experience, and not because he makes no effort to satisfy this desire, but because the reason Parfit forms it is that its satisfaction would be better for the stranger. Parfit's desire is irrelevant to Parfit's well-being, because the reason he holds it has something to do with the stranger's life, and not his.

Of course, Parfit could hold this desire for the reason that *his* own life go better, but there would be something strange with that reason—especially given the extremely low likelihood that he ever meets the stranger again. Perhaps holding the desire for that reason is not impossible or inappropriate, in which case Parfit's life might go better or worse depending on the stranger's fate. But a more natural reaction to this case would be to say that Parfit probably mistakenly believes that the reason he formed this desire for is that his own life go better, and in fact he holds it for some other reason. Maybe he has not given enough attention to clarifying his reasons for himself; maybe he deceives himself or he is deceived by others; maybe he is unable to clarify his reasons because of some psychological deficit. People

can be mistaken about the reasons for their desires in many ways.

Because of this problem, it might seem that a "reason-based account" of self-regarding desire is no improvement over the other proposals. But this problem can easily be remedied. All we have to add is that the reasons which delineate self-regarding desires are the reasons the person would have were she free of distorting influences. That is, we can amend the account with certain counterfactual constraints to avoid the problem. These counterfactual constraints are *epistemic* and *cognitive*: they require that the person is fully informed and ideally rational. The fully informed and ideally rational "counterpart" of the person would be able to adequately clarify the reasons of the person: she is fully informed, she can give enough attention to the matter, and cannot be influenced by external manipulation; she is ideally rational, hence free of psychological deficits; and self-deception is conceptually impossible for such ideal agents. Therefore, this account proposes the following definition:

Self-regarding desire. A self-regarding desire is a desire that the person, were she fully informed and ideally rational, would have for the reason that the satisfaction of this desire would promote her own well-being.[10]

In my view, a desire or preference satisfaction theory of well-being must be formulated as a preference satisfaction theory, and not a desire satisfaction theory. It must also employ some constraints on the relevant preferences, but preference laundering constraints are inadequate for various reasons. One reason is that only a preference satisfaction theory with counterfactual constraints can distinguish between self-regarding and other-regarding desires, and hence preferences. In particular, only a theory that uses both epistemic and cognitive constraints can do this job. This theory is the ideal advisor theory.[11]

6.4 Towards the Ideal Advisor Theory

In this chapter, I have tried to make a case for the ideal advisor theory through examining a number of constraints that a preference satisfaction theory of well-being has to impose on the preferences whose satisfaction is relevant to well-being.

[10] One may object that this account of self-regarding desire is circular. I will have something to say about this point on page 146.

[11] Defining self-regarding desire in terms of the reasons for desires the person in ideal conditions would have perhaps also helps to explain why people have widely diverging intuitions about the relevance of desires to welfare in cases in which a desire is satisfied, but the satisfaction does not affect the person's life (cases of the stranger on the train, the fate of the children of the exile, desires satisfied after one's death). Perhaps our intuitions are different because it is often unclear what the reason is for which the heroes in these stories have the desires they do—or what the appropriate reason for having these desires would be.

The ideal advisor theory holds the well-being consists in the satisfaction of the (self-regarding) preferences the person would have were she adequately informed and appropriately rational. Of course, there are many ways of spelling out the information and rationality requirements—as I call them later, the epistemic and the cognitive constraints. These will be examined in Chapter 7.[12] My conclusion so far is that if you want to be a preference satisfaction theorist, you need to impose both cognitive and epistemic constraints on the relevant preferences. That is, you have to move to a theory like the ideal advisor theory.

There are, however, philosophers who deny that if you want to be a preference satisfaction theorist, you have to accept an idealization theory. In particular, it has been argued that the information requirement—the epistemic constraint—is superfluous. This constraint is attacked, for instance, by Mark Murphy (1999). He notes that there are two major considerations that push philosophers accepting a "DF-theory" (the desire satisfaction or fulfillment theory of well-being) towards a "Knowledge-Modified" version of the theory (that is, one with a counterfactual epistemic constraint for specifying the desires which are relevant to well-being). These are: (1) that desires can be based on false beliefs, and (2) that desires can be absent due to a lack of true beliefs. According to Murphy, the proper account of the individuation of desires entails that neither consideration justifies the need for a counterfactual epistemic constraint. As he argues:

> Of all the conditions that DF theorists have incorporated within their hypothetical desire situations, the information condition has been the most common and has been thought to be the weakest and most in the spirit of DF theory. A successful argument against the Knowledge-Modified form of DF theory is thus an excellent, if still *prima facie*, case against all Modified versions of DF theory. (1999:249)

Consider (1). Murphy argues that there are only two senses in which a desire can be based on false beliefs. In one sense, the *causal history* of the desire contains some false belief. Murphy believes that the falsity of beliefs which have played a role in the formation of a desire does not make the desire irrelevant to well-being. This is so both if the desire persists when the person learns that the belief is false, and also if the desire is dropped if the person learns that the belief is false.

A preference satisfaction theory with an epistemic constraint can handle these cases easily. The theory asks what would happen if the person learnt that the belief was false. If the corresponding desire persists, then it is a desire whose satisfaction makes the person better off, since the desire withstands idealization. Thus, so far there is no disagreement between Murphy and this theory. If, however, the desire

[12] In the discussion of the theory, I will drop the "self-regarding" qualifier in front of preference. Unless indicated otherwise, adequately informed and appropriately rational preferences are always understood to be self-regarding.

would be dropped, its satisfaction does not make the person better off. Murphy denies this; he thinks that a false belief somewhere down along the causal chain does not make a difference to the relevance of the desire. I think it does, and later I will present some kinds of desire for which this seems to be the case.

But first consider the second sense of a desire being based on a false belief. A desire may be based on beliefs which are *specificatory* or *instrumental*. Take specificatory beliefs. Murphy imagines that he desires a baseball signed by a famous player, and he has the false belief that *this* ball was signed by that player, so he is motivated to get this ball; therefore, in virtue of that motivation, he has a desire to get this ball. On a view which employs a counterfactual epistemic constraint, his desire to get this ball would not count among those whose satisfaction promotes his well-being, since if he did not have a false specificatory belief, he would not have a desire for this ball. But Murphy believes that we do not need to appeal to the person's counterfactual or hypothetical desires in order to make sure that the satisfaction of the desire for this ball—a specificatory desire—is irrelevant to the person's well-being. We do not need such an appeal, because on a plausible account of desire individuation, *there are no specificatory desires*.

The case is similar with instrumental desires: to take an example by Parfit (1984:117), if I desire to meet a beautiful librarian, and I want to go to the library merely in order to meet her, and you introduce me to this librarian, then I have no unsatisfied desire. Again, this is because on a plausible account of desire individuation, *there are no instrumental desires*.

That account of the individuation of desires is this. Since we ascribe desires in order to explain action, a principle for the individuation of desires is governed by the needs of an explanatory theory of action. Thus, a person has a desire if and only if the desire plays a role in the *scope* and *power* of the person's motivation. Two putative desires, A and B, are different if and only if either there is a state of affairs toward which A and B motivate the person, but toward which neither A nor B alone motivates the person; or there is a state of affairs toward which A and B together motivate the person to a greater degree than A or B alone does. In the first case, the additional desire changes the scope of the motivation; in the second case, it changes its power. If a putative desire plays no role in the scope or the power of the person's motivation, then there is only one desire (1999:253–4).

Thus, when Murphy desires to have a baseball signed by a famous player, and he believes that *this* ball is signed by that famous player, he does not have a separate desire for this ball: we explained his motivation by citing the desire and the belief. Similarly, when I desire to meet the beautiful librarian, and I believe that by going to the library I can meet her, I don't have a separate desire to go the library. If Murphy falsely believes that this ball was signed by the famous player, or I falsely believe that I will meet the librarian in the library, we do not need to appeal to what he and I would desire if we had true specificatory and instrumental beliefs: we

would desire no differently. There is no desire to be filtered out by a counterfactual epistemic constraint. Rather, the reason we are not better off going for the ball and to the library is simply that our desires to have the ball signed by the player, and to meet the beautiful librarian, are not satisfied due to our false specificatory and instrumental beliefs.

Consider now (2): the rationale for accepting a counterfactual epistemic constraint that desires can be absent due to a lack of true beliefs. One way acquiring true beliefs might make a difference is by causing us to form new desires. Suppose I learn that stamp-collecting is a relaxing pastime, and I come to have a desire to start a stamp collection (perhaps giving up the relaxing pastime of reading, such that I have a separate desire for stamp-collecting). Murphy (1999:260) believes that in the absence of this belief, stamp-collecting would not contribute to my well-being, thus there is no advantage to be gained by asking whether I would start a stamp collection if I knew that stamp-collecting was relaxing. My reply is that this is posing the wrong question. The right question to ask is this: how would the desires the person has now change if she acquired new true beliefs? I think there are kinds of belief whose acquisition would make a difference to the person's desires. I mention some below.

Another way acquiring true beliefs might make a difference is by changing specificatory and instrumental beliefs. Thus, Murphy imagines that he has a desire for a baseball signed by a famous player, but he lacks the true belief that the ball in front of him is signed by that player. Similarly, I desire to meet the beautiful librarian, but I lack the true belief that I can meet her by going to the library. But, once again, in order to explain why the satisfaction of our desires for the baseball and to meet the librarian (respectively) would be good for us, we do not need to appeal to counterfactual desires, since if we had the true beliefs, we would not acquire new desires. All we need to appeal to is that the ball in front of Murphy is in fact signed by the player, and going to the library is, indeed, the way to meet the beautiful librarian.

Now, suppose we accept Murphy's principle for the individuation of desires, together with the implication that there are no instrumental and specificatory desires. Even so, his case against a counterfactual epistemic constraint is not conclusive.

On the one hand, there might be other reasons why a desire satisfaction theorist of well-being would want to employ some counterfactual constraint on a person's desires. She might be worried that the person has desires which are formed in a state of anxiety, depression, emotional disturbance, compulsion, or lack of proper "imaginative acquaintance."[13] The argument against an epistemic constraint is not a good *prima facie* case against such an idealization theory.

[13]See, for instance, Smith (1994:155–6). For a discussion of imaginative acquaintance, see Lewis (1989:121–3).

On the other hand, even if we put these other considerations aside, there remain reasons for employing a counterfactual epistemic constraint in a *preference* satisfaction theory of well-being. Murphy misses these, since he puts his case forward in terms of desires. But we need to use the concept of preference. When we form preferences, we use beliefs he does not consider. For instance, you need to *weigh* or *balance* your desires. Suppose you can either become a philosopher or a concert pianist, and you desire both: whichever career you choose might depend on the beliefs you have about the relative value, for you, of being a philosopher or being a concert pianist. These beliefs need neither be specificatory, nor instrumental. One kind of belief that might be relevant for deciding between these two alternatives concerns whether you think you will be more successful as a philosopher than as a musician, or a philosopher's life will be more rewarding for you than a life as a musician.

There are other sorts of relevant beliefs too, and some of these are instrumental or specificatory beliefs. Return to Parfit's example. I desire to meet a beautiful librarian, and I might be able to do that if I go to the library, since there's a fair chance that she is there. But there is also a chance that she will be at your party. Suppose I have no genuine desire to go to the library or to go to your party. Wherever I go depends on the probability I assign to where the beautiful librarian is. On Murphy's account, even though I do not form an additional desire to go to the library or to the party, my belief about where the beautiful librarian is makes a whole lot of difference to what I do: it determines whether I go to your party or to the library. These probabilistic beliefs are clearly instrumental. A similar story can be told about specificatory beliefs—from which I spare the reader.

These beliefs can clearly be false. An adequate preference satisfaction theory of well-being has to be able to exclude preferences based on such false beliefs, or lack of true beliefs. It has to appeal to counterfactual situations to determine what is good for us.

Chapter 7

In Defense of Ideal Advisors

7.1 Idealization Theory

Subjectivism about welfare has been a very influential position. Many subjectivists hold that welfare consists in the satisfaction of preferences. Most of them, however, agree that only the satisfaction of those preferences constitutes well-being which count as improvements over the person's actual preferences, since those preferences might be mistaken or irrelevant in several ways. There are many possible versions of this type of theory. One version holds that improved preferences are identified by whether they conform to some counterfactual constraint, or set of constraints. I use the expression *idealization theory* as a generic term for such theories:

Idealization theory. One thing x is better for a person, or promotes her well-being more, than another, y, if and only if the person would prefer x to y, were she in ideal conditions for preferring.[1]

The most familiar versions of the idealization theory employ *epistemic* and *cognitive* constraints for the selection of improved preferences. The epistemic constraint requires that the person is informed about her circumstances, range of options, and the possible consequences of her choices. The cognitive constraint requires that when the person evaluates her circumstances and options, she does not make any mistakes of representation of facts or errors of reasoning. In short, these constraints require that the person is informed and rational in some appropriate sense, and she is in ideal conditions for forming preferences if and only if she is. That is, these theories involve a causal-counterfactual process of idealization on the preferences of the person: adding true information, we "observe" what changes in these preferences occur, making sure that the changes are not caused by errors of reasoning. The resultant, informed and rational preferences are improvements over the preferences of the person. These theories connect well-being to these improved preferences. They are sometimes called informed, or hypothetical, or counterfactual desire (or preference) theories, ideal advisor views, or, more often, full-information theories of welfare. I will refer to them collectively as the

[1]By "preference," I mean *weak preference* in the decision theorist's sense. Weak preference includes indifference, hence x and y are equally good if and only if the person is indifferent between them in ideal conditions for preferring. Also, "preference" is always understood as *self-regarding* preference, as explained in the previous chapter.

ideal advisor theory of well-being. My favored interpretation, which I call IRP, is spelled out in Section 8.1.

To avoid confusion, I will call the person placed in the hypothetical ideal conditions for preferring the *ideal advisor* of the person. But, of course, distinguishing between the (actual) person and her "ideal advisor" is no more than a useful metaphor which makes the discussion easier.

Many philosophers nowadays believe in some version of the ideal advisor theory. Recently, however, the plausibility of such views has been questioned repeatedly.[2] In this chapter, I first present some seminal versions of the theory, then I examine various counterarguments. As I argue, these arguments expound some common background ideas, and these are perhaps not as uncontroversial as one might have initially thought. Thus, the objections do not conclusively refute the theory—but they definitely help to work out its most plausible version.

7.2 Four Versions of the Ideal Advisor Theory

Idealization theories in general, and ideal advisor views in particular, can also be put forward as theories of *normative reasons*. A theory of normative reasons tells us what a person has the most reason to do. In these theories, what a person ought to do is established by discovering what she, in the relevant counterfactual situation, would be motivated to do. What she would be motivated to do as an ideal advisor is what she ought to be motivated to do in her actual situation: the motivating reasons of the ideal advisor are the normative reasons of the person. The role of idealization is to specify what criteria the person should aim at satisfying in her deliberation of what she has the most reason to do, and it explains how these reasons can be compelling to her—by being connected to her actual motivating reasons *via* a number of counterfactual steps.

Nevertheless, it is important to distinguish between an idealization theory of well-being and an idealization theory of normative reasons. Even if, for instance, an ideal advisor theory turned out to be both the best theory of well-being and the best account of normative reasons, the two would still be different, since what we have the most reason to do is not always the best thing for us to do. The failure to distinguish between these can lead to misplaced arguments.[3]

[2] Advocates include, for instance, Railton (1986*a*,*b*), Gauthier (1986), Darwall (1983:85–100), Harsanyi (1982), Hare (1981:101–6, 214–8), Brandt (1979), Rawls (1971:407–24), and Sidgwick (1907), among others. Some have interpreted Griffin (1986) as such a view as well, although he seems to have moved to an objectivist theory in Griffin (1996). See also his remarks in Griffin (2000). Recent opponents to different versions of idealization theory include Hubin (1996), Loeb (1995), Rosati (1995), Sobel (1994), Anderson (1993), Cowen (1993) and Velleman (1988).

[3] For instance, Gibbard (1990:18–22), in arguing against Brandt (1979), asks why he ought to do, when he is lost in a forest, what he would want himself to do if he was fully informed, instead of

Still, both kinds of theory have a common strategy: they take the actual preferences of the person and provide information pertaining to her situation—factual knowledge of the alternatives, possible consequences of her choices, and the like. The aim is to establish what preferences would be generated by making more and more information available, given that the person does not make mistakes of reasoning and avoids other sorts of cognitive error.

Therefore, on an ideal advisor theory of normative reasons, actual preferences provide no reasons, or provide reasons only insofar as they would remain intact throughout the idealization process. "Real" reasons are discovered in the hypothetical situation; that is, reasons are rooted in hypothetical preferences. On one type of theory using epistemic and cognitive idealization, we can be sure that these reasons are capable of being *authoritative* for the person because they are established in accordance with a requirement that there must be an internalist link between hypothetical and actual motivation—between the motivating reasons of the ideal advisor and the motivating reasons of the actual person. These are ideal advisor views of normative reasons with the *internalist requirement*. On other versions, the authority of the reasons uncovered in the hypothetical situation are secured by the *convergence requirement*—the requirement that these reasons are shared by all persons placed in ideal conditions.

Ideal advisor theories of well-being can also appeal to the internalist or the convergence requirement to ensure the authority of the ideal advisor's preferences in determining what promotes the welfare of the person. One ideal advisor view with the internalist requirement is the theory of Richard B. Brandt (1979). His aim is to work out a theory of what is rational to want and to do—which he identifies with what is good for the person (1979:15). He describes idealization the following way:

> some intrinsic desires and aversions would be present in some persons if relevant available information registered fully, that is, if the persons repeatedly represented to themselves, in an ideally vivid way, and at an appropriate time, the available information which is relevant in the sense that it would make a difference to desires and aversions if they thought of it. (1979:111)
>
> By "ideally vivid way" I mean that the person gets the information at the focus of attention, with maximal vividness and detail, and with no hesitation or doubt about its truth. I mean by "available information" … relevant

pursuing one of the standard strategies for finding one's way out of the woods. I agree that what he has the most reason to do is to pursue such a strategy, given that he does not have full information; but it would be still better for him to be fully informed, and act upon the preferences he would then have. His argument is presented as an argument against the idea of idealization in general—but it is compelling as an argument against an ideal advisor theory of normative reasons only, and not so convincing against an ideal advisor theory of well-being. In context, whether it is successful depends on how we interpret Brandt's project. (I thank David Sobel for helping me clarify the distinction between idealization theories of well-being and of normative reasons. See also his 2001.)

> beliefs which are a part of the "scientific knowledge" of the day, or which are justified on the basis of publicly available evidence in accordance with the canons of inductive or deductive logic, or justified on the basis of evidence which could now be obtained by procedures known to science. (1979:111–2)

Relevance of information should be judged by its content. Information should be presented repeatedly until it "registers," a sign of which may be that we have reason to suppose that further repetition would induce no change in the person's desires; and it should be presented at the time when it could make a difference to deliberation. Brandt sums his view up this way: "This whole process of confronting desires with relevant information, by repeatedly representing it, in an ideally vivid way, and at an appropriate time, I call *cognitive psychotherapy*" (1979:113, his emphasis).

That is, in the idealization process, we present information to the person in order to find out whether her motivational structure would alter due to the acquisition of new beliefs (or being vividly reminded of old ones). The presentation is subject to certain conditions. One is that the information should be presented repeatedly. This invites the question of how much repetition is needed. We cannot simply say that information should be repeated until it induces a change in motivation; this at least presumes that the more repetition, the likelier the change. This may not always be so, but even if it was, the point of repeating information is to enable the person better understanding, not to induce a change of motivation. The criterion for the number of sufficient repetitions, therefore, cannot be a function of motivational change. But are there any independent criteria then?

Brandt gives the general criterion that information should register. But there is more to registration than repetition. Maximal vividness and detail of information also enhance the epistemic conditions. But just as before, a case can be made against the presumption of "the more the better" behind this view. It is not clear why more vivid information would imply epistemic advancement. Moreover, vividness may depend on the psychology of the person—what counts as vivid presentation of information for you may be quite vague for me. Once again, vividness and detail cannot be measured by the motivational change they induce; they are to be measured by whether they help information to register fully. But since it is not entirely clear how one can make sure that information has registered, how far the idealization process should go is indeterminate. The same point applies to the the extent of available information. Brandt only requires information that is publicly available *via* scientific methods or reasoning. This does not ensure that all information that could make a difference will be at the person's disposal, and also introduces an element of indeterminacy. In addition, obtaining some pieces of information may be prohibitively costly. But it is not that these cannot be used to criticize what the person prefers (1979:13). So some indeterminacy is unavoidable

in specifying relevant information.[4]

As a matter of fact, Brandt's theory is not a "full information" theory, since he does not require all relevant information. Moreover, it cannot reply to the arguments from the psychological shortcomings of specific persons, since it does not abstract away from such shortcomings. These problems may be, however, overcome by tightening both the epistemic and cognitive constraints: by requiring that all relevant information must be at the disposal of a cognitively perfect ideal advisor.

Peter Railton's theory avoids the problems of the presentation, vividness, and amount of information by doing precisely that. It posits a duplicate super-agent, possessing extra-human cognitive capacities:

> Give to an actual individual A unqualified cognitive and imaginative powers, and full factual and nomological information about his physical and psychological constitution, capacities, circumstances, history and so on. A will have become A^+, who has complete and vivid knowledge of himself and his environment, and whose instrumental rationality is in no way defective. We now ask A^+ to tell us not what *he* currently wants, but what he would want his non-idealized self A to want—or, more generally, to seek—were he to find himself in the actual condition and circumstances of A. (Railton, 1986*b*:173–4, his emphasis)

Railton's theory yields the following theory of welfare:

> an individual's good consists in what he would want himself to want, or to pursue, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality. (Railton, 1986*a*:16)

Railton's theory is an improvement over Brandt's theory in another way as well. On this view, well-being is tied to the hypothetical preferences of the ideal advisor. That is, the person placed in ideal conditions forms preferences over her preferences in her actual conditions. These preferences are about what is good for the actual person, and not about what is good for the ideal advisor. Evidently, any ideal advisor theory must be put forward this way.

At the same time, however, the actual self and the fully informed and cognitively perfect ideal advisor may be very unlike, because Railton gives unlimited cognitive and imaginative powers to ideal advisors. But this may detach the preferences of the ideal advisor from those of the actual person's: if idealization can profoundly change the preferences of a person, how can we make sure that the internalist link between her preferences and the ideal advisor's preferences remains

[4]For similar arguments, see Loeb (1995) and Velleman (1988). See also Gibbard (1990:18–22). For broader discussions of Brandt's project, see Daniels (1983) and Sturgeon (1982). See also Brandt (1998) and contrast Kusser (1998).

in place? If you could, ideally, perceive and comprehend information in a way you cannot now, you could be totally alien to your ideal "self," and *vice versa*. And if your ideal advisor is totally alien to you, how could what they prefer you to prefer have any authority over you, here and now?

We cannot stipulate that the idealization process cannot "distance" the ideal advisor this way. If we did, we would have to show why an otherwise possible radical change would be inconsistent with improvement in the person's epistemic situation—why it would not be allowed in the process. But there is no *prima facie* reason to think that an epistemic or cognitive improvement could not result in breaking the internalist link.

Perhaps we do not need the internalist link. At least, this is what is proposed by Michael Smith (1994). His aim is to reconcile a Humean theory of motivating reasons with an anti-Humean theory of normative reasons. His theory replaces the internalist requirement between the motivating reasons of actual persons and ideal advisors with the *convergence requirement*—according to which the desires or preferences of an ideal advisor are authoritative for her actual counterpart if and only if all ideal advisors would form the same desires or preferences over the desires or preferences of the actual person.[5]

The epistemic and cognitive constraints play their usual role:

> what it is desirable for us to do is what we would desire that we do if we were fully rational. ... Thus, what it is desirable for us to do in our actual circumstances is what our more rational selves, looking down on ourselves as we actually are from their more privileged position, would want us to do in our actual circumstances. ... [F]acts about what it is desirable for us to do are constituted by the facts about what we would advise ourselves to do if we were perfectly placed to give ourselves advice. (1994:151–2)

The conditions for the more privileged position are that "(i) the agent must have no false beliefs; (ii) the agent must have all relevant true beliefs; (iii) the agent must deliberate correctly" (1994:156). Smith attributes these conditions to Williams (1980). With false beliefs, you may act in a way that fails to fulfill your preferences—for instance, you may think the stuff in front of you is gin and tonic, whereas it is petrol, and it would certainly not satisfy your thirst (Williams, 1980:102). With regard to (i) and (ii), Smith takes Williams' specification of the conditions of the idealization process.

However, (iii), that the agent must deliberate correctly, admits of different interpretations of correct deliberation. Smith's and Williams' analyses part company

[5]Ideal advisor theories making use of the internalist requirement are *naturalist* and *reductionist* about welfare, or reasons. Those which make use of the convergence requirement are reductionist, but not naturalist. (See Section 10.3.) Note also that Smith's theory is a theory of reasons, and not welfare, therefore it is more general. But it merits mentioning here since it is also very influential in discussions of welfare. For discussions, see, for instance, Hubin (1999) and Sobel (1999).

here. In general, Williams gives a more restricted account of the ways of deliberation. His account includes combination, time-ordering, weighing of desires, and finding constitutive solutions to which option would be the best; finally, reflection and imagination (1980:102). In contrast, Smith permits a more inclusive account of correct deliberation. He proposes that we drop "imagination," and substitute it with the more open-ended but also more determinate forms of "attempts of systematic justification" (1994:158–61). He also mentions that effects of "psychological compulsions, physical addictions, emotional disturbances," depression, etc., are to be excluded (1994:155–6).

After the idealization process, ideal advisors would be left with identical desires. Smith establishes this through giving an analysis of what normative force in general means. His argument is this: there is no "relative" normative force in the sense of a reason being justifiable-for-you but not justifiable-for-me. One of the things the relativity of reasons could mean is that given our differing preferences, what may be a good reason for me in light of my preferences may not be a good reason for you in the light of your preferences. For example, if I prefer beer over wine, and you prefer wine over beer, and you tell me there is a reason to go to the local wine bar for a drink after work—for they sell excellent wine—my reply that this may be a reason for you, but it is not a reason for me, is perfectly understandable. Smith has to argue that this use is imprecise (1994:170–3). He does this by arguing that the differing preferences in the example can be considered as data for the specification of the circumstances, and hence for the idealization process. Therefore, there are no relative reasons, given that the relevant facts include personal differences. So that "which desires *I* end up with, after engaging in such a process, thus in no way depends on what *my* actual desires are to begin with" (1994:173, his emphases). The preferences I end up with give me reasons if and only if all ideal advisors would embrace those preferences for me.

The final theory I present comes from the work of John C. Harsanyi (1982). Harsanyi bases his utilitarian ethical theory on formal decision theory. He distinguishes between the *personal* and the *moral* preferences of a person. The former are preferences that concern the person's own interests (which I called self-regarding preferences), the latter are the preferences she forms in an impartial position (1982:47). In order to show that personal preferences are relevant to welfare, Harsanyi appeals to a further distinction between *manifest* and *true* personal preferences. This latter distinction is needed because people are not always reliable judges of their own well-being:

> a person's true preferences are the preferences he *would* have if he had all the relevant factual information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice. (1982:55, his emphasis)

In contrast, manifest personal preferences are the (self-regarding) preferences the person actually has. Only the satisfaction of true personal preferences constitutes the person's well-being. We arrive at these preferences through an idealization process. There is, however, a further component to Harsanyi's theory. It is introduced through yet another distinction. Suppose that we want to evaluate whole lives, including the personal characteristics of the persons living these lives. Call the objects of this sort of evaluation *extended alternatives*. An extended alternative includes a whole "life-story," together with all the causal influences that shape the preferences of the person having that life-story. The further component to Harsanyi's theory is that ideal advisors have the same preferences among extended alternatives—that they have identical *extended preferences*. If you and I were fully informed and ideally rational, then you and I would have the same extended preferences, between, for instance, the life of a philosopher (together with the personal characteristics and causal history of that life), and the life of a concert pianist (together with that life's personal characteristics and causal history). This is another version of the convergence requirement: ideal advisors agree in their preferences, at least over extended alternatives.

Harsanyi gives a formal argument for the existence of extended preferences, which need not concern us here. In his and other versions of the extended preference theory, ideal advisors are required to have *imaginative empathy* in order to ensure the reliability of their judgments and establish the authority of their preferences. Imaginative empathy is a further cognitive constraint on ideal advisors.[6]

As these illustrations show, there are many ways to construct an ideal advisor theory. Fortunately for my purposes, most of the arguments against these theories—even when they are targeted at a particular author's version—apply to elements that are more or less common to all. Most often, the counterarguments target the very idea of idealization. I turn to some of these more general objections now.

7.3 Some Recent Counterarguments

All versions of the ideal advisor theory hold that what is good for a person is the satisfaction of the preferences she would form in an epistemicly and cognitively more privileged standpoint. These theories make use of a causal-counterfactual process. They go from the actual preferences of a person to a hypothetical set of preferences by proposing counterfactuals about the person's preference changes. The idea is that by evaluating these counterfactuals according to some best theory for the evaluation of counterfactuals, we can determine what the person would

[6]See Harsanyi (1977*a*:51–60), and compare Arrow (1977). For discussion, see Broome (1998); compare Schüssler (1998). For Harsanyi's ethical view, see Mongin (2001).

prefer herself to prefer in ideal conditions for preferring. In such conditions, the person becomes her own "ideal advisor," who is fully informed and possesses perfect cognitive capacities.

One problem may be that the notion of "being fully informed" is incoherent, because informing a person does not lead to the appropriate motivational changes. In other words, there is indeterminacy or there arise unwanted side-effects in the causal-counterfactual process. A version of this argument is what David Velleman (1988) calls the *problem of representation*:

> To ask simply about the motivational impact of the facts, then, is not to ask a determinate question. There is no single motivational impact associated with the facts in themselves. The facts would exert various impacts, when presented in various media, perspectives, and vocabularies. Consequently, I cannot resolve my practical dilemma by asking what I would want after exposure to the facts, since the only accurate answer to that question is, "It would depend on how I looked at them." (1988:366)

Velleman targets Brandt's theory only, but this argument applies to other versions as well. For instance, Loeb (1995) argues more broadly that informing people may have motivational consequences that have nothing to do with the information itself. That is, there could be "motivational side-effects" that defeat the purpose of putting the person in a situation that counts as epistemicly privileged. Since people react to information differently, the resultant ideal advisor might not end up in an epistemic situation that we could consider more appropriate to determine what is good for the person.

It is true that different representations of the very same facts in different media can have different impacts. But as long as the information conveyed is equivalent, the only reason they can have different motivational impacts is that insofar as they are sensitive to different degrees to the same facts, conveyed in different forms, in different media, people fall far from being rational. This is precisely why we want to abstract away from these individual differences of sensibility and perception. One of the reasons why you ought to take the recommendations of your ideal advisor as normatively salient is that your ideal advisor is free from your idiosyncratic susceptibility of perception, vividness of imagination, and sensibility to representation. The point of the theory is to filter out these effects, and the assumption of improved cognitive and imaginative capacities is meant to do precisely this. If it is harder for you to represent information vividly than it is for me, our ideal advisors nonetheless should be equals in this respect.

The problem of representation targets the epistemic constraint. It holds that variations of representation of information affect the way we perceive it and the way it influences motivation. But even if this assumption is warranted, it provides no objection to the ideal advisor theory. It is no objection because as long as the counterargument concedes that it is the selfsame piece of information that

is conveyed, it does not make a difference. The point of idealization, among other things, is to get rid of the baggage different representations may carry with them. Ideal advisors will not exhibit individual shortcomings of perceiving, representing, and being sensitive to different extents to the selfsame piece of information. On the other hand, if the counterargument comes with the assumption that identical pieces of information are in fact *different* pieces of information when conveyed in different media—that there are no identity criteria for telling when different representations purport to convey the selfsame piece of information—then the argument is trivially off-target, since then ideal advisors will just gather all these different representations.

So it is not a good objection that indeterminacy in the idealization process arises because actual people have certain shortcomings for representing information, or because what it is for some representations to carry the same piece of information is radically indeterminate. But perhaps giving coherence to the notion of being fully informed is more than just a matter of overcoming these shortcomings. Perhaps there are special additional conditions for information to "sink in," or register: perhaps it is not enough just to *provide* information. The person must "appreciate" the relevance of information in order for it to have the desirable impact. Connie S. Rosati (1995) raises this point as the *problem of appreciation*, and argues that what lies behind it is "what it is like to be a particular person" (1995:307). Whether any fact is "informing for" a person depends on that person's psychological and intellectual make-up. Different pieces of information can have different impacts on different people, and some information can be informing, or can register, only if the person undergoes radical changes, and experiences different kinds of life. In short, we all have our own "perspective," and what can be informing for us depends on our perspective.

Rosati has in mind, primarily, "fundamental" choices, like those of choosing a career, or contemplating a major change of lifestyle. In such fundamental choices, our choice is both influenced by our previous choices, and it will influence our subsequent choices. The effects are temporally extended. So we should have full information of all the different resulting lives in order to form preferences between them, but these lives all involve their particular perspective. It would surely surpass conceivable human capacities to keep in mind all the relevant information of all the relevant experiences, from all the relevant perspectives.

The problem of appreciation is exacerbated if we consider the distinction made by David Sobel (1994) between the *report model* and the *experiential model* of representing information and comparing options. On the former, the ideal advisor does not directly experience different possible lives: she is only told what they are like and what they feel like. On the latter, she is not denied direct experience, and has first-hand acquaintance with these lives.

The report model does not provide the ideal advisor with enough, or accu-

rate enough, information to form preferences among the options. This is because "some experiences are *revelatory* in the sense that they alter our responses to facts and descriptions" (1994:797, emphasis added). This model cannot incorporate information about such experiences, so it cannot incorporate everything that may be relevant. The experiential model suffers from comparable problems. In this model, the ideal advisor lives through all the relevant lives. But she cannot experience these different lives independently if she serially moves through them: having lived one type of life will influence how she experiences the living of another type of life. That is, the way she experiences and evaluates these lives will depend on factors like the *sequence* of experiencing them. If that is the case, ideal advisors have to experience all possible lives, in all possible orders.

But even then, the differences in the way information can register for each of the persons in each of the lives may cause a further difficulty: the chooser would have to retain those features from each of the lives that allowed her to appreciate information pertaining to that life, but these features might mutually be incompatible with one another. This could happen especially if radically different lives would have to be experienced, for it is possible that one would be unable to appreciate, say, life as a nun, after having tried life as a Marxist atheist (Rosati, 1995:315). Since being a person involves having a particular perspective, it may be impossible to experience other lives together with the particular perspective those other lives involve. If so, being fully informed is conceptually impossible.

In order to assess this counterargument, return to the starting point. Suppose now we are concerned only with fundamental choices. Because of the characteristics of such choices, there are not many options at issue. After all, we do not choose from dozens of careers or ways of life; we usually have just a few to consider. This is precisely because we bring to these choices the effects of all the previous choices we have made—if you like, we bring our specific perspective. Given who I am, that is, a set of facts about me that idealization takes into account, many ways of life are simply not relevant. For this reason, some options, although actually feasible, do not enter my deliberation about my fundamental choices. I could decide to go to the jungles of the Amazon as a missionary, but I most certainly would not. This option is quite irrelevant to my deliberation. Given all the data about me, it is hard for me to imagine how *that* would be the form of life my ideal advisor would advise me to adopt. Hence it does not seem inevitable that he would need to experience that life.

What's more, fundamental choices are not made for once and all. We normally do not decide on square one what we attempt to do with our lives. That is, there is no unique stage of development when the idealization would have to be carried out.[7] We also change throughout our lives, and that includes changing our perspec-

[7]For some reason, Rosati thinks there would have to be such a stage (1995:310).

tive. That is, if we are at least fairly malleable beings—and surely we are—then it is unreasonable to suppose that we possess a fixed perspective. As we move through life, not only our goals and plans, but our perspective changes as well, and it changes at least partly in response to the choices we make.

Still, if having even a "flexible" and changing perspective is incompatible with fully experiencing another life, then the task of ideal advisors would indeed be difficult. They would have to experience all the options, with all their perspectives, stepping outside of incompatible ones, and then stepping back, uniting all the various experiences in their mind or memory. Thus, they may not be able to evaluate all the alternatives consistently.[8]

In different ways, the point Sobel and Rosati make against ideal advisor theories is that even if ideal advisors can experience all possible lives, in all possible orders, they cannot compare those lives due to the specific *perspectives* involved in those lives. That is, even if ideal advisors travel a myriad possible worlds, they cannot become fully informed, for each life involves a unique perspective.

I think, however, this argument badly backfires. There are reasons to doubt that perspectives are fixed—and even if they were, they would be taken into account in the idealization process as another relevant fact about the person. Therefore, ideal advisors do not face a problem due to incompatible perspectives—because the actual person already *has* a perspective. If having a perspective plays such a central role in human psychology, then any life or experience with an incompatible perspective is *infeasible*. It is irrelevant to what you have reason to do, since you couldn't do whatever it would involve doing, and it cannot be good for you, for it, as it were, cannot *be* for you.

But perhaps I misrepresent their objection. This is not easy to decide, since Sobel and Rosati tell us precious little about what a perspective is. Worse yet, sometimes they speak about what a life "feels like." I must admit I have not known before that lives are supposed to feel like anything.[9] Sometimes, Rosati equates a

[8]In the course of making her argument, Rosati says some extraordinary things. For instance, "if a person must have certain traits in order to experience something in a certain way, it seems she must also have those traits (or at least ones which are similar enough to allow her access to the same information) in order to remember what it is like to experience that thing in that way. In order for the obtuse person to be fully informed about her life as a sympathetic person, she must take on those qualities and have the requisite experiences" (1995:320). The first sentence is about whether it is possible to remember what it was like having another perspective, having had that perspective. Surely it is. I might have been very selfish and self-centered as an adolescent. As a more mature person now, I surely don't have to become selfish and self-centered again to recall my experiences as a selfish and self-centered person. The second sentence denies we can, as it were, place ourselves into the shoes of others without becoming like those others. But surely we can do that too. If we couldn't, for one thing, it would be practically impossible to be a sympathetic person.

[9]On a more sober note, preference satisfaction theorists of well-being do not care about how experiences and lives may feel. Hedonists do. But an ideal advisor theory holds that what is good for you is the satisfaction of the preferences you would have in some epistemicly and cognitively ideal

person's perspective with her psychological and intellectual make-up (I have used the concept in this sense so far). Other times, a perspective seems to be a collection of personality traits. The idea is, then, that being fully informed is incompatible with having these traits. Once again, though, these traits are just facts about the individual, and they will be taken into account in idealization.

Or perhaps personality traits cause some other difficulty. Perhaps the point is that ideal advisor theories fail to take into account that people make choices from specific circumstances, and it is not impossible that the idealization process fails to preserve the person's unique outlook on her circumstances and options. I have already mentioned this worry in connection to Railton's theory on page 89. This may be called the *alienation problem*: a person may fail to take the recommendations of her ideal advisor as authoritative because those recommendations may be incompatible with her "perspective."

On this interpretation, the counterarguments target the *normative adequacy* of ideal advisor views:

> What a given individual's motivational system will be like once she is fully informed will depend upon what it was like before she was fully informed and how it has changed as a result of idealization. In order for us to be sure that we can regard the fully informed individual as authoritative, we must have a conception of what it would be for an individual's motivational system to change for the better, and thereby a more substantive conception of an ideal advisor—one that incorporates an ideal of the person. (Rosati, 1995:312)

There are at least two possible readings of this quote. On the one hand, the point might be that in order to build an ideal advisor theory, you need a conception of what it is to be ideally placed to be an *advisor*. If this is the correct reading, the argument is circular. Ideal advisor theories give carefully worked out conceptions of that. For versions with the internalist requirement, the ideally placed advisor is your own fully informed and rational self, whose preferences are adequately connected to your actual preferences. For versions with the convergence requirement, the ideally placed advisor is your own fully informed and rational self, whose recommendations are shared by similar, ideally placed advisors of other persons. Before the argument can be made, it must be shown that these conceptions are deficient.

On the other hand, the point might be this. Suppose a person is placed in ideal conditions. There she is likely to have different preferences than those she actually has. If so, these preferences determine what, in her actual conditions, is good for her. Once "back" in her actual conditions, however, she may not accept

condition for forming preferences, and the satisfaction of your preferences does not have to result in any feeling.

the recommendations based on the preferences of her ideal advisor. She may think what her ideal advisor prefers does not tell her what she ought to do when she aims to do what is best for her. She, of course, may be irrational to think that. But it is also possible that she cannot be judged irrational: she does not deserve blame for rejecting the recommendations of her ideal counterpart. She may be not unjustified to reject the authority of the ideal advisor: she may not recognize her as having anything to do with herself. In Rosati's words, "the 'fully informed' person, though purportedly you, may not be someone whose judgments you would recognize as authoritative; thus, Ideal Advisor views lack normative force" (1995:299).

The reason the actual person may be "alienated" from the ideal advisor in terms of accepting her advice as authoritative for herself in her actual circumstances is that the process of idealization may induce radical changes, removing the ideal advisor too far from the person. Thus, in an ideal advisor theory with the internalist requirement, there may be a "missing link" in the chain between the preferences (or motivating reasons) of the person, and the preferences (or motivating reasons) of the ideal advisor.

Just as in the case of the concept of a perspective, I find it very hard to evaluate this argument. Part of the reason is that the concept of internalism is itself highly ambiguous and controversial. But perhaps the problem of alienation can ultimately be developed into a good argument against ideal advisor views—but it is unclear whether it would apply to versions with the convergence requirement. To be on the safe side, preference satisfaction theorists of well-being can opt for an ideal advisor theory with the convergence requirement, like (a suitably modified version of) those of Smith's and Harsanyi's. Or perhaps they can try and modify their theory in anticipation of the argument. The next section examines a proposal for this.

7.4 The Concept of Integrity

Robert Noggle (1999) introduces the concept of *integrity* to defend internalism. A person's integrity is constituted by her central projects, commitments, character traits, and the like, which have an essential role in constituting the person she is. Noggle's proposal is that an *integrity requirement* serves as a constraint on the extent of idealization to avoid the problem of alienation. That is, the preferences of the ideal advisor must be arrived at a way that is "integrity-preserving" for the person, otherwise she may justifiably complain that her ideal counterpart ignores, in forming her preferences, something that is essential about her. Noggle characterizes this new requirement the following way:

> a hypothetical situation preserves a person's *integrity* if and only if it preserves those concerns, attitudes, and other mental states that are constitutive of her identity or her self. (1999:314–5, his emphasis)

Which particular concerns, attitudes, and mental states are constitutive of a person's identity? It seems a plausible reply that this should be left to the jurisprudence of the person: she is to ultimately decide how many of these can be taken away from her without losing her identity.

The integrity requirement also gives a reply to the problem of appreciation. Recall that the problem was one of experiencing different and incompatible lives, whose evaluation cannot be united in a single consciousness: it is impossible for the ideal advisor to have first-hand experience of all conceivably possible lives of the person, while retaining all those aspects that may be relevant to the appreciation of these lives. Now, if the ideal advisor is constrained by the integrity of the person, she does not have to travel through many possible worlds—she only has to wander around neighboring ones. The ideal advisor is also constrained temporarily, since how far she is to go into the future and how many "time-branches" she has to travel for collecting information is also a function of the integrity of her counterpart.

Nevertheless, there is a fundamental problem with the integrity requirement. A person's identity is defined, at least partly, in terms of her central concerns. But in establishing what is good for the person, we are also interested in these central concerns. Any recommendation that is not integrity-preserving may be rejected by the person, but what warrants that the elements which are constitutive of her identity are relevant to her well-being? That is, the problem with Noggle's account is that identity-constituting central projects do not admit of criticism. Here is an example.[10] Suppose I am a mafia leader. Members of the rival gang have murdered my brother. I make *vendetta* my central concern in life. On the integrity account, my central project cannot be criticized. My ideal advisor, if he wants his recommendations to move me, must accept my thirst for revenge as part of my identity. That is, the advice he gives me I may reject by appealing to what constitutes my identity. Nevertheless, it seems that I can be wrong in doing that. The integrity requirement is overly restrictive, and it is defective on the most tender spot—criticism of the central projects or goals of a person.

As I said above, I find that it is not easy to evaluate the objections to the ideal advisor theory that I have discussed in this chapter. They carry background assumptions which are controversial, ambiguous to interpret, or rest upon answers to unsettled problems external to idealization theories. The problems of internalism, identity, and what it is to have a perspective touch upon broader issues within philosophy, and whether the arguments based on them are ultimately successful will be decided, so to speak, in a different ballpark. At this point, however, it seems that the ideal advisor theory has the resources to counter the objections marshaled against it.

[10] I was challenged with this example by James Griffin.

In light of this, if one wants to argue against the ideal advisor theory, it may be worth its while to look for some argument with less controversial background assumptions—that is, of course, if anything can be less controversial in philosophy. In the next chapter, I first present my favored interpretation of the ideal advisor theory, then I attempt such an argument against this theory. Ultimately, however, I think that my argument, rather than refuting that version of the ideal advisor theory, gives an opportunity to revise it. The revision is the theory of well-being I propose.

Chapter 8

Why You Shouldn't Listen to Your Ideal Advisor

8.1 IRP

Let me take stock at this point. Theories of well-being have traditionally been grouped into *subjective* and *objective* accounts. One way of drawing the distinction between these is in terms of *preference*. Thus, on a subjective theory, what is good for you is what you prefer: some thing, x, contributes to your well-being if and only if you prefer x. In contrast, on an objective theory, what is good for you is independent of your preferences: x contributes to your well-being in virtue of something else. It is claimed that this division has the advantage of providing an exhaustive and mutually exclusive classification of theories of well-being.

Subjective theories are intuitively attractive, and they have been very influential. It has also been recognized, however, that they need to be modified: the relevant preferences cannot be your *actual* or given preferences, since it is possible that you prefer what does not turn out to promote your well-being. Furthermore, it also seems possible that something you do not prefer can be good for you. For these reasons, many philosophers agree that actual preferences cannot be sufficient for determining what promotes well-being. But subjectivists insist that even if preferring does not provide a sufficient condition, it must provide a necessary one. That is, they concede that since people can be mistaken about the sources of their well-being, actual preferences do not determine what is good for them; but, they claim, if people were in ideal conditions for preferring, their preferences under these conditions would indeed determine what promotes their well-being.

I called this type of theory the idealization theory of well-being on page 85. There are many versions of this type of view, depending on how they construct the ideal conditions. On its most popular versions, the ideal conditions are *epistemic* and *cognitive*: the person in ideal conditions for preferring forms adequately informed and appropriately reasoned preferences. For instance, she is informed about her circumstances, range of options, and the possible consequences of her choices, and she does not make any mistakes of representation of facts or errors of reasoning when she evaluates her circumstances and options. I called such accounts the ideal advisor theory. This theory, although usually couched in terms of desire rather than preference, is currently perhaps the most popular theory of well-being in ethics. It is also prevalent in more formal approaches to ethics and in welfare economics.

Since the epistemic and cognitive conditions can be interpreted in different

ways, the ideal advisor theory has further sub-versions. I presented some of these in Section 7.2. On my favored interpretation, the person in ideal conditions for preferring is *fully informed* and *ideally rational*. I will refer to this theory with the abbreviation IRP (for "informed and rational preference"):

(IRP) One thing, x, is at least as good for person i as another, y, if and only if were i fully informed and ideally rational, i would weakly prefer x to y.

The aim of this chapter is to present an argument *against* IRP. In Section 8.2, I give what seems to me the best interpretation of the "full information" and "ideal rationality" conditions. My argument is presented in Section 8.3. It targets the idea that in order to determine what promotes the well-being of a person, it is sufficient to establish what the person would prefer if she was fully informed and ideally rational. I argue that either full information and ideal rationality are not sufficient to determine the preferences of the person placed in ideal conditions, or the person placed in ideal conditions might not prefer what is better for her. Thus, in the former case, it may be underdetermined what you would prefer if you were fully informed and ideally rational; in the latter case, what you would prefer if you were fully informed and ideally rational would possibly not promote your well-being. If I am right, IRP needs either to be rejected, or to be revised. I conclude in Section 8.4 by suggesting how the theory could be revised. The main advantage of the revision is that the revised IRP turns out to be a theory that is faithful both to the subjectivist and the objectivist intuitions (see page 5). However, my revision makes IRP a partly objective account of well-being, hence I suggest that the distinction between subjective and objective theories of well-being cannot be mutually exclusive. The remaining chapters of this work further explore the revised IRP.

8.2 What Is It Like to Be an Ideal Advisor?

On the ideal advisor theory, a person's well-being is constituted by the satisfaction of that person's *hypothetical* preferences. The theory involves specifying a number of *counterfactuals* about preferences: it takes the person's actual preferences and establishes how these preferences would change if the person was given information pertaining to her situation—factual knowledge of the alternatives, possible consequences of her choices, and the like—given that the person does not make any mistake of reasoning and avoids other sorts of cognitive error. It is usually assumed—and I will also assume this—that the counterfactuals about the person's preference changes can be evaluated on some best theory for the evaluation of counterfactuals, whatever that theory is. By evaluating these counterfactuals, we determine what the person would prefer if she was adequately informed and reasoned appropriately, and the satisfaction of these preferences promotes the person's

well-being. On IRP, these preferences are the preferences of the fully informed and ideally rational "ideal advisor" of the person.

Ideal advisors fulfill both the epistemic and the cognitive conditions. In other words, an ideal advisor theory places epistemic and cognitive *constraints* on the preferences whose satisfaction is relevant to well-being. But how should we interpret these constraints? Consider first the epistemic constraint. On its most popular reading, being "fully informed" means that *all relevant* information is available to ideal advisors. Any piece of information is relevant which could make a difference to the preferences of the person in the idealization process; and all such information should be made available, since the recommendations based on the preferences of the ideal advisor would have less normative force if she was to work with limited information only. Restricting the information accessible to the ideal advisor would introduce the possibility of error into her preference formation.

Nevertheless, ideal advisors should not have *too much* information. An ideal advisor knows what options are open to her actual counterpart, the relevant features of the choice situation, and the probabilities with which the possible outcomes might obtain—so she knows the objective, *a priori* probabilities involved in the choice situation of the actual person, and, since she is ideally rational, she forms and handles subjective probabilities appropriately, when objective probabilities of some options cannot be obtained.

That is, ideal advisors cannot be omniscient. They do not have certitude of what *will* happen given their actual person's choice; they only know what *is likely to* happen, given that choice. IRP claims that something is good for us in virtue of our preferring it in ideal conditions for preferring; but fully informed and ideally rational preferences ought to be preferences *for* the actual person, taking into account the limitations of actual persons. Otherwise the theory would only tell us what would promote the well-being of omniscient beings—something we are not interested in. Rather, we are interested in what would promote *our* well-being, and in how to weigh things which promote our well-being in our lives, given our limited time, resources, and the uncertainties we have concerning the outcomes of our actions and the influence of the choices of others. A theory of omniscient ideal advisors would be a useless device to determine what would promote the well-being of actual persons. Therefore, ideal advisors need to be short of omniscient, but they need to be knowledgeable of our alternatives, their outcomes, and the probabilities of these. Accordingly, no version of the ideal advisor theory that I am familiar with supposes that the ideal advisor is omniscient in the sense I am using the term.

Consider now the cognitive capacities of ideal advisors. IRP treats ideal advisors as ideally rational; but it is controversial what rationality is, and what it is to be ideally rational. For ideal advisors, rationality may consist in forming rational preferences, and representing and processing information appropriately. Or it may also consist in some further cognitive capacities. I propose therefore to make the

following distinction. The cognitive capacities of ideal advisors include that

(a) they form their preferences according to the canons of a fully developed theory of rational choice;
(b) in addition, they have further cognitive capacities.

By a "fully developed theory of rational choice," I mean a formal theory that tells rational agents how to order their preferences under conditions of certainty, uncertainty, and risk. This theory also involves norms for handling and updating probabilities. I will call such a complete theory of rational choice "$\mathcal{R}$" for short. Needless to say, we do not now have such a fully developed theory, but rather we have a number of competing theories. Nevertheless, an intuitive idea of what such a formal theory would look like in broad outline is this: $\mathcal{R}$ tells a rational agent how to solve decision problems, including problems in which her choice may be influenced by states of nature or the consequences of the choices of other agents, and it tells her only this much. In contrast, by "further cognitive capacities," I mean cognitive capacities that are not part of that theory, even though they are necessary for an agent to have in order to be able to employ that theory. These further cognitive capacities are needed by the agent to be able to understand her situation—to describe and represent the options and possible strategies, the influences of the choices of other players, and so on. In short, (a) enables a rational agent to make a choice, while (b) enables an agent to understand what is involved in making that choice.

Therefore, we can think of the difference between (a) and (b) this way. The former says that rational advisors follow the norms of rationality, hence their preferences must be mathematically representable on theory $\mathcal{R}$. The latter describes what capacities an agent must have to count as rational—what it takes to be able to follow the norms of rationality and to have preferences representable by $\mathcal{R}$. In what follows, I am going to bracket (b). In order to make my case against IRP, all I need to suppose is that ideal advisors are ideally rational in the sense given by (a)—that they follow the norms, and their preferences satisfy the axioms, of $\mathcal{R}$.[1]

8.3 A Conspiracy Against Ideal Advisors

Consider the following example. I am faced with the choice of what career to pursue in my life. For simplicity, I assume that only my success in my chosen career

[1]The ideal advisor theory has recently received a lot of criticism—some of these were discussed in Section 7.3. Note that all of those objections targeted (b). Moreover, I concluded that the objections either raise no special difficulty for the theory, or they rely on background assumptions which are not as uncontroversial as they might initially seem. My hope is that by relying exclusively on (a), I can avoid making some of the assumptions which seem unsubstantiated or question-begging to me.

determines how well my life goes. Now suppose that due to my circumstances, inclinations, and talents, the two relevant options open for me are becoming a philosopher or becoming a concert pianist. In order to decide which of these would be better for me, I turn to my ideal advisor. My ideal advisor knows the following. I have talent for both pursuing an academic career in philosophy and a career in the performing arts as a pianist. But he also knows that my talent for philosophy is somewhat modest: I can become a reasonably successful, average philosopher, and therefore have a reasonably good life. If, on the other hand, I pursue a career in music, I have the ability to become an exceptionally good pianist, and have an immensely rewarding life.

There is, however, a problem. If I do decide to pursue the career in music, there is a high likelihood that I will develop rheumatoid arthritis in my fingers in a few years—which will destroy my career completely, and I will end up with a miserable life. My ideal advisor knows that the probability that I develop this condition after a few years is in fact 0.9—since he fulfills the epistemic condition, that is, has full information of the possible consequences of my choice and the relevant probabilities. As I suppose now, he also uses a fully developed theory of rational choice, $\mathcal{R}$, to form his preferences over what I should prefer and choose. In other words, he knows that the decision problem I face is the one depicted on Figure 8.1 on the next page.

The numbers 0, 1, and 10 represent, on a ratio scale, my well-being: how well my life goes overall if I choose to become a philosopher or a concert pianist.[2] Node I shows my move, and node N shows Nature's "move." If I move "down," that is, become a philosopher, my life will be alright, although not great. If, on the other hand, I move "across," it is Nature's move—"she" will either move down, with the consequence that I develop rheumatoid arthritis, or she will move across, in which case I do not develop the condition. There is a 0.9 probability that Nature moves down, and a 0.1 probability that she moves across. If she moves down, my career is ruined, and my life will be miserable. Should she, however, move across, my life will be extraordinarily good. Note that my *expectations* of well-being are equal in the two prospects I face. If I move down, I realize a life with 1 "unit" of well-being. If I move across, my expectation of well-being is $(0.9 \times 0 + 0.1 \times 10 =)$ 1 as well.

The preference my ideal advisor settles on constitutes which prospect is better for me—whether it is the *sure prospect* I can choose by moving down (that is, by becoming a philosopher), or the *lottery prospect* I can choose by moving across (that is, trying to become a concert pianist). So what will his recommendation be?

Now, we know that my ideal advisor forms his preference based on the principles and axioms of $\mathcal{R}$. But in order to form his preference concerning the prospects

[2]We can think of these numbers as indices of income, Rawlsian primary goods, vectors of capabilities, quality-adjusted life years or some other quality of life index, and so on.

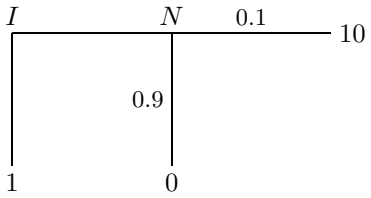

Figure 8.1

I am facing, he has to use some principle of $\mathcal{R}$ that tells him how to form his preferences when I have to choose between a sure and a lottery prospect with the same expected values. This will be given by a principle that tells him what risk-attitude he ought to take towards well-being—more precisely, this will be settled by some *principle of reasonable levels of risk-taking towards well-being*, which I will abbreviate as P.[3]

Let me ask the following question: will principle P be a part of a fully developed theory of rational choice, $\mathcal{R}$, and what are the consequences of its inclusion or omission for IRP?

First, suppose that P is *not* part of $\mathcal{R}$. In this case, my ideal advisor will not be able to form a preference in cases when I face prospects with equal expected values. Since nothing tells him which prospect he ought to prefer, he cannot give recommendations to the actual person. He cannot give recommendations since he cannot compare the prospects the actual person has to choose from with $\mathcal{R}$. And he cannot say that the prospects are equally good since their expected values are equal. That is, he is not indifferent between the prospects, since that would presuppose a principle P: that you ought to be indifferent between prospects with equal expected values.

It might be objected that I may already have some risk-attitude towards these prospects, and it will be taken into account as just another fact about me in the idealization process. That is, my ideal advisor has the same risk-attitude towards well-being as I do. But when we want to assess our preferences, we want to assess preferences over such prospects as well. What I am asking my ideal advisor to do in this case is precisely to tell me whether my preference based on my risk-attitude would be one I could embrace in ideal conditions, and, if not, what sort of risk-attitude I ought to have when forming a preference over such prospects. Consequently, if P is not part of $\mathcal{R}$, IRP is underdetermined: when the actual person has to choose from risky prospects, the theory does not specify what the person

[3]Actually, the expected values of the prospects do not have to be equal: principles for reasonable levels of risk-taking may be relevant even if these values are unequal. I use the simplest case only for purposes of illustration.

would prefer were she fully informed and ideally rational. Hence, the theory fails to specify what would promote the person's well-being.

In order to avoid this problem, it is natural to assume that P is part of $\mathcal{R}$. Thus, in the remainder of this section, I test the hypothesis that P is part of $\mathcal{R}$ for different versions of P. I argue that if it is, then ideal advisors will have preferences whose satisfaction does not promote the well-being of their actual counterparts. I show this by arguing that if P is part of $\mathcal{R}$, then actual persons might reject, with good reason, the recommendations based on the preferences of their ideal advisors.

I will assume, for now, that if P is part of $\mathcal{R}$, then it can be any of three simple principles. (I will discuss the possibility of more refined principles later.) P might tell rational agents to be risk-averse towards well-being. Or it might tell rational agents to be risk-neutral towards well-being. Finally, it could tell rational agents to be risk-seeking towards well-being. However, I only mention this last possibility to discard it at the outset. I suspect that it would be quite extraordinary from our ideal advisors to tell us to take risks comprehensively. It is hard to see how a principle to seek risk could be a principle of rationality. Consider again the choice I have between becoming a philosopher and pursuing a risky career as a concert pianist. Suppose now I have the talent of a genius for playing the piano. If I do not develop rheumatoid arthritis, I will not only become a great concert pianist, but I will be the greatest concert pianist of the time: a talent like me is born only once in a century. Unfortunately, I am even more likely, on this scenario, to develop the condition in my fingers. Suppose the probability of this is 0.999 now. There is, however, a very low—0.001—probability that I do not develop the condition, and my life will be exceptional: its value will be, not 10, but 1,000. The expectations of becoming a philosopher and risking the career in music are again equal. It nonetheless seems, given the extremely low likelihood of pursuing the concert pianist career successfully, that my ideal advisor would not recommend to take that risk. But, in any case, I will give a general argument against *any* principle later.

Let me now consider the remaining two cases. Suppose, first, that principle P of $\mathcal{R}$ tells rational agents to be risk-averse towards well-being. In particular, it tells rational agents that when they are faced with sure prospects and lottery prospects of the same value, they ought to prefer and choose the sure prospect—in other words, rational agents play it safe.

I will, once again, argue through an example. In the example of the choice between becoming a philosopher or a concert pianist, the outcome of the choice of pursuing the latter was influenced by factors outside of my control—by the state that may result following a "move" by Nature. But our choices are not influenced by states of nature only. They may also be influenced by the consequences of the choices other people make. Our ideal advisors, when giving us advice for what would be better for us, must take these influences into account as well.

Look at Figure 8.2 on page 109 now. In this situation, there are two individuals,

A and B. I suppose A is female and B is male. The first number at the endpoints stands for the value of the outcome for A, and the second number stands for the value of the outcome for B. Thus, A has a choice at node A: she can either move down, in which case she receives 1 unit of well-being and B receives 0; or she can move across. If A moves across, there is a 0.5 probability that she will receive 0 and B will also receive 0, but there is also a 0.5 probability that B now gets to make a choice: he can also move down or across. (What p and $(1-p)$ stand for will become clear later.) If B moves down, A receives 0 and B receives 3 units of well-being. If, on the other hand, he moves across, there is, once again, a 0.5 probability that they will both receive 2, and it is equally likely that A receives 2 and B receives 4.

I will assume that initially both A and B take the recommendations of their respective ideal advisors as authoritative—that they recognize the preferences of their ideal advisors as reason-giving—and they mutually know that they do. Furthermore, their ideal advisors form their preferences according to the canons of $\mathcal{R}$, and $\mathcal{R}$ contains P: a principle that tells rational agents to be risk-averse towards well-being. What will the recommendations of the ideal advisors be?

Look at the situation from the perspective of B first. His ideal advisor reasons that if B moves down at his node, he will receive 3 for certain; if he moves across, he can receive either 2 or 4 with equal probabilities. The expectations of these two prospects are equal. But B's ideal advisor follows P, which says that you ought to be risk-averse towards well-being. Hence, if B ever gets a chance to make a move, he ought to move down.

Consider A now, who will definitely have a chance to make a move. She can move down or across. If she moves down, she will get 1 for sure. In order to find out whether she ought to move across, she will reason the following way:

> If I move across, Nature will either "move" down or across. If Nature moves down, I receive 0. If Nature moves across, B will make a move. He will either move down or across. If he moves down, I again receive 0. If he moves across, Nature will move again, but that move is irrelevant, since no matter what happens, I receive 2. So what I can expect if I move across partly depends on B. Suppose B moves across with probability p, and he moves down with probability $(1-p)$. My expectation if I move across therefore is:
>
> $$0.5 \times 0 + 0.5\left((1-p) \times 0 + p(0.5 \times 2 + 0.5 \times 2)\right) = p.$$
>
> So whether I ought to move across depends on what B is likely to do, whether he is willing to move across. But I know that he will take the preference of his ideal advisor as the reason for his move. And I also know that his ideal advisor forms his preference according to a principle of risk-aversion towards well-being, that is, he will prefer him to move down, should he get a chance to move. Hence I know that he would move down, that is, I know that $p = 0$. So I ought to move down myself.

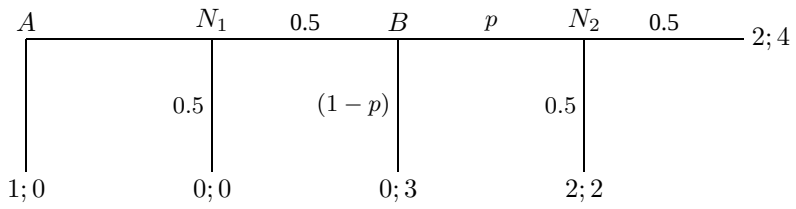

Figure 8.2

This assumes, of course, that the preferences of the ideal advisors, as well as the reasons for those preferences, are known by the actual persons. This assumption enables us to check whether A and B can endorse the preferences of their ideal counterparts. The argument I make is that they have reasons not to. They have reasons to "conspire" against the recommendations of their ideal advisors.

For return to B. He realizes that if they both act in accordance with the preferences of their ideal advisors, he will never have a chance to move. But getting a chance to move would be at least as good for him as not getting a chance to move. That is, if only he got a chance to move—even if he *actually* could not move because Nature moved down at N_1—he would not be worse off, and possibly he could end up being much better off. In short, he would not be worse off if A moved across, irrespective of what happens afterward. And B starts to think now, and comes up with an idea.

Suppose A and B can communicate, and do it without unduly high costs. Then B can make the following offer to A: "I promise to move across if I get a chance to make a move." B has nothing to lose with promising this, since if A accepts the offer, he might end up better off, and if she rejects it, he ends up no worse off. The idea behind the offer is that by cooperating in ways not embraced by their ideal advisors, they might fare better than by strictly following the recommendations of their ideal advisors.

When B makes his offer, he promises that he will not act in accordance with the preferences of his ideal advisor. In effect, he promises that at his move he will *not* be risk-averse towards well-being. In other words, he promises to reject the reasoning based on P. Instead of being risk-averse, he becomes risk-seeking, and he makes it the case that $p = 1$. We can think of B's offer as choosing a *risk-disposition* towards well-being at the start of the choice problem: if he makes the offer, he promises to become risk-seeking, and if he declines to make the offer, he remains risk-averse. Similarly, we can think of A's decision whether to accept the offer as choosing a risk-disposition which determines which way she moves at node A: on the one hand, if she accepts the offer, she becomes risk-seeking towards well-being and moves across; on the other hand, if she rejects the offer, she becomes risk-averse towards well-being and moves down.

In order to model the offer, a move by B may be inserted before node A in Figure 8.2, representing B's making the offer or declining to make the offer.[4] For the sake of the argument, assume that the agents are *transparent*, that is, their risk-dispositions are known with certainty. Of course, in many situations agents are not transparent, thus their risk-dispositions are not known with certainty. In such cases, whether A can accept the offer depends on how she evaluates the risk of accepting it, given her probability assessment of B's risk-disposition—thus, whether she accepts the offer depends on the *degree* to which she is willing to become risk-seeking. However, at least in this case B has a reason to become transparent—as a way of assuring A that the promise of moving across at his move will be honored.

Assume also that B's offer to commit himself to be a risk-seeker is credible: once he chooses his risk-disposition, he sticks with it, and he does move across at node B. Of course, prior commitments are not always credible. When the time for action comes, agents may find that they are better off breaking a prior promise. However, at least in this case, once B has chosen his risk-seeking disposition, he has no obvious reason to change it later. In other words, we may suppose that B's risk-disposition is stable, in which case A can count on B to move across at node B. If risk-dispositions are less than perfectly stable, whether A can accept the offer depends on how she evaluates the risk of accepting it, given her probability assessment of B's stability of risk-disposition—thus, whether she accepts the offer once again depends on the *degree* to which she is willing to become risk-seeking. Hence, B has a reason to develop a stable risk-disposition.

But should A accept the offer after all? If she moves down, she will get 1 for sure. If she moves across, she also expects 1, since now she knows that B will move across—that $p = 1$. Why would she reject the offer? One consideration is that her ideal advisor, who is ideally rational in the sense given by $\mathcal{R}$, tells her to be risk-averse towards well-being, so she should still move down at node A, regardless of B's promise. On the other hand, however, this consideration is relevant only if she continues to believe that P is a principle of rationality. But B now rejects P, and for sound reasons. With the offer, her situation has changed. So should she now listen to her ideal advisor, or to B? Figure 8.3 on the next page illustrates how her original choice problem is simplified, given that the offer is made.[5]

In a way, B's offer is quite "conspirative," since it requires that the persons cooperate by harmonizing their risk-attitudes towards well-being in ways not embraced by (and not open to) their ideal advisors. The offer works only if B gives

[4] The subsequent branches are the same on both branches leading from this node representing B's opportunity to make the offer. The difference is in the *preferences* that B forms over the prospects at node B, given the choice of his risk-disposition at this initial node.

[5] Note that Figure 8.3 shows only her perspective of the choice problem—assuming that B is transparent and his risk-disposition is stable—with the probabilities and payoffs relevant to deciding whether she should accept the offer (move across) or reject it (move down).

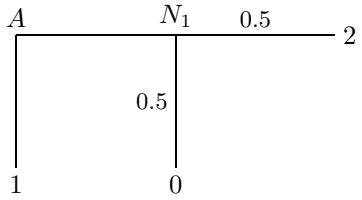

Figure 8.3

up the recommendation based on the preferences of his ideal advisor by rejecting P, and A too gives up the recommendation based on the preferences of her ideal advisor, also rejecting P. The offer requires that they both become risk-seekers—that they both become irrational in light of $\mathcal{R}$. If they do, B can end up better off: once A moves across and if he gets lucky, he is guaranteed a payoff of at least 2, and, with some further luck, 4. Arguably, A is no worse off, since the expectations of moving down and moving across are now equal, and she might now have a reason to reject P, becoming a risk-seeker.[6]

Their ideal advisors, in contrast, cannot cooperate in the same way. They cannot transform their situation in order to open up the possibility of realizing higher gains. Since ideal advisors, by definition, are "in the grip" of their rationality, and their rationality, by hypothesis, prescribes risk-aversion towards well-being, their preferences cannot yield the recommendation to cooperate. A and B will realize this, since their ideal advisors are transparently risk-averse and their risk-dispositions are fixed.

B has a reason to reject the recommendation of his ideal advisor and make this known to A. His reason is that if he follows that recommendation, he forgoes benefits he might otherwise be able to obtain through cooperation. His ideal advisor, since he is ideally rational and this is known, cannot credibly commit himself to move across at node B. Thus, his offer would not be accepted. Actual persons therefore may be able to come to agreements which would be foreclosed to them if they were ideally rational.

A has a reason to reject the recommendation of her ideal advisor because she may fail to see why she ought to be risk-averse towards well-being in a situation like the one depicted on Figure 8.3. She notices that there are many situations in which it is better not to follow the recommendation given by $\mathcal{R}$ (B is in such a situation), and she may start to wonder why P should be considered as a norm of rationality at all—or, if it is a norm of rationality, why a norm of rationality should

[6]Actually, she might play a mixed strategy by tossing a fair coin to decide whether to move down or across, given B's offer. But that also means that she rejects P, since now she has become risk-neutral towards well-being.

determine what is better for her in such situations.

Notice that the argument is not that ideal advisors can *never* employ commitment and incentive mechanisms to come to advantageous agreements. Rather, the argument is that in virtue of their ideal rationality, certain mechanisms whose use can make people better off are foreclosed to them. Therefore, their preferences may fail to determine what would make less than ideally rational persons better off.

Finally, let me ask what happens if principle P prescribes risk-neutrality towards well-being instead of risk-aversion. Actually, nothing changes in this case. A and B may similarly reject the recommendations of their ideal advisors. In order to see this, return to Figure 8.2 on page 109. For B, the expectations of moving down and across at node B are equal, and if he adheres to the preferences of his ideal advisor, then he will, say, toss a fair coin to decide which move to make. That is, $p = 0.5$. Hence A will move down, because by doing so she expects 1. Once again, B will realize that this way he will never get to make a move, and he is thus eager to give up P and convince A to move across. He tells her that he will not follow the principle, so the expectations of A are now equal. If A listens to the recommendation of her ideal advisor, she will also toss a fair coin to decide whether to move down or across. But that is not good enough, for why jeopardize their cooperation by staying risk-neutral? She could make sure their cooperation gets a better chance to kick off if she also abandons P, and acts as a risk-seeker. She once again has reason to think that principle P does not determine what is better for her, given that there are situations when it is more beneficial to give it up.

An idealization theory of well-being identifies what is good for a person with the satisfaction of the preferences that person would have were she in ideal conditions for preferring. The ideal advisor theory identifies the relevant preferences as the preferences of the ideal counterpart of the person who reasons in privileged epistemic and cognitive conditions. On the best interpretation of these conditions, these preferences are the preferences the person's fully informed and ideally rational advisor forms over the preferences of the person. But what is the use of establishing the preferences formed in these privileged conditions if actual persons may gain by doing something different than what they would advise themselves to do from these conditions? If we find that contrary to our best efforts at specifying these conditions, following the recommendations of ideal advisors may still leave the actual persons worse off, there is reason to suspect there is something amiss with the theory.

The moral: sometimes it is better not to listen to what you would advise yourself to do, were you ideally placed to give yourself advice.

In the next section, I consider the possibility that $\mathcal{R}$ contains a more refined principle of reasonable levels of risk-taking. I argue that no such principle is possible, at least not within the context of rationality.

8.4 Well-Being and Principles of Risk-Taking

Of the three basic possible candidates for P, we found that one, risk-seeking, is implausible on its own right, and it is possible to construct situations in which the two other principles also become implausible—because actual persons may find incentives to abandon them. This means that if $\mathcal{R}$ contains any of these, actual persons may fail to take the preferences of their ideal advisors as reason-giving, since rejecting these principles can be better for them. At the same time, if $\mathcal{R}$ does not contain a principle that prescribes preference formation in risky situations, then ideal advisors cannot form preferences in these situations and the prospects remain incomparable. Either way, IRP fails to determine what is better for actual persons. It fails to tell us, in these cases at least, what promotes our well-being.

On the one hand, perhaps one may be prepared to bite the bullet and concede that IRP is incomplete. One could say: "So what? It is here, in preferences over risky prospects, that we find examples of incommensurability within the context of well-being." My problem with this proposal is that nothing seems to be incommensurable about these prospects. After all, when I am pondering whether I should try to be a reasonably good philosopher or an exceptionally good piano player (with a predisposition to develop rheumatoid arthritis), my complaint is not that I cannot compare these prospects—my complaint is that what attitude of risk-taking I ought to have towards these prospects does not seem to be a matter resolvable by a principle of rationality.

On the other hand, one could propose that principle P, in a fully developed theory of rational choice, will be much more complex. It will specify some particular *level* of risk-taking. So, for instance, it will tell me to try to become a concert pianist if the likelihood of developing rheumatoid arthritis is within tolerable limits, but choose the career in philosophy if its probability is too high. That is, it would prescribe a rational level of risk-taking. What you ought to do, then, would depend on the riskiness of the prospects you face.

But it is doubtful that you ought to have the same level of risk-taking in all situations. So a complex principle must differentiate between different reasonable levels of risk-taking for different *objects* of preferences. For example, the principle could say that you ought to be risk-averse when making a career choice, to ensure that your life does not turn out to be very bad. Hence I ought to choose to become a reasonably good, although not great, philosopher. But this complex principle could also tell me to be more risk-seeking when I am faced with the choice between staying at my current academic post or accepting a job offer from another country, where I may do my best work, but there is a fair chance that I will not be able to integrate into the academic community there, and my work will go poorly.

In giving advice in real life, we do distinguish between reasonable levels of risk-taking. We believe people ought not to jeopardize their health with smoking,

and that they ought to save up money for their old age. But very often, we also advise people to take risks. We say, "In order for life to have quality, one has to take risks sometimes." We think it is a good thing to travel to other countries, to take reasonable financial risks, we admire people choosing risky professions. Hence, the proposal goes, a fully developed theory of rationality will incorporate a principle of reasonable levels of risk-taking for different objects of preferences.

The problem with this proposal is twofold. On the one hand, given that there is no consensus on reasonable levels of risk-taking in everyday situations, and it is controversial what these levels should be, it is hard to see where we could find the resources to formulate this complex principle. Most likely, the "principle" would turn out to be an infinitely long conjunction, associating reasonable risk-taking indices with descriptions of possible objects of preference, without any principled way to calculate the indices! On the other hand, it is hard to see why such a complex principle would be a principle of *rationality*. Whether and how much you ought to save up for your old days, or whether you should be risk-averse in your career choices are substantive questions (in the sense I introduced the term on page 7), which admit of no answer that can be derived from a theory of rationality.

As I said earlier, we do not have a fully developed theory of rational choice. In Section 2.2, I distinguished between those versions of utility theory which accept the expected utility hypothesis, and modern cardinalism which rejects the hypothesis. What do these tell us about risk-taking? On the former, risk-attitudes are *exogenous*: the theory does not specify what risk-attitude (or level of risk-taking) a rational agent ought to have. Risk-attitudes are given and beyond criticism within the theory (see page 21). In contrast, modern cardinalism establishes utility functions without reference to the risks involved in choices, appealing instead directly to the person's valuations of the final outcomes. For this theory, risk-attitudes are irrelevant in determining utility (see page 24). Hence, modern utility theories do not allow for the criticism of preferences based on unreasonable risk-attitudes. Indeed, if IRP interprets the ideal rationality condition based on any of these theories, it is not able to determine what preferences ideal advisors ought to have in risky situations. It needs to appeal to something else.

Where can we go from here? We have two alternatives. One is to reject IRP; another is to revise it to include some condition, besides the epistemic and cognitive constraints, about reasonable levels of risk-taking. Since it seems to me that there is something fundamentally correct about the idea that what promotes your well-being has something to do with the preferences you would have were you in more privileged epistemic and cognitive conditions, my tentative suggestion is to revise the theory. Here's one proposal for doing that.

We can start by borrowing a famous notion from Rawls (1971). According to him, each person has her own "conception of the good." A person's conception of the good involves convictions about what is valuable, what pursuits are worth-

while, what goals and achievements the person ought to strive for in her life. We can extend the list of the convictions that are part of a conception of the good by including convictions about reasonable levels of risk-taking in different situations in that person's life. Thus, people have a "conception of reasonable risks" as part of their conception of the good. It seems to me that people do indeed have such convictions about reasonable risks. For instance, if one of the pursuits you find worthwhile is mountaineering, your convictions about what risks are reasonable to take in the context of your conception of the good are evidently different from somebody else's convictions of reasonable risks—someone who, for instance, might believe that playing chess is a more valuable pursuit. Nevertheless, in the context of a given conception of the good, people will often agree what risks are reasonable to take. They are likely to agree that different risks are reasonable for you if you prefer to spend your free time mountaineering than the risks which are reasonable for a less adventurous person. There is even reason to think that people would converge on their judgments about what risks are reasonable to take for a person—given her otherwise defensible conception of the good.

Thus, on this proposal, the revision of IRP takes the following form. A person's well-being consists in the satisfaction of the preferences that person would have were she fully informed, ideally rational, and formed her preferences in accordance with convictions about reasonable levels of risk-taking which would be agreed to by all fully informed and ideally rational agents, in the context of the person's conception of the good.

This "revised" IRP is similar to the "standard" IRP in the sense that the person's preferences over alternatives and outcomes can be criticized on the basis of the epistemic and cognitive conditions. Since these preferences are influenced by the person's conception of the good, indirectly a person's conception of the good is also assessed on the basis of those conditions. But the revised theory is also different: with respect to preferences over risky prospects, a *convergence requirement* can be employed to assess and criticize convictions about reasonable risks in the context of the person's conception of the good. This is necessary, since the resources for the criticism of convictions about risk-taking cannot be found within rationality as such.

The revised IRP is faithful to both the objectivist and the subjectivist intuitions. On this theory, for some thing x to promote your welfare, it is necessary that you would prefer x were you fully informed and ideally rational—but it is not sufficient. This is compatible with the subjectivist intuition. At the same time, if there are risks involved, in order for the satisfaction of your fully informed and ideally rational preferences to be good for you, it is also necessary that all fully informed and ideally rational agents agree with your preference—which is just to say that x is worthwhile independently of *your* preferences. This is compatible with the objectivist intuition. Thus, whether the life of a reasonably good philosopher or

an exceptionally good concert pianist with a predisposition to develop rheumatoid arthritis in the fingers is better for me depends on which of these fully informed and ideally rational advisors would prefer me to prefer, in the context of my conception of the good.

Note that the revised ideal advisor theory is difficult to place on the conceptual landscape delimited by an exhaustive and mutually exclusive distinction between subjective and objective theories of welfare. This theory is subjective, since it ties welfare to the satisfaction of preferences. But it is also objective, since some of these preferences must be based on substantive judgments about the reasonableness of risks. Perhaps the reason this theory has not been mapped on this landscape is that in order to discover it one must reject the distinction.

The remaining two chapters develop and defend my revision of IRP. Chapter 9 presents a powerful objection to the ideal advisor theory and examines whether my suggested revision can cope with the objection better than other versions of the theory. Chapter 10 discusses the core ideas of the revised theory and gives further reasons why it is preferable to its rivals.

Chapter 9

Well-Being, Autonomy, and Paternalism

9.1 Scanlon's Dilemma

Many philosophers accept a preference satisfaction theory of well-being. Different versions of this type of view can be identified by their answer to the question: "Which preferences matter for well-being?" An actual or unrestricted preference satisfaction theory gives a straightforward answer: all preferences do. But this cannot be true: the objects of some of our preferences have nothing to do with how well our lives go; our preferences can be satisfied without having any causal impact on our lives; and many of our preferences are based on insufficient information, lack of experience, or faulty reasoning.

This last problem prompts many preference satisfaction theorists of well-being to move to some version of the idealization theory—particularly the ideal advisor theory. This theory answers the "Which preferences?" question by imposing epistemic and cognitive constraints on the relevant preferences. These preferences must be based on all relevant information and the person should not make any mistake of reasoning while forming them. On this theory, only the satisfaction of these informed and rational preferences matters for a person's well-being. What is good for the person is determined by what she would prefer were she adequately informed and appropriately rational.

At the same time, one of the reasons an actual preference satisfaction account of well-being seems initially attractive is that it fits well with the idea that individual preferences are normatively salient. This idea is often called *preference autonomy* or preference sovereignty: unless we have sufficiently strong countervailing reasons, people's preferences ought to be respected.

The ideal advisor theory implies that when we are interested in promoting well-being, we take into account only adequately informed and appropriately reasoned preferences. Only these are normatively salient for the theory. But if only the satisfaction of those preferences contributes to your well-being which you would form in an epistemicly and cognitively privileged counterfactual situation, then it is possible that few or even none of your actual preferences would pass the tests of hypothetical preference formation. In short, it is possible that you end up better off if you allow a knowledgeable third party to substitute their judgments of what is good for you for your actual preferences. Whereas an actual preference satisfaction account gives proper weight to the preferences of a person, once we accept an idealized version of the preference satisfaction theory, our account of well-being does

not sit comfortably with preference autonomy. That is, there seems to be a conflict between preference autonomy and the idealization theory—or, in particular, the ideal advisor theory.

One reply an ideal advisor theorist can give is to admit that there is indeed such a conflict, but claim that it is not a problem for *this* theory, but for our more general moral theory. Autonomy and well-being are separate values. Sometimes preferences ought to be respected even if their satisfaction does not promote the well-being of the person, because sometimes we ought *not* to promote well-being, but instead let people satisfy their preferences. That is, sometimes what matters most is not what is good for people, or sometimes what matters most is not the consequences of our choice. Of course, in such cases we have to explain why the respect for autonomy should take precedence over the promotion of well-being.

Another reply an ideal advisor theorist can give is to claim that the conflict is more apparent than real. Preferences ought to be respected *precisely* because they have a close connection to well-being. Since people are best placed to find out what is good for them, their own preferences are most likely to reflect their well-being (as opposed to someone else's judgments). That is, the basis for respecting preferences is that they reliably indicate or "track" welfare. This claim is related to the *best judge principle* (see page 26), and the success of this reply is ultimately dependent on the truth of an empirical hypothesis.

A third reply an ideal advisor theorist can give is to claim that there is no conflict at all. People's preferences ought to be respected, because autonomy is *constitutive* of well-being. A person's preferences determine what is good for the person, and that is the basis for respecting them. By respecting the person's preferences, we respect the person's autonomy, and respecting the person's autonomy promotes the well-being of the person.

Of these three replies, the second is the least satisfactory. Since people are often not reliable judges of their welfare, the conflict may be more real than apparent after all. Thus, I take it that ideal advisor theorists want to choose between the first and the third replies. Moreover, my impression is that they prefer giving the third reply. There are two reasons for this. First, the idea that autonomy is constitutive of well-being is intuitively attractive. Second, it is not seldom pointed out that objective theories of well-being can incorporate this idea: autonomy is one of the "objectively valuable" goods for people. Ideal advisor theorists thus may be attracted to explore the possibility of incorporating it into their theory as another way of showing that whatever an objective theory of well-being can deliver, their theory can deliver too.

One attempt to incorporate the idea of preference autonomy into an ideal advisor theory is to be found in the theory of John Harsanyi (1982), which I presented on page 91. Harsanyi holds that welfare consists in the satisfaction of "true personal preferences"—those self-regarding preferences which the person would form

if she had all the relevant information and reasoned appropriately. Moreover, Harsanyi explicitly appeals to a principle of preference autonomy:

(A) "In deciding what is good and what is bad for a given individual, the ultimate criterion can only be his own wants and his own preferences." (1982:55)

Harsanyi believes that one reason to accept his ideal advisor theory is given by (A). He admits that it is possible that a person prefers what is worse for her. In his terminology, a person's "manifest" (revealed or actual) and "true" (informed and rational) preferences can come apart. Since only the true preferences of the person are relevant to her well-being, it is possible that others can judge better what is good for the person. But, Harsanyi claims, this is not in conflict with (A), since, when determining what is good for the person, we ultimately still appeal to her preferences.

Harsanyi's theory is attacked in this respect by Thomas Scanlon (1991). Although Scanlon targets Harsanyi's theory only, his argument is clearly intended to apply to all versions of the ideal advisor theory. This is how it goes in its generalized form. Take any purported ideal advisor theory. For simplicity, call that theory $\mathcal{T}$. Then:

(1) $\mathcal{T}$ holds either that a person's adequately informed and appropriately reasoned preferences *determine* what is good for the person, or that a person's adequately informed and appropriately reasoned preferences are merely *indicators* of what is good for the person.
(2) If $\mathcal{T}$ holds that a person's adequately informed and appropriately reasoned preferences are merely indicators of what is good for the person, then $\mathcal{T}$ is not a genuine ideal advisor theory.
(3) If $\mathcal{T}$ holds that a person's adequately informed and appropriately reasoned preferences determine what is good for the person, then $\mathcal{T}$ implies that the person's actual preferences are not the ultimate criterion of what is good for this person.
(4) If $\mathcal{T}$ implies that a person's actual preferences are not the ultimate criterion of what is good for this person, then $\mathcal{T}$ violates (A).
(5) If $\mathcal{T}$ violates (A), then $\mathcal{T}$ is paternalistic.
(6) Therefore, either $\mathcal{T}$ is not a genuine ideal advisor theory, or $\mathcal{T}$ is paternalistic.[1]

[1] I must admit I am not entirely sure I interpret Scanlon correctly. This is my best attempt. If my interpretation leaves a bit to be desired, that's not due to lack of trying. Perhaps it is better to consider the argument not as a piece of Scanlon scholarship, but as a representative for an argument that often appears in the contemporary literature on the ideal advisor theory. I chose Scanlon's formulation not because it is the clearest, but because he at least discusses it more than in passing.

Let's look at premise (1) first. According to Scanlon, there are two ways one can construct an ideal advisor theory. The more familiar way is to start from the claim that a person's adequately informed and appropriately reasoned preferences *determine* what is good for the person. But there is also another way to conceive of the role of preference in the theory. On this version, the person's adequately informed and appropriately reasoned preferences are taken as *evidence* for what would promote her welfare: such preferences *indicate* what is good for the person.[2] Thus, Harsanyi and his theoretical allies can see the role of preference in their theory in two different ways. Whichever they choose, however, they face one of the horns of a dilemma.

Consider the second alternative. If your preferences merely indicate what is good for you, then there is something else your preferences are the indicators of, something which determines what is good for you. If ideal advisor theorists were to accept this view, their theory would in effect cease to be a preference satisfaction theory of well-being, as asserted in (2). Elsewhere, Scanlon argues that such theories are in effect objective theories, or "substantive goods theories" (1993). Furthermore, if the connection between adequately informed and appropriately reasoned preferences and well-being is constructed this way, the ultimate criterion of what is good and bad for a person is not her preferences. This implies that preferences are not normatively salient for this view. Of course, this does not mean that ideal preferences are not useful in a theory of well-being, but their role is only derivative. As Scanlon says:

> Someone who accepts a substantive goods theory, according to which certain goods make a life better, will no doubt also believe that these goods are the objects of informed desire—that they would be desired by people who fully appreciated their nature and the nature of life. (1993:190)

An ideal advisor theory, therefore, needs to insist that adequately informed and appropriately reasoned preferences *determine* what is good for the person. But since actual and ideal preferences can be very different, the person's actual preferences are not the ultimate criterion for determining what is good or bad for the person on this view (3). That is, the connection between a person's manifest preferences and her true preferences might be severed. If so, the ideal advisor theory is in conflict with (A). Thus, the dilemma Scanlon presents to the ideal advisor theory is this: it either holds that adequately informed and appropriately reasoned preferences determine what is good for a person, or it takes such preferences to be mere indicators of the person's welfare. In the former case, the theory violates the principle of preference autonomy. In the latter case, the theory ceases to be a preference satisfaction theory. Thus, Scanlon concludes:

[2] One example for such an ideal advisor theory is suggested by Mongin and d'Aspremont (1999) in their discussion of the relation of utility theory and ethics.

> The conflict between the principle of Preference Autonomy and the move to "true" preferences reflects a fundamental moral tension, not just an inconsistency in Harsanyi's theory. (Scanlon, 1991:29)[3]

What shall we make of Scanlon's dilemma? His argument, as I reconstructed it, clearly begs the question against the ideal advisor theory.[4] Premise (4) implicitly presupposes that on the principle of preference autonomy, (A), autonomy must be interpreted on the domain of *actual* preferences. But ideal advisor theorists like Harsanyi can insist that (A) can be interpreted on the domain of *ideal* preferences. That is, they can deny (4) by giving an interpretation of (A) in terms of the person's adequately informed and appropriately reasoned preferences—by showing that only such preferences are properly the "person's own." They can argue that only these preferences are *autonomous*, and the principle of preference autonomy defends only autonomous preferences.

In Section 9.2, I discuss the interpretation of the principle of preference autonomy in terms of ideal preferences. I argue that if ideal advisor theorists give this interpretation, they can successfully undermine premise (4). Nevertheless, the worry behind the objection that the ideal advisor theory is paternalistic might remain: isn't there something paradoxical in the idea that when we promote people's well-being on the ideal advisor view, we can interfere with their actual preferences yet maintain that we respect their autonomy at the same time, since we respect the preferences they would have were they adequately informed and appropriately rational? In order to dispel this worry, Section 9.3 turns to the literature on paternalism. But in this literature, I find a basis for the worry. Thus, Section 9.4 argues that there is a problem for the ideal advisor theory, and checks whether the version of the ideal advisor theory I propose, the revised IRP, can avoid the problem.

9.2 The Problem of Malleability

Scanlon believes that Harsanyi moves to true preferences from manifest preferences and departs from the principle of preference autonomy in response to what

[3]"Moral tension," Scanlon says, because he believes any restriction placed on the preferences which are relevant to well-being is introduced in order to retain the idea that the satisfaction of preferences is morally important (see 1993:187–8). In later work, he argues that preference satisfaction, and well-being in general, are not fundamental values (1998:108–43). That is to say, he thinks we ought to give less moral weight to preference satisfaction and well-being. Harsanyi briefly replies to Scanlon's criticism in his 1997:140–1. Note that Harsanyi excludes from social choice not only the uninformed manifest preferences of the person, but also her "antisocial" true (informed and rational) preferences. (Antisocial preferences are those arising from sadism, envy, resentment, malice, etc.) These, however, are excluded not because they are irrelevant to well-being, but for explicitly moral reasons (1982:56). Scanlon does not object to this move.

[4]And that makes me suspect that I am likely to have misinterpreted it. Nevertheless, we are now in the position to examine the ideal advisor theory's defense.

he calls the "problem of malleability" (1991:28–9). This is the problem that preferences can be "adaptive": reflecting the effects of indoctrination and manipulation, instead of being based on informed and reflected judgment. It is not impossible that manifest preferences are formed under such influences. If they are, they might not adequately be the person's "own" preferences. Scanlon seems to think that the satisfaction of such preferences does not make the person better off, therefore an ideal advisor theory must be able to exclude these preferences.

How can an ideal advisor theory exclude such non-autonomous preferences? It can employ some constraint in order to determine which preferences are adequately the person's "own." There are two familiar ways of excluding non-autonomous preferences. One way is to look at the *causal history* of the preference. If it has an inappropriate causal background, it is not autonomous. This account is sometimes called the *historical* account of preference autonomy. A rival account looks not at the causal history of the preference, but whether it would be affirmed by the person in her higher-order preferences. On this account, preferences are autonomous if and only if the person would prefer to have these preferences. Such accounts are sometimes called *hierarchical* accounts of preference autonomy.

One version of the hierarchical account requires that the higher-order preferences are formed on the basis of all relevant information and without committing any sort of cognitive error. An ideal advisor theorist is most likely to accept this account for excluding non-autonomous preferences. This is because it is a natural extension of her theory. The theory holds that welfare consists in the satisfaction of adequately informed and appropriately reasoned preferences, identified by epistemic and cognitive constraints. Someone who accepts this theory can argue that these constraints not only determine which preferences are relevant to a person's well-being, but they also determine which of the person's preferences are autonomous.

Consequently, even if Scanlon is right in thinking that the satisfaction of non-autonomous preferences cannot make a person better off, he has not provided an argument against the ideal advisor theory. Proponents of the theory can argue that the constraints they impose upon preferences filter out non-autonomous preferences. Furthermore, when they propose to promote well-being, they do not violate the principle of preference autonomy, since they propose to promote only the satisfaction of autonomous preferences.

One may object that the epistemic and cognitive constraints do not ensure that adequately informed and appropriately reasoned preferences have the appropriate causal history, since they do not look at the history behind these preferences. But an ideal advisor theorist can deny, in reply, that the inappropriate causal history of a preference makes that preference non-autonomous. She can point out that insofar as an actual preference is endorsed by the higher-order, adequately informed and appropriately reasoned preferences of the person, it is *very likely* that this prefer-

ence does indeed have the right sort of causal history. After all, it is hard to imagine forms of manipulation which result in preferences that you would have formed if you had all the relevant information at your disposal and you reasoned correctly.

For illustration, consider the stock example of the problem of malleability: the battered housewife. This woman's actual preferences are very modest due to manipulation by her husband: her major aim in life is to subserviently make her husband's life comfortable, she has low self-esteem, and she sincerely denies that she could have any other accomplishments than the modest ones she is indoctrinated to believe she can strive for. She has adopted her preferences to her situation. Would this housewife continue to have these preferences if she knew all the relevant information and reasoned correctly? It is very doubtful. Her husband's manipulating her must have been, partly, a matter of giving her false beliefs about herself, her abilities, and her circumstances; it must have been, partly, an attempt to suppress her reasoning, valuation, or deliberation.

Scanlon's implicit assumption that (A) must be interpreted on the domain of actual preferences is all the more surprising given that most philosophers agree that in order for a preference to be autonomous it must pass some requirement: typically, either an historical or an hierarchical one. Of these two, the latter is much more plausible, and it fits well with the version of the ideal advisor theory which maintains that autonomy is constitutive of well-being. But even if Scanlon's premise (4) was unobjectionable, premise (5) would still be problematic. Apparently, Scanlon believes that paternalism is a threat to the ideal advisor theory because of the problem of malleability. This problem prompts the move to ideal preferences; the promotion of ideal preferences violates preference autonomy; and the violation of preference autonomy is paternalism.

But consider the following example. In country A, the government is headed by a manipulative populist. People's preferences are cunningly manipulated to be in accord with the aims of the government. Consequently, the government never overrides these preferences—and it boasts in its propaganda that it always respects the principle of preference autonomy. Meanwhile, in country B, the government is headed by a paternalistic dictator. He always does what people would prefer the government to do if they were adequately informed and appropriately rational, even when—as it may often happen—people's actual preferences are different. Moreover, the government in country B never manipulates its people. We would object to the governments of both of these countries—however, the reasons for our objections are different in the two cases. Country A's government is objectionable because it manipulates its people. Country B's government is objectionable because it is completely unresponsive to people's preferences. The government in country A exploits the malleability of preferences—the government in country B does no such thing. Consequently, manipulation does not imply paternalism, or *vice versa*.

The objection that the ideal advisor theory is "paternalistic" in some sense is typically mentioned merely in passing in the critical literature on the ideal advisor theory. Scanlon is one author who actually tries to make a case for this claim. He does not succeed. But I think what fuels his attempt to make the case is the worry that when we promote well-being on the ideal advisor theory, we promote the satisfaction of adequately informed and appropriately reasoned preferences, which may be different from actual preferences. On this theory, it might be possible to interfere with people's lives with the objective of promoting their well-being, *and*, at the same time, argue that the interference also enhances their autonomy. In order to examine this worry, I briefly review the literature of paternalism.

9.3 Justifications for Paternalism

The most influential account of paternalism is from Gerald Dworkin (1972, 1983). Dworkin once defined paternalism as

> interference with a person's liberty of action justified by reasons referring exclusively to the welfare, good, happiness, needs, interests, or values of the person being coerced. (1972:20)[5]

As critics have pointed out, and as Dworkin himself admits now, this definition is too restrictive.[6] A paternalistic action cannot always be understood as an infringement on liberty, and it is not necessarily coercive. For instance, if my child very much wants to be a concert pianist, but I realize that she does not have an ear for music and would be unsuccessful in pursuing a musical career, I may decide not to pay for her music lessons. My decision is paternalistic, but does not involve violating her liberty. Moreover, I may realize that she is very good in abstract thinking, and I may offer to pay for extra math lessons instead. My decision is paternalistic, but does not involve coercion.

The definition of paternalism, therefore, needs to be refined. Dworkin suggests that we should concentrate on judgment, instead of liberty of action or coercion. A paternalistic act, then, is defined as "an attempt to substitute one person's judgment for another's, to promote the latter's benefit" (1983:107). Substitution of judgment

[5]Other definitions are more narrow, focusing on well-being; see, for example, Bok (1980:204): "To act paternalistically is to guide and even coerce people in order to protect them and serve their best interests." VanDeVeer (1986:12) puts the definition in terms of interference: "A paternalistic act is one in which one person, A, interferes with another person, S, in order to promote S's own good." Buchanan (1978:372) includes the dissemination of information in the definition: "Paternalism is interference with a person's freedom of action or freedom of information, or the deliberate dissemination of misinformation, where the alleged justification of interfering or misinforming is that it is for the good of the person who is interfered with or misinformed."

[6]See Dworkin (1983:105), Gert and Culver (1976:45–7), Fotion (1979), and VanDeVeer (1986:25, 1980:188–9).

can be carried out both by disregarding the decision another person has made, and by influencing the process of deliberation by which the person arrives at a decision. Thus Dworkin now holds that "there must be a violation of the person's autonomy (which I conceive as a distinct notion from that of liberty) for one to treat another paternalistically" (1983:107). On this view, paternalism is defined in terms of autonomy, and this has become the standard way in the literature to look at paternalism.

When is a paternalistic act permissible? Before discussing this question, it is worth pointing out that many paternalistic practices may be permissible but not on paternalistic grounds. That is, justifications for paternalistic interference may be non-paternalistic, paternalistic, or a mixture of the two. For instance, interference may be justified by an appeal to the harm an action would cause to third parties.

For permissible paternalism, Donald VanDeVeer (1986) distinguishes between two types of justification:

> The proposed candidates (for a *paternalistic* justification) seem to fall into two broad categories:
>
> (1) those which, in some fashion, appeal to actual, predicted, or hypothetical *consent on the part of the subject* of the paternalistic act, or hypothetical consent of a "fully rational person"; and
> (2) those which do not.
>
> The latter typically suppose that the *consequences* of the paternalistic interference ... are so important as to override any presumption against the act in question. (1986:41, his emphases)

Consider first justifications for paternalism based on (2). One type of such justifications is consequentialist: it holds that paternalistic interference is justified if and only if the consequences of interference will result in more well-being for the person than her own action would result in. This view is unattractive. Sometimes we ought to let people choose what they prefer, even if what they choose is worse for them. Furthermore, on this view, it is possible to justify widespread interference with people's lives, even when people are not mistaken about what would promote their welfare. People often rationally choose what would not promote their welfare, because their own well-being is not the only value they pursue in their lives. Consequentialist justifications of paternalism seem, for these reasons, unacceptable.[7]

[7]For an example of a consequentialist justification of paternalism, see Brock (1983). Another type of justification in this category is based on the notion of personal identity. Its discussion, however, is beyond the scope of this work. See Regan (1974) and Kleinig (1983:67–73). A completely different approach to paternalism is to be found in New (1999); see also Calcott (2000) and Leonard *et al.* (2000). For general discussions of paternalism in the context of politics, see Weale (1978) and Goodin (1991, 1993, 2002:48–72).

Consider now purported justifications of paternalistic intervention falling into (1). A view on the justification of paternalism in this category might tie permissible paternalistic interventions to the *ex ante* or prior consent of the actual person. Such a view, however, is too restrictive—it excludes all interventions to which the person has not explicitly consented. A different view may appeal to the (predicted) *ex post* or subsequent consent of the person. But it is hard to ascertain what a person would consent to after the paternalistic act has been done. We may also be unable to get her subsequent consent for some contingent reason—even though the interference continues to seem permissible. Finally, appeals to hypothetical consent, according to VanDeVeer, can take the form of appeals either to the hypothetical consent of the *actual* person, or to the hypothetical consent of the person were she adequately informed and appropriately rational. The former appeals to what the person would consent to as she is, while the latter appeals to what the person would consent to, not as she is, but as she would be if she was in ideal conditions to give her consent.

Although the appeal to the hypothetical consent of the actual person may seem a more attractive way to justify paternalism, there are cases when the appeal to what the actual person would consent to does not justify interference that intuitively seems permissible. After all, the actual person may be incompetent in various ways, and in such cases the interference to promote her welfare must be justified by an appeal to what she would prefer, not as she is, but as she would be if she was adequately informed and appropriately rational. For this reason, a consent theory of permissible paternalism is more plausible if it is based on the *hypothetical consent* of the adequately informed and appropriately rational "counterpart" of the person.

Dworkin's account of permissible paternalism falls into this version of (1): it appeals to the hypothetical consent of adequately informed and appropriately rational persons. Thus, his justification of paternalistic interference appeals to some sort of *incompetence* on the part of the person. One kind of incompetence is *epistemic*: the person is not aware of some relevant and important fact. Another kind of incompetence is *cognitive*: the person is not in the position to make a calm, considered, reflected choice, or she makes some mistake of practical reasoning. On this view, permissible paternalistic interference is justified on the basis of epistemic or cognitive incompetence.[8]

[8]Dworkin discusses a further sort of incompetence. Consider cases of weakness of will. Odysseus commands his men to tie him to the mast when they approach the island of the sirens: he knows that he would not be able to withstand their song, thus he makes it impossible for himself to act on subsequent desires he does not endorse. His men are justified to ignore his later pleas to be released. Such cases involve "motivational incompetence" (the term is mine, not Dworkin's). Here the appeal is made to the consent of the actual person, and not to the consent she would give if she was in ideal conditions to give her consent. Thus, Dworkin's hypothetical consent view justifies paternalism by reference to cognitive, epistemic, or motivational incompetence (1972:28–33). In what follows, I bracket motivational incompetence for two reasons: first, it raises issues beyond the scope of this work; second, I am not sure there is genuine substitution of judgment in cases of motivational

What sort of paternalistic interferences are permissible on an hypothetical consent account? In order to see this, consider the distinction introduced by Joel Feinberg (1971) between *soft* and *hard* (or *weak* and *strong*) paternalism. Feinberg argues that paternalistic intervention is permissible if and only if it takes the soft or weak form. By soft paternalism, Feinberg means that interference is justified because the choice the person would make is not fully *voluntary*. Hard paternalism, on the other hand, allows interfering with the fully voluntary choices of the person. A fully voluntary choice is one that satisfies certain epistemic and cognitive conditions (among others, which need not concern us here): one is "fully informed of all relevant facts and contingencies, with one's eyes wide open, so to speak, and in the absence of coercive pressure"; and "to whatever extent there is compulsion, misinformation, excitement or impetuousness, clouded judgment (as from alcohol), or immature or defective faculties of reasoning, to that extent the choice falls short of perfect voluntariness" (1971:7).

What Feinberg calls voluntary choice closely corresponds to a choice made in accordance with an adequately informed and appropriately reasoned preference. The conditions he gives for voluntariness describe the epistemic and cognitive constraints of an hypothetical consent theory of permissible paternalism. In this respect, it closely resembles Dworkin's theory.[9]

Fully voluntary choices, based on knowledge of all the relevant information and the correct exercise of one's cognitive capacities, "not only have their origin 'in the agent,' they also represent the agent faithfully in some important way: they express his or her settled values and preferences" (Feinberg, 1971:7). Such choices can be considered the autonomous choices of the person. Compare the way Dworkin defines soft paternalism:

incompetence. The weak-willed subject agrees that it would be better for her to do what she cannot bring herself to do, and this is the reason others might be permitted to interfere. For different hypothetical consent views, see VanDeVeer (1986:45–94) and Gert and Culver (1979). Compare Husak (1981:30–5).

[9]Feinberg and Dworkin are interested, however, in different kinds of paternalistic interferences. While Feinberg discusses legal paternalism, Dworkin is concerned, more generally, with paternalistic policies and paternalistic actions. The similarity between Feinberg's and Dworkin's accounts has also been pointed out by Donald VanDeVeer: "The principle that Feinberg finds acceptable, weak paternalism, is not explicitly couched in terms of hypothetical consent. However, it is possible to do so. For example, weak paternalism seems equivalent to the view that paternalistic interference is permissible, and only permissible, if the subject would consent to the interference if *he* were making a fully voluntary choice" (1980:200, his emphasis). Note, however, that Feinberg changed his view in his 1986 book. In his later view, for some interference to be permissible it is sufficient that the choice is not "voluntary enough." Voluntariness, in turn, depends on the context of the choice. Thus, his later view is much more restrictive in sanctioning interference, since interfering with less than fully voluntary choices of the person may also be instances of hard paternalism. See Feinberg (1986:113–8); see also Scoccia (1990).

> By soft paternalism, I mean the view that (1) paternalism is sometimes justified, and (2) it is a necessary condition for such justification that the person for whom we are acting paternalistically is in some way not competent. This is the view defended by Feinberg … By hard paternalism, I mean the view that paternalism is sometimes justified even if the action is fully voluntary. (1983:107)

Thus, Dworkin and Feinberg agree that soft paternalism is justified in terms of some form of incompetence, and it is only soft paternalism that can be justified. Hard paternalism would warrant interference with persons who are competent in forming their preferences. But can the hypothetical consent account of permissible paternalism exclude instances of hard paternalism?

9.4 Paternalism and Risk

Recall that part of Scanlon's criticism of the ideal advisor theory was that if it holds that the person's adequately informed and appropriately reasoned preferences determine what is good for the person, then it is paternalistic. One way to understand this objection is this: if our aim is to promote well-being and we accept the ideal advisor theory, then we are permitted to interfere with people's preferences on the ground that they would have different preferences if they were adequately informed and appropriately rational. In some cases—when the interference would be an instance of soft paternalism—this is not a problem; but the objection may be that instances of hard paternalism are also warranted.

This would be a problem, because nowadays there is widespread agreement among philosophers that the only instances of permissible paternalistic interventions are those which are sanctioned by the principle of soft paternalism. According to this principle, interference with a person's life in order to promote the person's well-being—on the basis that her judgment about what is good for her is mistaken—can be justified only if the person is incompetent in certain ways. This was argued by Feinberg, and Dworkin is at pains to show that his hypothetical consent account of permissible paternalism yields this principle.

In reply to this interpretation of the objection, ideal advisor theorists may choose the strategy of combining their theory of well-being with the hypothetical consent account of permissible paternalism. These two components fit together in the sense that they both employ the same sort of epistemic and cognitive constraints for determining the preferences which are relevant to well-being on the one hand, and for specifying the conditions in which the consent of the person may justify interfering with her pursuit of satisfying her preferences, on the other. If the hypothetical consent account of permissible paternalism sanctions only instances of soft paternalism, the ideal advisor theory does not make unjustified paternalistic interferences with people's lives possible.

But consider another aspect of Feinberg's discussion of voluntary and non-voluntary choice. In many cases, the harm a person's action would inflict on herself is not unavoidable, but only likely. That is, the person creates some *risk* for herself by her choice. For example, not wearing a seatbelt in a car does not directly cause harm to you, but raises the probability that you will be harmed, should an accident happen.

Many controversial cases of paternalistic interference concern risks. The controversies are about *reasonable* and *unreasonable* risks. Even though people often agree on which risks are reasonable to assume, in other cases it is unclear what risks are reasonable to take. As Feinberg says:

> Certain judgments about the reasonableness of risk-assumptions are quite uncontroversial. We can say, for example, that the *greater* the probability of harm to self (1), and the magnitude of the harm risked (2), the *less* reasonable the risk; and the *greater* the probability the desired goal will result (3), the importance of the goal to the doer (4), and the necessity of the means (5), the *more* reasonable the risk. But in a given difficult case, even where questions of probability are meaningful and beyond dispute, and where all the relevant facts are known, the risk-decision may defy objective assessment because of its component personal value judgments. (1971:6–7, his emphases; see also 1986:103)

Feinberg actually sets up the distinction between fully voluntary and not fully voluntary choices in the context of assumptions of risk. A fully voluntary choice is one which the person makes aware of all relevant information and avoiding all sorts of cognitive error.[10] Paternalistic interference can only be justified if the person's choice is not fully voluntary. It is not clear, however, what to say in cases when otherwise well-informed, competent and rational people choose to take unreasonable risks. Feinberg thinks that in such cases governments cannot override people's risk assumptions on the ground that they are unreasonable. What governments can do is to make sure that people are aware of all relevant information and do not exhibit any sort of incompetence in deciding what to do. If interference is justified, it is justified on non-paternalistic grounds, since overriding people's risk assumptions would be instances of hard paternalism.

Dworkin also discusses unreasonable risks. He believes that the principle of soft paternalism is able to render problematic cases of risk-taking tractable. For instance, if someone does not fasten her seatbelt on the ground that she is a risk-seeker, the full appreciation of the possible consequences of an accident would convince the person otherwise, and she would consent to the limitations imposed by traffic regulations.

[10] And in the absence of coercion or various other forms of compulsion. But this constraint is not relevant for my purposes here, so it is understood to be satisfied.

Nevertheless, neither of these strategies are successful in handling "hard cases" arising from unreasonable risk assumptions, like the prohibitions on using certain drugs, protective helmet and safety belt laws, limitations on dangerous activities. In Feinberg's case, one may wonder whether all of the possible hard cases can be given a justification on non-paternalistic grounds (but see his 1986:98–142). In Dworkin's case, it is doubtful that people who fully appreciate the possible consequences of some activity would always give their consent to limitations on that activity. Consider highly hazardous activities, like mountain climbing. It is doubtful that the full appreciation of the possible consequences of an accident by adequately informed and appropriately rational persons implies that they would consent to limitations on mountain climbing. Such limitations would indeed seem paternalistic. Dworkin agrees: "there are risks—even very great ones—that people are entitled to take with their lives" (1972:33). So here he changes strategy: he argues that what limitations on risk-taking would be hypothetically consented to also depends on the role the activity plays in people's lives. A requirement to fasten your seatbelt is trivial, but a ban on mountain climbing would not respect the autonomy of persons.

But this is insufficient to establish the distinction between reasonable and unreasonable risks. On the one hand, what counts as an important activity for people also depends on the limitations imposed on engaging in that activity. On the other hand, it is unclear whose judgment of importance is relevant here. Professional drivers are not exempt from safety belts laws, even though driving plays an important role in their lives.

It seems that an hypothetical consent account of permissible paternalism has difficulties with making the distinction between reasonable and unreasonable risks. Such an account ties permissible interference with people's pursuit of satisfying their preferences to some sort of incompetence. But what sort of incompetence is the taking of unreasonable risks? Or, equivalently, what sort of failure is the having of an unreasonable risk-attitude when forming preferences?

It is certainly not an epistemic shortcoming: if you have an unreasonable risk-attitude (but you are otherwise adequately informed), your mistake is not lacking some relevant information. If there is a reasonable level of risk-taking for your situation, then it is determined by a normative principle telling you how you ought to form your preference in that situation, and not a piece of information that describes your situation. Thus, the only possibility that remains is that when you have an unreasonable risk-attitude, your failing is not an epistemic, but a cognitive one. What risks you ought to assume, then, must be a requirement of rationality.

Predictably, the argument I am about to make parallels the one I made against the ideal advisor theory in Section 8.4. There I argued that preferences over risky prospects cause a difficulty for the ideal advisor theory. On the one hand, if the theory does not incorporate some principle of reasonable levels of risk-taking, it is

often impossible to determine what the person, were she fully informed and ideally rational, would prefer herself to prefer. On the other hand, if the theory does incorporate such a principle, then that principle will either be formal or substantive. If the principle is formal—prescribing levels of risk-taking without reference to the content of preferences—then an actual person, with her own conception of reasonable risks, may reject the recommendations of her ideal counterpart as reason-giving for her. In contrast, if the principle is substantive, making reference to the objects of preferences, then it is not a principle of rationality.

Now, if the cognitive constraint of the hypothetical consent account of permissible paternalism does not incorporate some principle of reasonable levels of risk-taking, then it is unable to make the distinction between reasonable and unreasonable risks. This is because if nothing tells fully informed and ideally rational "consenters"—to use the most plausible interpretation of the epistemic and cognitive constraints, and to introduce another metaphor—how to form their preferences when their actual counterparts are faced with risky prospects, then the theory cannot determine what interferences with the pursuit of the satisfaction of their preferences people would hypothetically consent to.

Thus, the cognitive constraint for hypothetical consent must include some principle of reasonable levels of risk-taking. Suppose first that the principle is formal: it assigns a given risk-taking level, or levels, to formally equivalent choice situations, independently of what is at stake—what the objects of the choice are. The problem with such a principle is that otherwise informed and rational people will often not find it acceptable. This is because in formally equivalent choice situations with different objects of preference, competent people find different risk-attitudes reasonable. They may, for example, make judgments about risks based upon the importance of the objects of their preferences for them. Suppose a road-trip and a mountain climbing expedition are equally risky. While people may think that it is unreasonable not to use a seat-belt on the road-trip, they may reject not going on a mountaineering expedition just because similar risks are involved. A formal principle is insensitive to the different roles buckling up and mountaineering play in people's lives. A hypothetical consent account with a formal principle for determining what risks are reasonable therefore sanctions hard paternalism—it permits the overriding of the risk-assessments of competent persons in different situations.

Consequentially, the principle of reasonable levels of risk-taking must be substantive: it must determine what risks or levels of risk-taking are reasonable based on the *objects* of preferences. It should prescribe different levels of risk-taking when driving a car and when going mountain climbing—an activity which is attractive to people partly because of the very risks involved. Such a principle would be a long, open-ended conjunction mapping objects of preferences to levels of risk-taking. However, if there is such a principle for determining what risks or levels of risk-taking are reasonable, then rationality alone does not have the resources to

uncover it. That is, a substantive principle would not be a principle of rationality.

Where have we got? The objection to the ideal advisor theory I have been exploring is that this theory violates preference autonomy, therefore it is paternalistic in some sense. The best reading of the "some sense" clause is that on this theory, the promotion of well-being sanctions hard paternalism. One possible reply to this objection is to argue that the fully informed and ideally rational preferences of a person are the person's autonomous preferences, therefore the principle of preference autonomy defends only these preferences, rather than the actual or manifest preferences of a person.

A promising way of showing this is to supplement the ideal advisor theory with an hypothetical consent account of permissible paternalism. I have argued, however, that in order to be able to make the distinction between reasonable and unreasonable risks, that account must incorporate some principle of reasonable levels of risk-taking. But that principle must be a substantive principle—that is, it cannot be part of the cognitive constraint on hypothetical consent. Because both the ideal advisor theory of well-being and the hypothetical consent account of permissible paternalism must appeal to something else besides rationality and information, the defense against the objection which combines the two is not satisfactory.

Nevertheless, perhaps fully informed and ideally rational persons would agree on what risks are reasonable to take. In a given person's situation, taking into account that person's characteristics, goals and plans, the probabilities involved in the choice, and the importance of the outcomes for the person, they may be able to come to an agreement about the risk-attitude the person ought to choose in forming her preference. Furthermore, fully informed and ideally rational persons, taking into account the role and importance of certain activities in people's lives, the probabilities of harm from engaging in those activities, and the social consequences of allowing or restricting those activities, may be able to agree on general limitations and regulations for potentially harmful self-regarding activities. That is, they may converge on their judgments about risks.

In Section 8.4, I proposed to revise the ideal advisor theory by incorporating a *convergence requirement* into it in order to be able to distinguish between reasonable and unreasonable risks. I called the revision the "revised IRP" theory. I suggest a hypothetical consent account of permissible paternalism can use the same convergence requirement to be able to distinguish between reasonable and unreasonable risks. The revised IRP, supplemented with that hypothetical consent account, can meet the objection from hard paternalism. On this theory, when we promote the well-being of a person, we promote the satisfaction of the preferences she would have were she fully informed, ideally rational, and all fully informed and ideally rational "advisors" would agree on her preferences. These ideal preferences are her autonomous preferences—thus, the promotion of her welfare does not violate preference autonomy. Neither does it allow hard paternalism, since only those in-

terferences with her pursuit of satisfying her preferences are permissible which she would consent to were she in ideal conditions to give her consent: were she fully informed, ideally rational, and all fully informed and ideally rational "consenters" would agree on her preferences.

Of course, the question remains open whether fully informed and ideally rational advisors and consenters would agree on reasonable risks. Thus, it is time to discuss in more detail the revised IRP in general, and this question in particular.

Chapter 10

The Revised IRP

10.1 Contours of a Theory

Each person has a conception of the good. A conception of the good is the collection of the person's deeply held convictions and beliefs about what goods are valuable and what goals are worth their while to pursue in her life. It provides the framework in which a person makes her choices and it gives structure and coherence to those choices. It contains the beliefs which are relevant to the person for forming her preferences, and it determines the hierarchy of the person's preferences. Even though, as I said on page 76, I doubt that most people have a life plan in the sense philosophers use the term, they undoubtedly have some more or less complex hierarchy of values and goals, determined by their conception of the good.

A conception of the good is not the same thing as a theory or conception of well-being, although philosophers sometimes seem to use the two concepts interchangeably.[1] While the latter is an account of that in virtue of which something contributes to welfare, the former is a view about what particular goods contribute to welfare, among other beliefs. A person may have a conception of the good without having a theory of well-being—indeed, I suspect most people do have a conception of the good without having a theory of welfare. That is, they have beliefs about what goods contribute to their welfare without having an account of in virtue of what those are good for them.

As I suggested in Section 8.4, a person also has a "conception of reasonable risks" as part of her conception of the good. A conception of reasonable risks is the collection of the person's convictions and beliefs about what risks are reasonable to take in the context of various goals, pursuits and goods. Such beliefs are necessary to weigh and balance between the person's options and to give coherence to her choices. They are relevant to many of our most important choices, and they are related to the hierarchy of our preferences: a person has different views on what risks are reasonable to assume for choices of varying importance to her.

When we make a judgment about what would promote the welfare of a particular person, we normally have to take into account her conception of the good—not because it is beyond criticism, but because it provides the context in which the person's options can be evaluated. In order to determine what a person ought to choose

[1] Kymlicka (1990) does this, at least at certain places.

to promote her well-being, you must know something about the goals, values, and overarching preferences of that person. On the other hand, a conception of the good must be defensible. People can be mistaken in their convictions about what is valuable, what pursuits are worthwhile, and what goals and achievements they ought to strive for in their lives. Different theories of well-being provide *one* sort of basis for the criticism of conceptions of the good. They provide one sort of basis only, since conceptions of the good may also be criticized from other perspectives—for instance, from a moral perspective.

On the ideal advisor theory, conceptions of the good can be assessed in terms of the epistemic and cognitive constraints. Their assessment is indirect: the theory evaluates the preferences of the person, some of which, in turn, are at least partly determined by her conception of the good. Thus, on this theory, a person's conception of the good is defensible if the person would have this conception, along with the preferences which correspond to this conception, were she adequately informed and appropriately rational. In contrast, on an hedonist theory of well-being, for example, a person's conception of the good is defensible if the values and goals it includes are valuable from an hedonist perspective. Arguably, the ideal advisor theory is more attractive in this respect, because the way it evaluates conceptions of the good is very similar to what we do in real life when we deliberate about what goods or pursuits would be good for us. On such occasions, we try to form an informed and rational preference: we attempt to gather as much relevant information as possible, and we attempt to reflect on that information in circumstances in which we are less likely to make cognitive and other sorts of error, including those which are due to emotional disturbances, psychological compulsions, and the like. Moreover, when we ask other people for advice, we tend to give more weight to the advice of those who are more informed, experienced, and rational. The ideal advisor theory proposes constraints on the preferences that are relevant to well-being which in many ways correspond to our everyday deliberation about well-being.

Nevertheless, even the version of the ideal advisor theory which I believe gives the most plausible interpretation of the epistemic and cognitive constraints, IRP, falls short of being able to determine what is good for a person when the person faces choices which involve risks. So I propose to augment the epistemic and cognitive constraints with a *convergence requirement*. The role of the requirement is to ensure that the fully informed and ideally rational preferences of the person incorporate reasonable levels of risk-taking—that is, that the preferences of the "ideal advisor" of the person which determine what the person ought to do to promote her well-being involve only reasonable risks. The idea is that the agreement of fully informed and ideally rational ideal advisors guarantees the authority of the preference of the person's ideal advisor for the person in choices involving risk. The requirement comes into play only when the person has to form preferences for such choices. Thus, the theory I propose, the "revised IRP," is a "two-level"

theory: when the person faces alternatives which lead directly, or with certainty, to an outcome (also called "pure" alternatives), only the epistemic and the cognitive constraints come into play. On the other hand, when the person faces alternatives which lead to each one of a number of outcomes with a given probability (also called "mixed" alternatives or prospects), the convergence requirement comes into play in addition to the epistemic and the cognitive constraints. Since virtually all of our choices involve risk, practically almost always only the latter "level" is relevant.[2] That is to say, our choices are almost always between *prospects*, which are complex objects including a set of outcomes and the probabilities with which those outcomes might obtain. Therefore, putting the rather rare cases of choice under certainty aside, the theory I am proposing can be defined the following way:

(Revised IRP) Some prospect promotes a person's welfare if and only if that person would self-regardingly prefer that prospect to other available prospects were she fully informed and ideally rational, and other fully informed and ideally rational persons would strictly converge on her preference.[3]

One could object that since choice under certainty is a special case of choice under risk (that is, the case in which the probability of some outcome's obtaining is 1), the convergence requirement should apply to such cases as well. But recall that the convergence requirement was introduced specifically for risky choices, since it is in the case of such choices that rationality and information are insufficient to determine what a person would prefer were she in ideal conditions for forming her preferences. The ideal conditions need to be supplemented by some requirement that can distinguish between reasonable and unreasonable risks. The point is that in the case of choice under certainty, there is no scope for fully informed and ideally rational ideal advisors to disagree about what a person ought to prefer, whereas in the case of choice under risk there is scope for disagreement. Hence, in the former case, a convergence requirement is superfluous.

Note also that whether some risk is reasonable or unreasonable is context-dependent. For example, suppose you have to weigh prospects which include very good outcomes with very low probabilities and very bad outcomes with very high probabilities. Perhaps one would say that these prospects contain only unreasonable risks. But given that they might be the only available options, the relevant question is which one of risks is *more* reasonable (or less unreasonable). Thus, we can think about reasonableness as a kind of ordering of risky prospects.

[2] A special case of choice under risk is choice under uncertainty. In such choices, the probability with which some of the outcomes might obtain is not known to the decision maker. But since ideal advisors are fully informed, they know the relevant probabilities. Of course, perhaps there are cases of genuine uncertainty—when the relevant probabilities *in principle* cannot be known. If there are such cases, ideal advisors form the best possible subjective probabilities.

[3] What strict convergence is will be explained below.

The convergence requirement also seems to mirror our everyday deliberation about what would be good for us as far as that deliberation involves asking for the *advice* of other people.[4] We tend to think that a piece of advice is better advice if it is backed up by the opinion of more people than that which is not. That is, if the people whose advice we seek in order to settle our practical problem agree on what we ought to do, we naturally give more weight to that advice. Convergence is a good *prima facie* consideration for a recommendation to be correct. Therefore, on the revised IRP, conceptions of the good can be assessed indirectly, by assessing the person's preferences, which are at least indirectly determined by the person's conception of the good, including her conception of reasonable risks. Thus, the theory evaluates a person's conception of the good in terms of the epistemic constraint, the cognitive constraint, and the convergence requirement.

The most familiar objection to the idea of convergence is a sort of *open question argument*: according to the objection, we have no reason to suppose that ideal advisors would converge on their preferences.[5] Claims of convergence always have an "open feel." But the objection usually stops at this point. We are given no reason why ideal advisors would *not* converge on their preferences.

Perhaps the objection is that we have no reason to expect that *all* fully informed and ideally rational persons would converge on their preferences. Perhaps some ideal advisors would "dissent" from the rest. But why would they do so? The convergence requirement tries to capture the idea—once again, prevalent in everyday reasoning—that if it is a fact that some particular thing is better for a person than some other thing, then all would agree to this fact unless there was a special reason for their disagreement—for instance, if some lacked relevant information or suffered from some form of cognitive error. Furthermore, it is important to note that a convergence requirement does not impose the very same preferences on ideal advisors: it imposes a particular structure on their preferences when they form preferences over what would be good for a particular person. That is, the claim is not that all people, were they fully informed and ideally rational, would have identical preference orderings, but it is rather that their preferences would take a particular structure *vis-à-vis* one another.

This is because "converging on preferences" can have different meanings depending on the interpretation of the convergence requirement. Unfortunately, both the proponents and the opponents of the requirement tend to leave it uninterpreted. The different interpretations give versions of the convergence requirement with varying strengths. In order to see this, suppose that ideal advisor j has to form a preference over prospects x and y. The preference she forms determines which prospect is better for her actual counterpart—but only if other ideal advisors con-

[4] On this point, compare Smith (1994:151–2).

[5] See Hubin (1999) and Sobel (1999).

verge on that preference. That is, the preferences of all the other ideal advisors $N = 1, \ldots, i, \ldots, n$ ($j \notin N$) *collectively* determine what ideal advisor j should prefer—and that is just to say, they determine which prospect promotes the welfare of j's actual counterpart.

Perhaps the most obvious interpretation of the convergence requirement is what could be called the *strong convergence requirement*:

$$\forall i \in N,\ (u_i(x) > u_i(y) \Rightarrow x \succ_j y) \vee (u_i(x) = u_i(y) \Rightarrow x \sim_j y).$$

This says that if all the other ideal advisors strictly prefer x to y, then ideal advisor j must also strictly prefer x to y; or if all other ideal advisors are indifferent between x and y, then j must also be indifferent between them. This is a strong requirement in the sense that j's preference must exactly correspond to the collective preference of the other ideal advisors. I believe that the appeal of the objection that we have no reason to expect ideal advisors to converge on their preferences derives, to a large extent, from the strong interpretation of the convergence requirement. This interpretation leaves no room at all for even mild disagreement between ideal advisors. Hence, this version of the convergence requirement is likely to be too strong.

A related problem is that this interpretation implies that if some of the other ideal advisors strictly prefer x to y, while some others are indifferent between them, then j's preference cannot be determined. And a case can be made that such a scenario is possible. In Section 8.4, I argued that judgments of the reasonableness of risk-taking in any situation are substantive judgments—that is, they appeal to the objects of the preferences in the situation. Moreover, I quoted Feinberg's claim on page 129 that some judgments about risk-taking are relatively uncontroversial. For instance, if the probability of the worse outcome is sufficiently high, or the worse outcome is sufficiently bad, the risk is more likely to be unreasonable; and if the probability of the best outcome is sufficiently high, or the best outcome is very good, the risk is more likely to be reasonable. Furthermore, the context of the choice influences the reasonableness of the risk—for instance, whether there are other alternative course of actions to secure the desired outcome may make a difference. Thus, judgments about the reasonableness of risks are quite complex: they might involve considerations about the probabilities with which the different outcomes may obtain, about the value of the possible outcomes, and about the context of the choice. This suggests that it is unrealistic to expect that in any given situation there will be one particular level of risk-taking that is reasonable: it is more likely that there will be a *range* of risk propensity within which risk-taking is reasonable. This dovetails with our intuition from everyday reasoning about risks: we think that different levels of risk-taking, given that they are within appropriate limits, are acceptable in a given situation. If you go mountaineering, there is a

range of risk-attitudes with which you might pursue that activity within the limits of being reckless or being overcautious to the extent that you become a spoilsport.

Consequentially, the revised IRP may incorporate the *weak convergence requirement*:

$$\forall i \in N,\ u_i(x) \geq u_i(y) \Rightarrow x \succcurlyeq_j y.$$

This says that if none of the other ideal advisors (strictly) prefer prospect y to prospect x, then neither does j (strictly) prefer y to x. In effect, all the requirement says is that if everyone else rejects y, then j also rejects y. In particular, independently of whether the other ideal advisors strictly prefer x or they are indifferent, j can strictly prefer x or be indifferent. The weak convergence requirement is compatible with some degree of disagreement between ideal advisors, although they must still be in broad agreement.

Several cases are possible. First, if some ideal advisors strictly prefer x while others are indifferent between x and y, j can either strictly prefer x or she can be indifferent. Even though there is no complete agreement between the other ideal advisors, they broadly agree on the reasonableness of the risks involved in these prospects, and j is free to make her own judgment within this broad agreement. Second, if all the other ideal advisors are indifferent between x and y, then j can either be indifferent between these prospects, or she can prefer x to y. The interpretation of the former case is straightforward: if everybody else thinks that with respect to the reasonableness of the risks involved the two prospects are equivalent, then the ideal advisor whose preferences are to be determined also has reason to think that the prospects are equivalent in this respect.

The interpretation of the latter case, in which even though all the other ideal advisors are indifferent between the prospects but j nonetheless strictly prefers one of the prospects, is a bit puzzling. It is a bit puzzling because it is unclear whether we can say that the ideal advisors converge on j's preference given that j strictly prefers x to y, while all the others do not. Perhaps converging on j's preference in this case means that the other ideal advisors, by virtue of their indifference between the two prospects, essentially agree that it does not make a difference what preference j forms with respect to the reasonableness of the risks involved; that is, she can choose to prefer either of them. But perhaps their indifference entails that j should also be indifferent between the two prospects. (I will return to this issue below.)

Third, it is possible that all the other ideal advisors strictly prefer one of the prospects. In this case, j can either strictly prefer the same prospect or she can be indifferent between them. Once again, the interpretation of the former case is straightforward: if everybody else thinks that with respect to the reasonableness of the risks involved prospect x is preferable, then the ideal advisor whose preferences

are to be determined also has reason to think that this prospect is preferable.

The latter case, however, causes problems. On the weak convergence criterion, even if all the other ideal advisors strictly prefer one of the prospects, j can still be indifferent between the prospects. If all the other ideal advisors agree that prospect x is better than prospect y with respect to the reasonableness of the risks involved, but ideal advisor j remains indifferent between x and y, then the ideal advisors cannot be said to converge on their preference. Therefore, the weak convergence requirement cannot exclude a problematic sort of disagreement between ideal advisors. If we choose this requirement to ensure the authority of the recommendations of ideal advisors, then the possibility that ideal advisors do not "genuinely" converge on their preferences remains open.

What we have found is that the strong convergence requirement is too strong and the weak convergence requirement is too weak. On the former, ideal advisors who are in broad, although incomplete, agreement are not considered to be converging on their preferences; on the latter, ideal advisors who are in complete agreement about the preference the person should have were she fully informed and ideally rational can be considered to converge on the person's fully informed and ideally rational preference, even if the person's fully informed and ideally rational preference is different from theirs. Perhaps strengthening the weak convergence requirement a bit might avoid these problems. One possibility would be this:

$$\forall i \in N,\ u_i(x) \geq u_i(y) \Rightarrow x \succ_j y.$$

This says that if all of the other ideal advisors strictly prefer x to y, or if some of them strictly prefer x to y, while some others are indifferent, or even if they are all indifferent, then j should strictly prefer x. Whether this is a plausible strengthening of the weak convergence requirement depends on the interpretation of the last case. One way of reading it is that if all ideal advisors are indifferent, j can freely choose her preference—save for being indifferent. But it seems implausible that j cannot be indifferent herself, given that all other ideal advisors are indifferent. All of them believe that the risks are equally reasonable, and j has no reason to prefer either.

Therefore, instead of strengthening the weak convergence requirement, we can combine the plausible forms of convergence the requirements we have so far surveyed warrant. I will call the resultant requirement the *strict convergence requirement*:

$$\forall i \in N,\ (u_i(x) \geq u_i(y) \Rightarrow x \succ_j y) \vee (u_i(x) = u_i(y) \Rightarrow x \sim_j y).$$

On this version, "converging on a preference" can take the following forms. If all ideal advisors strictly prefer one prospect over another, than j should also strictly prefer that prospect. Similarly, if all ideal advisors are indifferent between the two prospects, than j should also be indifferent. In these cases, ideal advisors

are in complete agreement about their preference. On the other hand, if some of them strictly prefer one prospect and some others weakly prefer that prospect, j should strictly prefer that prospect. In this case, ideal advisors agree that with respect to the reasonableness of the risks involved, one prospect is at least as good as another. But since some of them believes that this prospect is strictly better, j has reason to prefer it: after all, no-one believes it is worse, and some believes it is better.

Perhaps the strict convergence requirement could be weakened by allowing j to be indifferent between prospects x and y if no ideal advisor strictly prefers y, and the number of those who strictly prefer x is below a certain threshold. Requiring that j strictly prefer x even if, at the extreme, only one ideal advisor prefers it strictly reflects a "conservative" strategy of forming judgments about the reasonableness of risks. The general idea is that in any situation, if many fully informed and ideally rational agents would agree that the risks involved are all reasonable—that the person is permitted to choose any of the options open to her—but some believe that some risks are more reasonable, then the person ought to choose the prospect involving those risks.[6]

Of course, nothing of what I have said ensures that ideal advisors will in all cases agree. Perhaps we still have no reason to expect that their preferences will converge. But the strict convergence requirement sets broad conditions for convergence. This reduces the scope for disagreement. Moreover, recall that in the cases relevant here, ideal advisors form their preferences for a particular person, in the context of that person's conception of the good, including that person's conception of reasonable risks. They do not form their preference for themselves or for their own actual counterparts. Their judgments are formed in the context of one particular life with respect to the values and goals of the person living that life. In that context, it is not unlikely that they could agree on their preferences.

10.2 Welfare Judgments and Risk

On my revision of the ideal advisor theory, a person's well-being is promoted if and only if the preferences she would have were she fully informed and ideally rational are satisfied, and, insofar as promoting her well-being involves risks, other fully informed and ideally rational persons would strictly converge on those preferences.

The notion of convergence employed in this theory is *normative*. It is normative because ideal advisors are envisaged as being engaged in a process of reaching

[6]I leave the possibility of further weakening the strict convergence requirement open. Incidentally, note that when I call this strategy of determining fully informed and ideally rational preferences for risky prospects "conservative," I do not mean to say that they should reflect risk-aversion or any other risk-disposition.

at least partial agreement on the reasonableness of risks. There is, however, another notion of convergence which has been employed in ideal advisor theories. This notion of convergence is *descriptive*. One theory which incorporates it is John C. Harsanyi's theory of well-being, which I discussed on pages 74 and 91. On his theory, fully informed and ideally rational persons have the same *extended preferences* over extended alternatives. Extended alternatives are whole lives, including the personal characteristics of the persons living those lives, and the causal variables which determine the preferences and characteristics of the persons living those lives. Fully informed and ideally rational persons would rank such alternatives the same way.

In Harsanyi's view, there are two reasons why fully informed and ideally rational persons would converge on their extended preferences. First, he thinks that human beings have highly similar biological and psychological needs, and hence the same set of basic desires. This is an *empirical fact* about human nature. Second, human beings are governed by the same psychological laws, and even if our current knowledge of these psychological laws is far from perfect, the individual differences in preferences must be attributed to variables which are, in principle, empirically measurable. Indeed, sound scientific practice requires that we do not attribute differences of preference and behavior to unobservable hidden variables, and we do not explain the effects of some observable variable by appealing to some unobservable variable. Thus, ultimately, individual differences of preference are susceptible to scientific explanation.[7]

Extended preferences play a crucial role in establishing the possibility of interpersonal comparisons of utility. In modern utility theory, interpersonal comparisons are not possible, since individual utility functions have no common origin and unit. Thus, if utility theory is used in a preference satisfaction theory of well-being, the types of *welfare judgments* which can be made on the theory are rather limited. In particular, even if expected utility functions are used to represent the preferences of fully informed and ideally rational persons, judgments comparing the welfare of different persons cannot be made. But if ideal advisors converge on their extended preferences, then a common utility function can be assigned to them, which expresses their *common* judgments about the values of states of affairs. Thus, based on extended preferences, such judgments as "state of affairs x is better for person i (with his preferences, circumstances, psychology, etc.), than state of affairs y for person j (with her preferences, circumstances, psychology, etc.)" can be made. Ultimately, interpersonal comparisons of utility, in virtue of

[7]For the first reason, see Harsanyi (1992, 1995, 1997). As I argued on page 75, perhaps human beings do have similar basic desires, but a theory of well-being cannot be based on those desires only. For the second reason, see Harsanyi (1955:316–9). He calls the former requirement on scientific practice the *principle of unwarranted differentiation*, and the latter the *principle of unwarranted correlation*. See also Weymark (1995).

these common judgments, are reduced to *intrapersonal* comparisons of utility.[8]

The possibility of convergence on extended preferences is widely rejected today.[9] But, in any case, the "open question" objection is especially relevant to the descriptive notion of convergence that Harsanyi uses. Even if human beings share basic desires, these desires must be extremely general. And even if we can expect to understand preference formation better as our knowledge of psychology grows, there is no reason to expect that the alleged psychological laws governing preference formation unambiguously determine preferences.

Another attempt to make the problem of interpersonal comparisons of utility (and welfare) tractable is to develop measures based on *preference intensities*. The notion of preference intensity is understood here as a *psychological primitive*, on the basis of which people compare the desirability of different possible outcomes. It does not need to have an hedonist interpretation. Arguably, if such intensities of preference could be measured, the problem arising from the lack of common unit and origin of utility scales in modern axiomatic expected utility theory may be circumvented. Perhaps some common scale for these intensities for different persons could be established.[10] But such projects face the following problem. As I explained in Section 2.2, expected utility functions are invariant under positive affine transformations. Such transformations preserve the ratios between utility differences. However, these ratios cannot be interpreted to express intensities of preference. One reason is that the utility values are established by reactions to risk. Thus, even if intensity of preference has an influence on a person's preferences in establishing her utility function, her risk-attitude also plays a role. In this framework, it is impossible to delineate the effects of these two factors.

One way of solving this problem may be to argue that expected utility functions represent intensity of preference, even though they are established by measuring reactions to risk. If such an argument could be made, cardinal utility in the sense it is used in modern utility theory and the representation of outcomes in terms of intensity of preference as a psychological primitive would be co-extensive. Interestingly, Harsanyi does make such an argument. He says,

> it is the decision makers' *cardinal utilities* (outcome utilities) for various alternatives that determine their (instrumental) *willingness to take risks* in

[8] See Harsanyi (1975*a*,*b*). Note that Harsanyi's discussion takes place within the context of the Neumann-Morgenstern version of expected utility theory. Of course, there are many extensions of that theory, and many versions of modern axiomatic expected utility theory in general. But their discussion is beyond the scope of this work.

[9] See, for example, Hammond (1990, 1991:221–4) and Broome (1998). Note also that extended preferences are not the same as *fundamental preferences*, at least not in the sense the notion is used by Kolm (1994), for whom it is a representation of happiness. See also Broome (1993, 1994).

[10] For an overview, see Hammond (1991:215–8). On intensity of preference, see Schoemaker (1982:533–5).

> order to obtain some desirable alternatives. These cardinal utilities determine their attitude toward risk taking, rather than the other way round. (1993:318, his emphases)[11]

In effect, what Harsanyi claims is that a person's intensities of preference between outcomes is equivalent to the *relative importance* that she places upon those outcomes, and her cardinal expected utility function represents both (1993:315). But, on the most natural understanding of the concepts of "relative importance" and "intensity of preference," this would be extraordinary. Insofar as a decision maker compares the values of various outcomes by her reactions to risk, she can be interpreted as ranking the outcomes by their relative importance to her. But insofar as intensity of preference is taken to be a psychological primitive, available through introspection, it must be a very different sort of quantity.

Psychologists trying to measure that "psychological primitive" have found that even if people can report their preferences in terms of intensity, the resulting representation is very different from the expected utility representation of their preferences in terms of reactions to risky choices.[12] This is not surprising, given that asking subjects for their introspective judgments about the desirability of certain goods, especially in terms of the intensity of the desirability of those goods for them, seems to be a very different exercise from asking for their judgments of desirability given that the goods are presented as possible "prizes" of gambles. In the former case, subjects are asked to arrive at their judgments by introspection; in the latter case, they are asked to arrive at their judgments by reflection.

In general, people *reason* about the risks they face and the relative importance of goods in risky situations. They do not simply form their preferences by introspectively comparing the desirability of goods. Their risk-attitudes are not fixed or predetermined, but, more often than not, the result of reasoning.

A third strategy to tackle the problem of welfare judgments may be to note that there is a relatively high degree of consensus on particular *goods* which promote welfare, and to build on that consensus. Even though philosophers are sharply divided over in virtue of what something is good for a person, they are in broad agreement, along with people in general, about which particular things are good for people. Most people would agree that health, income, and strong social relations, for example, contribute to a person's welfare. Thus, we have reason to think that there are goods which are correlated with welfare and can serve as its proxies.

Several branches of the social sciences are interested in designing and applying indicators for the measurement of welfare. They include several branches of

[11] "Outcome utilities" are the utilities that the person derives from the various outcomes, as opposed to "process utilities," which arise from the psychological experiences due to the act of gambling (sometimes called "the utility of gambling"). For a similar interpretation of Harsanyi's 1993 argument, see Ng (1999); see also Weymark's account of the Harsanyi-Sen debate (1991).

[12] See Kahneman (1999:17–9).

economics, sociology, and psychology. But even though attempts to develop indices to measure welfare go back several decades, there is little agreement on the methodology of welfare measurement, and a lot of work remains to be done.

The problems reflect the uncertainties about welfare judgments in philosophy. No theory of welfare that I am familiar with can claim that it has solved the problems involved in welfare judgments and welfare measurement. The revised IRP fares no worse in this respect. And perhaps, at least in certain cases, it can fare better. If people can agree on some indices of goods for representing well-being, perhaps they can also agree on the reasonable risks with respect to those goods. Thus, in specific situations, they can come up with a common scale for representing the relative importance of those goods, and welfare judgments can be made in terms of that scale. This method would not assume that people have the same preferences in an empirical sense, and it does not introduce a mysterious psychological entity. Of course, this suggestion is tentative, but it is a possible direction for further research.

10.3 Conclusion

In Section 1.1, I set out to develop a theory of well-being that is faithful to both of the subjectivist and the objectivist intuitions. These intuitions, I claimed, underlie the distinction between subjective and objective theories of welfare, and the initial plausibility of both sorts of theory stems from them. The subjectivist intuition holds that well-being must be connected to preferences; the objectivist intuition holds that what is good for us is good for us due to some factor other than our preferences. I claimed that the theory that I had developed, the revised IRP, is indeed faithful to both intuitions: it connects well-being, on the one hand, to the preferences a person would have were she fully informed and ideally rational, and, on the other hand, to the agreement of all fully informed and ideally rational agents—thus, what is good for a person is ultimately independent of *that person's* preferences.

But what sort of theory is the revised IRP? Like other versions of the ideal advisor theory, it is *reductionist*: it holds that judgments about what is good for a person can be reduced to norms of rationality, information, and convergence. But, in contrast to many other versions of idealization theories—and the ideal advisor theory in particular—this theory is not *naturalist*. Since it incorporates a convergence requirement, it appeals to a constraint on the preferences which are relevant to well-being that, on my interpretation of the requirement, is explicitly normative. Thus, perhaps the theory can be characterized as a reductionist and *constructivist* theory of well-being.

One objection might be that such a theory is *circular*. It is circular because even though it purports to be a preference satisfaction theory of well-being, it includes substantive judgments—that is, it appeals to the objects of at least some

preferences. It also incorporates a normative idea of convergence. Furthermore, it identifies self-regarding preferences in a circular manner, since it appeals to what a person, were she fully informed and ideally rational, would prefer for the reason that the satisfaction of that preference would make that person better off.

But I think the charge of circularity is only impressive from the subjectivist perspective. Once we realize that a preference satisfaction theory must appeal to constraints on preferences beyond the epistemic and the cognitive constraints, it is unclear why the alleged circularity involved in other constraints is threatening. And once we reject the distinction between objective and subjective theories, the perspective from which the revised IRP looks circular is undermined.

As I see it, preference has an indispensable role in a theory of well-being. Even if you accept an objective theory, you must appeal to preferences—because it is hard to see how informed and rational preferences would not be useful indicators of what is good for a person on such a theory. Preferences, on all theories of welfare, must have at least an indirect role.[13]

There are many questions which this work leaves unanswered. These provide starting points for further research. Two of these I have already mentioned: the problem of welfare judgments, and the possibility of working out more precise norms for the reasonableness of risks. Both are necessary to be able to explore the implications of the ideas I have explored for the promotion of welfare and the relations of well-being and risk to social policy.

Well-being has often been discussed by philosophers. The problem of risk, however, has been largely neglected in philosophy.[14] This work has argued that there are interesting connections between well-being and risk.

[13]Compare the quote from Scanlon on page 120. Incidentally, note that if informed and rational preferences do have such a role, then some of the arguments I discussed in Section 7.3, if successful, would cause problems to objectivist theories of welfare as well.

[14]For exceptions, see Altham (1984), and Broome (1991*b*).

Bibliography

Allais, Maurice. (1953) “Le Comportement de l’Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l’École Americaine.” *Econometrica* 21(4), 503–46.

——. (1984*a*) “Determination of Cardinal Utility According to an Intrinsic Invariant Model.” In Allais and Hagen (1994), 31–64.

——. (1984*b*) “The Foundations of the Theory of Utility and Risk: Some Central Points of the Discussions at the Oslo Conference.” In Hagen and Wenstøp (1984), 3–131.

——. (1988) “Cardinal Utility: History, Empirical Findings, and Applications—An Overview.” In Allais and Hagen (1994), 65–103.

——. (1994) “Absolute Satisfaction.” In Allais and Hagen (1994), 1–29.

Allais, Maurice, and Ole Hagen, eds. (1994) *Cardinalism: A Fundamental Approach*. Dordrecht: Kluwer Academic Publishers.

Altham, J. E. J. (1984) “Ethics of Risk.” *Proceedings of the Aristotelian Society* 84, 15–29.

Anderson, Elizabeth. (1993) *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Arneson, Richard J. (1999) “Human Flourishing versus Desire Satisfaction.” *Social Philosophy & Policy* 16(1), 113–42.

Arrow, Kenneth J. (1977) “Extended Sympathy and the Possibility of Social Choice.” *The American Economic Review* 67(1), 219–25.

Ashmore, Malcolm. (1993) “The Theatre of the Blind: Starring a Promethean Prankster, a Phoney Phenomenon, a Prism, a Pocket, and a Piece of Wood.” *Social Studies of Science* 23(1), 67–106.

Attfield, Robin. (1987) *A Theory of Value and Obligation*. London: Croom Helm.

Baldwin, Thomas. (1993) “Editor’s Introduction.” In Moore (1903), *ix–xxxvii*.

Barberà, Salvador, Peter J. Hammond, and Christian Seidl, eds. (1999) *Handbook of Utility Theory, Volume 1: Principles*. Boston: Kluwer Academic Press.

Bentham, Jeremy. (1789) *An Introduction to the Principles of Morals and Legislation*. London: Athlone Press, 1970. Ed. by J. H. Burns and H. L. A. Hart.

Bernoulli, Daniel. (1738) “Exposition of a New Theory on the Measurement of Risk.” *Econometrica* 22(1), (1954), 23–36. Trans. by Louise Sommer. Repr. in Page (1968), 199–214.

Bernstein, Mark. (1998) “Well-Being.” *American Philosophical Quarterly* 35(1), 39–55.

Binmore, Ken. (1994) *Game Theory and the Social Contract, Volume I: Playing Fair*. Cambridge, MA: MIT Press.

Binmore, Ken, Alan Kirman, and Piero Tani, eds. (1993) *Frontiers of Game Theory*. Cambridge, MA: MIT Press.

Bok, Sissela. (1980) *Lying: Moral Choice in Public and Private Life*. London: Quartet Books.

Bond, E. J. (1983) *Reason and Value*. Cambridge: Cambridge University Press.

Brandt, Richard B. (1959) *Ethical Theory*. Englewood Cliffs, NJ: Prentice-Hall.

——. (1979) *A Theory of the Good and the Right*. Oxford: Clarendon Press.

——. (1982) “Two Concepts of Utility.” In Brandt (1992), 158–75.

——. (1992) *Morality, Utilitarianism, and Rights*. Cambridge: Cambridge University

Press.
——. (1998) “The Rational Criticism of Preferences.” In Fehige and Wessels (1998), 63–77.
Breault, K. (1981) “Modern Psychophysical Measurement of Marginal Utility: A Return to Introspective Cardinality?” *Social Science Quarterly* 62(4), 672–84.
Brink, David O. (1989) *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.
Broad, C. D. (1934) “Is ‘Goodness’ a Name of a Simple Non-Natural Quality?” In Cheney (1971), 106–23.
Brock, Dan W. (1983) “Paternalism and Promoting the Good.” In Sartorius (1983), 237–60.
Broome, John. (1991*a*) “Utility.” In Broome (1999*a*), 19–28.
——. (1991*b*) *Weighing Goods: Equality, Uncertainty and Time*. Oxford: Basil Blackwell.
——. (1993) “A Cause of Preference is not an Object of Preference.” *Social Choice and Welfare* 10(1), 57–68.
——. (1994) “Reply to Kolm.” *Social Choice and Welfare* 11, 199–201.
——. (1998) “Extended Preferences.” In Fehige and Wessels (1998), 271–87. Repr. in Broome (1999*a*), 29–43.
——. (1999*a*) *Ethics Out of Economics*. Cambridge: Cambridge University Press.
——. (1999*b*) “Introduction.” In Broome (1999*a*), 1–18.
Buchanan, Allen. (1978) “Medical Paternalism.” *Philosophy & Public Affairs* 7(4), 370–90.
Butler, Joseph. (1726) “Upon the Love of Our Neighbour (*Fifteen Sermons*).” In Raphael (1969), 364–77.
Calcott, Paul. (2000) “New on Paternalism.” *Economics and Philosophy* 16(2), 315–21.
Camerer, Colin. (1995) “Individual Decision Making.” In Kagel and Roth (1995), 587–703.
Carley, Michael. (1981) *Social Measurement and Social Indicators: Issues of Policy and Theory*. London: George Allen & Unwin.
Cheney, David R., ed. (1971) *Broad’s Critical Essays in Moral Philosophy*. London: George Allen & Unwin.
Cowen, Tyler. (1993) “The Scope and Limits of Preference Sovereignty.” *Economics and Philosophy* 9(2), 253–69.
Crisp, Roger, and Brad Hooker, eds. (2000) *Well-Being and Morality: Essays in Honour of James Griffin*. Oxford: Clarendon Press.
Daboni, L., A. Montesano, and M. Lines, eds. (1986) *Recent Developments in the Foundations of Utility and Risk Theory*. Dordrecht: D. Reidel.
Daniels, Norman. (1983) “Can Cognitive Psychotherapy Reconcile Reason and Desire?” *Ethics* 93(4), 772–85.
Darwall, Stephen. (1983) *Impartial Reason*. Ithaca, NY: Cornell University Press.
——. (2002) *Welfare and Rational Care*. Princeton, NJ: Princeton University Press.
DePaul, Michael R. (2002) “A Half Dozen Puzzles Regarding Intrinsic Attitudinal Hedonism.” *Philosophy and Phenomenological Research* 65(3), 629–35.
Duncker, Karl. (1940) “On Pleasure, Emotion, and Striving.” *Philosophy and Phenomenological Research* 1(4), 391–430.
Dworkin, Gerald. (1972) “Paternalism.” In Sartorius (1983), 19–34.

——. (1983) “Paternalism: Some Second Thoughts.” In Sartorius (1983), 105–11.
Dworkin, Ronald. (1981) “What Is Equality? Part 1: Equality of Welfare.” *Philosophy & Public Affairs* 10(3), 185–246.
——. (2000) *Sovereign Virtue: The Theory and Practice of Equality*. Cambridge, MA: Harvard University Press.
Ellingsen, Tore. (1994) “Cardinal Utility: A History of Hedonimetry.” In Allais and Hagen (1994), 105–65.
Ellsberg, D. (1954) “Classic and Current Notions of ‘Measurable Utility.’ ” *The Economic Journal* 64(255), 528–56. Repr. in Page (1968), 269–96.
Elster, Jon, and Aanund Hylland, eds. (1986) *Foundations of Social Choice Theory*. Cambridge: Cambridge University Press.
Elster, Jon, and John E. Roemer, eds. (1991) *Interpersonal Comparisons of Well-Being*. Cambridge: Cambridge University Press.
Fehige, Christoph, and Ulla Wessels, eds. (1998) *Preferences*. Berlin: Walter de Gruyter.
Feinberg, Joel. (1971) “Legal Paternalism.” In Sartorius (1983), 3–18.
——. (1986) *Harm to Self: The Moral Limits of the Criminal Law*. Oxford: Oxford University Press.
Feldman, Fred. (1988) “Two Questions about Pleasure.” In Feldman (1997*b*), 82–105.
——. (1996) “Mill, Moore, and the Consistency of Qualified Hedonism.” In Feldman (1997*b*), 108–24. Repr. from *Midwest Studies in Philosophy* 20 (1996), 318–31.
——. (1997*a*) “On the Intrinsic Value of Pleasures.” In Feldman (1997*b*), 127–47. Repr. from *Ethics* 107(3), (1997), 448–66.
——. (1997*b*) *Utilitarianism, Hedonism, and Desert*. Cambridge: Cambridge University Press.
——. (2002*a*) “Comments on Two of DePaul’s Puzzles.” *Philosophy and Phenomenological Research* 65(3), 636–9.
——. (2002*b*) “The Good Life: A Defense of Attitudinal Hedonism.” *Philosophy and Phenomenological Research* 65(3), 604–28.
Finnis, John. (1980) *Natural Law and Natural Rights*. Oxford: Clarendon Press.
——. (1983) *Fundamentals of Ethics*. Oxford: Clarendon Press.
Fishburn, Peter C. (1968) “Utility Theory.” *Management Science* 14(5), 335–78.
——. (1981) “Subjective Expected Utility: A Review of Normative Theories.” *Theory and Decision* 13(2), 139–99.
Fisher, Irving. (1918) “Is ‘Utility’ the Most Suitable Term for the Concept It is Used to Denote?” *The American Economic Review* 8(2), 335–7.
Fotion, Nicholas. (1979) “Paternalism.” *Ethics* 89(2), 191–8.
Friedman, Milton, and Leonard J. Savage. (1948) “The Utility Analysis of Choices Involving Risk.” *The Journal of Political Economy* 56(4), 279–304.
——. (1952) “The Expected-Utility Hypothesis and the Measurability of Utility.” *The Journal of Political Economy* 60(6), 463–74.
Gallie, W. B. (1954) “Pleasure.” *Proceedings of the Aristotelian Society, Supplementary Volume* 28, 147–64.
Gauthier, David. (1986) *Morals by Agreement*. Oxford: Clarendon Press.
Gert, Bernard, and Charles M. Culver. (1976) “Paternalistic Behavior.” *Philosophy & Public Affairs* 6(1), 45–57.
——. (1979) “The Justification of Paternalism.” *Ethics* 89(2), 199–210.

Gibbard, Alan. (1990) *Wise Choices, Apt Feelings*. Oxford: Clarendon Press.
——. (1998) "Preference and Preferability." In Fehige and Wessels (1998), 239–59.
Glover, Jonathan. (1977) *Causing Death and Saving Lives*. Harmondsworth: Penguin Books.
Goldstein, Irwin. (1985) "Hedonic Pluralism." *Philosophical Studies* 48, 49–55.
Goldsworthy, Jeffrey. (1992) "Well-Being and Value." *Utilitas* 4(1), 1–26.
Goodin, Robert E. (1986) "Laundering Preferences." In Elster and Hylland (1986), 75–101.
——. (1990) "Liberalism and the Best Judge Principle." *Political Studies* 38(2), 181–95.
——. (1991) "Permissible Paternalism: In Defence of the Nanny State." *The Responsive Community* 1(3), 42–51.
——. (1993) "Democracy, Preferences, and Paternalism." *Policy Sciences* 26, 229–47.
——. (2002) *Reflective Democracy*. Oxford: Oxford University Press.
Gorovitz, Samuel, ed. (1971) *Utilitarianism with Critical Essays*. Indianapolis, IN: Bobbs-Merrill.
Griffin, James. (1986) *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press.
——. (1996) *Value Judgement: Improving Our Ethical Beliefs*. Oxford: Clarendon Press.
——. (2000) "Replies." In Crisp and Hooker (2000), 281–313.
Hagen, Ole. (1984) "Neo-Cardinalism." In Hagen and Wenstøp (1984), 145–64.
——. (1994) "The Short Step from Ordinal to Cardinal Utility." In Allais and Hagen (1994), 209–21.
Hagen, Ole, and Fred Wenstøp, eds. (1984) *Progress in Utility and Risk Theory*. Dordrecht: D. Reidel.
Hammond, Peter J. (1990) "Interpersonal Comparisons of Utility: Why and How They are and should be Made." EUI Working Paper ECO 90/3, European University Institute, Florence.
——. (1991) "Interpersonal Comparisons of Utility: Why and How They are and should be Made." In Elster and Roemer (1991), 200–54. Extended version of Hammond (1990).
Hardin, Russell. (1988) *Morality within the Limits of Reason*. Chicago: University of Chicago Press.
Hare, Richard M. (1952) *The Language of Morals*. Oxford: Clarendon Press.
——. (1981) *Moral Thinking: Its Method, Levels, and Point*. Oxford: Clarendon Press.
Harrod, Roy F. (1936) "Utilitarianism Revised." *Mind* 45(178), 137–56.
Harsanyi, John C. (1953) "Cardinal Utility in Welfare Economics and in the Theory of Risk-taking." *The Journal of Political Economy* 61(5), 434–5.
——. (1955) "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *The Journal of Political Economy* 63(4), 309–21.
——. (1975*a*) "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory." *American Political Science Review* 59(2), 594–606.
——. (1975*b*) "Nonlinear Social Welfare Functions: Do Welfare Economists have a Special Exemption from Bayesian Rationality?" *Theory and Decision* 6, 311–32.
——. (1977*a*) *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.
——. (1977*b*) "Rule-Utilitarianism and Decision Theory." *Erkenntnis* 11(1), 25–53.
——. (1978) "Bayesian Decision Theory and Utilitarian Ethics." *The American Economic*

Review 68(2), 223–8.
——. (1982) “Morality and the Theory of Rational Behaviour.” In Sen and Williams (1982), 39–62.
——. (1985) “Does Reason Tell Us What Moral Code to Follow and, Indeed, to Follow Any Moral Code at All?” *Ethics* 96(1), 42–55.
——. (1992) “Utilities, Preferences, and Substantive Goods.” WIDER Working Papers 101, Helsinki.
——. (1993) “Normative Validity and Meaning of von Neumann-Morgenstern Utilities.” In Binmore *et al.* (1993), 307–20.
——. (1995) “A Theory of Prudential Values and a Rule Utilitarian Theory of Morality.” *Social Choice and Welfare* 12(4), 319–33.
——. (1997) “Utilities, Preferences, and Substantive Goods.” *Social Choice and Welfare* 14(1), 129–45. Extended version of Harsanyi (1992).
Haslett, D. W. (1990) “What is Utility?” *Economics and Philosophy* 6(1), 65–94.
Hicks, J. R. (1939) “The Foundations of Welfare Economics.” *The Economic Journal* 49(196), 696–712.
Hubin, Donald C. (1996) “Hypothetical Motivation.” *Noûs* 30(1), 31–54.
——. (1999) “Converging on Values.” *Analysis* 59(4), 355–61.
Husak, Douglas N. (1981) “Paternalism and Autonomy.” *Philosophy & Public Affairs* 10(1), 27–46.
Jevons, William Stanley. (1871) *The Theory of Political Economy*. Fifth edn., 1957.
Kagan, Shelly. (1992) “The Limits of Well-Being.” *Social Philosophy & Policy* 9(2), 169–89.
Kagel, John H., and Alvin E. Roth, eds. (1995) *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
Kahn, Rober L., and F. Thomas Juster. (2002) “Well-Being: Concepts and Measures.” *Journal of Social Issues* 58(4), 627–44.
Kahneman, Daniel. (1999) “Objective Happiness.” In Kahneman *et al.* (1999), 3–25.
Kahneman, Daniel, Ed Diener, and Norbert Schwartz, eds. (1999) *Well-Being: The Foundations of Hedonic Psychology*. New York: Russell Sage.
Kahneman, Daniel, and Amos Tversky. (1979) “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47(2), 263–92.
Kawall, Jason. (1999) “The Experience Machine and Mental State Theories of Well-Being.” *Journal of Value Inquiry* 33(3), 381–7.
Kleinig, John. (1983) *Paternalism*. Manchester: Manchester University Press.
Kolm, Serge-Christophe. (1994) “The Meaning of ‘Fundamental Preferences.’ ” *Social Choice and Welfare* 11, 193–8.
Kraut, Richard. (1979) “Two Conceptions of Happiness.” *The Philosophical Review* 88(2), 167–97.
——. (1994) “Desire and the Human Good.” *Proceedings and Addresses of the American Philosophical Association* 68(2), 39–54.
Kreps, David M. (1988) *Notes on the Theory of Choice*. Boulder, CO: Westview Press.
Kusser, Anna. (1998) “Rational by Schock: A Reply to Brandt.” In Fehige and Wessels (1998), 78–87.
Kymlicka, Will. (1990) *Contemporary Political Philosophy: An Introduction*. Oxford: Clarendon Press.

Leonard, Thomas C., Robert S. Goldfarb, and Steven M. Suranovic. (2000) "New on Paternalism and Public Policy." *Economics and Philosophy* 16(2), 323–31.

Lewis, David. (1989) "Dispositional Theories of Value." *Proceedings of the Aristotelian Society, Supplementary Volume* 63, 113–37.

List, Christian. (2003) "Are Interpersonal Comparisons of Utility Indeterminate?" *Erkenntnis* 58(2), 229–60.

Little, Ian M. D. (1950) *A Critique of Welfare Economics*. Oxford: Oxford University Press. Second edn., 1957.

Loeb, Don. (1995) "Full-Information Theories of Individual Good." *Social Theory & Practice* 21(1), 1–30.

Lopes, Lola. (1986) "What Naive Decision Makers can Tell Us about Risk." In Daboni *et al.* (1986), 311–26.

Louden, Robert B. (1992) *Morality and Moral Theory: A Reappraisal and Reaffirmation*. New York: Oxford University Press.

Luce, R. Duncan, and Howard Raiffa. (1957) *Games and Decisions*. New York: John Wiley & Sons.

MacCrimmon, Kenneth R., and Donald A. Wehrung. (1986) "Assessing Risk Propensity." In Daboni *et al.* (1986), 291–309.

MacNiven, Don. (1993) *Creative Morality*. London: Routledge.

Marshall, Alfred. (1890) *Principles of Economics: An Introductory Volume*. Eighth edn., 1920.

McCloskey, Mary A. (1971) "Pleasure." *Mind* 80, 542–51.

Mill, John Stuart. (1861) "Utilitarianism." In Gorovitz (1971), 11–57.

Momeyer, Richard W. (1975) "Is Pleasure a Sensation?" *Philosophy and Phenomenological Research* 36(1), 113–21.

Mongin, Philippe. (2001) "The Impartial Observer Theorem of Social Ethics." *Economics and Philosophy* 17(2), 147–79.

Mongin, Philippe, and Claude d'Aspremont. (1999) "Utility Theory and Ethics." In Barberà *et al.* (1999).

Moore, G. E. (1903) *Principia Ethica*. Cambridge: Cambridge University Press, ed. by Thomas Baldwin; revised, 1993 edn.

Murphy, Mark C. (1999) "The Simple Desire-Fulfillment Theory." *Noûs* 33(2), 247–72.

Neufville, Judith Innes de. (1975) *Social Indicators and Public Policy: Interactive Processes of Design and Application*. Amsterdam: Elsevier Scientific Publishing.

Neumann, John von, and Oskar Morgenstern. (1944) *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press. Third edn., 1953.

New, Bill. (1999) "Paternalism and Public Policy." *Economics and Philosophy* 15(1), 63–83.

Ng, Yew-Kwang. (1999) "Utility, Informed Preference, or Happiness: Following Harsanyi's Argument to Its Logical Conclusion." *Social Choice and Welfare* 16, 197–216.

Noggle, Robert. (1999) "Integrity, the Self, and Desire-Based Accounts of the Good." *Philosophical Studies* 96(3), 303–31.

Nozick, Robert. (1974) *Anarchy, State, and Utopia*. New York: Basic Books.

——. (1989) *The Examined Life: Philosophical Meditations*. New York: Simon and Schuster.

Nussbaum, Martha C., and Amartya Sen, eds. (1993) *The Quality of Life*. Oxford: Claren-

don Press.
Overvold, Mark Carl. (1980) "Self-Interest and the Concept of Self-Sacrifice." *Canadian Journal of Philosophy* 10(1), 105–18.
Page, Alfred N., ed. (1968) *Utility Theory: A Book of Readings*. New York: John Wiley & Sons.
Pareto, Vilfredo. (1927) "Manuel D'Economie Politique (Ophélimité)." In Page (1968), 168–81, 375–83. Extracts from *Manuel D'Economie Politique*, Second Edn., 1927. Trans. by Ann Stranquist Schwier.
Parfit, Derek. (1984) *Reasons and Persons*. Oxford: Oxford University Press.
Penelhum, Terence. (1957) "The Logic of Pleasure." *Philosophy and Phenomenological Research* 17(4), 488–503.
Pennock, J. Roland, and John W. Chapman, eds. (1974) *The Limits of Law*. New York: Lieber-Atherton.
Perry, David L. (1967) *The Concept of Pleasure*. The Hague: Mouton & Co.
Qizilbash, Mozaffar. (1998) "The Concept of Well-Being." *Economics and Philosophy* 14(1), 51–73.
Quinn, Warren S. (1968) "Pleasure—Disposition or Episode?" *Philosophy and Phenomenological Research* 28(4), 578–86.
Railton, Peter. (1986*a*) "Facts and Values." *Philosophical Topics* 14(2), 5–31.
——. (1986*b*) "Moral Realism." *The Philosophical Review* 95(2), 163–207.
——. (1989) "Naturalism and Prescriptivity." *Social Philosophy & Policy* 7(1), 151–74.
Raphael, D. D., ed. (1969) *British Moralists*, vol. 1. Oxford: Clarendon Press.
Rawls, John. (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Raz, Joseph. (1986) *The Morality of Freedom*. Oxford: Clarendon Press.
Regan, Donald H. (1974) "Justifications for Paternalism." In Pennock and Chapman (1974), 189–210.
Richter, Marcel K. (1966) "Revealed Preference Theory." *Econometrica* 34(3), 635–45.
Robbins, Lionel. (1938) "Interpersonal Comparisons of Utility: A Comment." *Economic Journal* 48(192), 635–41.
Robertson, Ross M. (1951) "Jevons and His Precursors." *Econometrica* 19(3), 229–49.
Rohr, Michael David. (1978) "Is Goodness Comparative?" *The Journal of Philosophy* 75(9), 494–503.
Rosati, Connie S. (1995) "Persons, Perspectives, and Full Information Accounts of the Good." *Ethics* 105(2), 296–325.
Ryle, Gilbert. (1949) *The Concept of Mind*. Harmondsworth: Penguin Books.
——. (1954) "Pleasure." *Proceedings of the Aristotelian Society, Supplementary Volume* 28, 135–46.
Samuelson, Paul A. (1938*a*) "A Note on the Pure Theory of Consumer's Behaviour." *Economica* 5(17), 61–71.
——. (1938*b*) "A Note on the Pure Theory of Consumer's Behaviour: An Addendum." *Economica* 5(19), 353–4.
Sartorius, Rolf, ed. (1983) *Paternalism*. Minneapolis: University of Minnesota Press.
Savage, Leonard J. (1954) *The Foundations of Statistics*. New York: John Wiley & Sons.
Scanlon, Thomas M. (1991) "The Moral Basis of Interpersonal Comparisons." In Elster and Roemer (1991), 17–44.
——. (1993) "Value, Desire, and the Quality of Life." In Nussbaum and Sen (1993), 185–

200.
——. (1998) *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Schlick, Moritz. (1939) *Problems of Ethics*. New York: Prentice-Hall. Trans. by David Rynin.
Schoemaker, Paul J. H. (1982) "The Expected Utility Model: Its Variants, Purposes, Evidence, and Limitations." *Journal of Economic Literature* 20(2), 529–63.
Schüssler, Rudolf. (1998) "Wish You Were Me: A Reply to Broome and a Comment on Harsanyi's Extended Preference Theory." In Fehige and Wessels (1998), 288–97.
Schwarz, Norbert, and Fritz Strack. (1999) "Reports of Subjective Well-Being: Judgmental Processes and their Methodological Implications." In Kahneman *et al.* (1999), 61–84.
Scoccia, Danny. (1990) "Paternalism and Respect for Autonomy." *Ethics* 100(2), 318–34.
Seel, Martin. (1997) "Well-Being: On a Fundamental Concept of Practical Philosophy." *European Journal of Philosophy* 5(1), 39–49. Trans. by David Midgley.
Sen, Amartya. (1971) "Choice Functions and Revealed Preference." *Review of Economic Studies* 38(3), 307–17.
——. (1973) "Behaviour and the Concept of Preference." *Economica* 40(159), 241–59.
——. (1977) "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs* 6(4), 317–44.
——. (1980) "Equality of What?" In Sen (1982), 353–69. Repr. from *The Tanner Lectures on Human Values*, Vol. 1, 1980. Salt Lake City: The University of Utah Press.
——. (1982) *Choice, Welfare and Measurement*. Oxford: Basil Blackwell.
——. (1993) "Internal Consistency of Choice." *Econometrica* 61(3), 495–521.
——. (1994) "The Formulation of Rational Choice." *The American Economic Review* 84(2), 385–390.
Sen, Amartya, and Bernard Williams, eds. (1982) *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.
Sidgwick, Henry. (1907) *The Methods of Ethics* (Seventh edn.). Indianapolis, IN: Hackett (1981).
Silverstein, Matthew. (2000) "In Defense of Happiness: A Response to the Experience Machine." *Social Therory and Practice* 26(2), 279–300.
Smart, J. J. C. (1973) "An Outline of a System of Utilitarian Ethics." In Smart and Williams (1973), 3–74.
Smart, J. J. C., and Bernard Williams. (1973) *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.
Smith, Michael. (1987) "The Humean Theory of Motivation." *Mind* 96(381), 36–61.
——. (1988) "Reason and Desire." *Proceedings of the Aristotelian Society* 88, 243–58.
——. (1994) *The Moral Problem*. Oxford: Blackwell.
Smythe, Thomas. (1972) "Unconscious Desires and the Meaning of 'Desire.' " *The Monist* 56(3), 413–25.
Sobel, David. (1994) "Full Information Accounts of Well-Being." *Ethics* 104(4), 784–810.
——. (1998) "Summer on Welfare." *Dialogue* 37(3), 571–7.
——. (1999) "Do the Desires of Rational Agents Converge?" *Analysis* 59(3), 137–47.
——. (2001) "Subjective Accounts of Reasons for Action." *Ethics* 111(3), 461–92.
Sobel, David, and David Copp. (2001) "Against Direction of Fit Accounts of Belief and Desire." *Analysis* 61(1), 44–53.
Sober, Elliott. (1992) "Hedonism and Butler's Stone." *Ethics* 103(1), 97–103.

Stigler, George J. (1950*a*) "The Development of Utility Theory I." *The Journal of Political Economy* 58(4), 307–27.

——. (1950*b*) "The Development of Utility Theory II." *The Journal of Political Economy* 58(5), 373–96.

Strack, Fritz, Leonard L. Martin, and Norbert Schwarz. (1988) "Priming and Communication: Social Determinants of Information Use in Judgments of Life Satisfaction." *European Journal of Social Psychology* 18(5), 429–42.

Sturgeon, Nicholas L. (1982) "Brandt's Moral Empiricism." *The Philosophical Review* 91(3), 389–422.

Sumner, L. Wayne. (1992*a*) "Two Theories of the Good." *Social Philosophy & Policy* 9(2), 1–14.

——. (1992*b*) "Welfare, Happiness, and Pleasure." *Utilitas* 4(2), 199–223.

——. (1995) "The Subjectivity of Welfare." *Ethics* 105(4), 764–90.

——. (1996) *Welfare, Happiness, and Ethics*. Oxford: Clarendon Press.

——. (2000) "Something in Between." In Crisp and Hooker (2000), 1–19.

Thomson, Garrett. (1987) *Needs*. London: Routledge & Kegan Paul.

VanDeVeer, Donald. (1980) "Autonomy Respecting Paternalism." *Social Theory and Practice* 6(2), 187–207.

——. (1986) *Paternalistic Intervention: The Moral Bounds on Benevolence*. Princeton, NJ: Princeton University Press.

Velleman, J. David. (1988) "Brandt's Definition of 'Good.' " *The Philosophical Review* 97(3), 353–71.

Weale, Albert. (1978) "Paternalism and Social Policy." *Journal of Social Policy* 7(2), 157–72.

Weymark, John A. (1991) "A Reconsideration of the Harsanyi-Sen Debate on Utilitarianism." In Elster and Roemer (1991), 255–320.

——. (1995) "John Harsanyi's Contributions to Social Choice and Welfare Economics." *Social Choice and Welfare* 12(4), 313–8.

White, Alan R. (1958) *G. E. Moore: A Critical Exposition*. Oxford: Basil Blackwell.

Wiggins, David. (1987) "Claims of Need." In Wiggins (1998), 1–57, 314–28.

——. (1998) *Needs, Values, Truth*. Oxford: Clarendon Press, third edn.

Williams, Bernard. (1980) "Internal and External Reasons." In Williams (1981), 101–13.

——. (1981) *Moral Luck*. Cambridge: Cambridge University Press.

Wright, Georg Henrik von. (1963) *The Varieties of Goodness*. London: Routledge & Kegan Paul.

Zapf, Wolfgang. (2000) "Social Reporting in the 1970s and in the 1990s." *Social Indicators Research* 51(1), 1–15.

List of Citations

Allais (1953), 23
Allais (1984*a*), 23, 25
Allais (1984*b*), 23, 24
Allais (1988), 23, 25
Allais (1994), 23–25
Altham (1984), 146
Anderson (1993), 86
Arneson (1999), 1, 5
Arrow (1977), 92
Ashmore (1993), 38
Attfield (1987), 29
Baldwin (1993), 43
Bentham (1789), 15–17, 43
Bernoulli (1738), 17
Bernstein (1998), 1, 30
Binmore (1994), 23
Bok (1980), 124
Bond (1983), 1
Brandt (1959), 44, 45
Brandt (1979), 1, 40, 44, 45, 86–89, 93
Brandt (1982), 49–52
Brandt (1998), 89
Breault (1981), 23
Brink (1989), 1, 29, 44
Broad (1934), 44
Brock (1983), 125
Broome (1991*a*), 17, 48
Broome (1991*b*), 146
Broome (1993), 143
Broome (1994), 143
Broome (1998), 92, 143
Broome (1999*b*), 72
Buchanan (1978), 124
Butler (1726), 28
Calcott (2000), 125
Camerer (1995), 24
Carley (1981), 60
Cowen (1993), 86
Daniels (1983), 89
Darwall (1983), 86
Darwall (2002), 1
DePaul (2002), 32
Duncker (1940), 43, 44
Dworkin (1972), 124, 126–130
Dworkin (1981), 2
Dworkin (1983), 124–128, 130
Dworkin (2000), 3
Ellingsen (1994), 18
Ellsberg (1954), 19, 48, 51
Fehige and Wessels (1998), 72
Feinberg (1971), 127–130, 138
Feinberg (1986), 127, 129, 130
Feldman (1988), 28, 43
Feldman (1996), 28
Feldman (1997*a*), 28, 45
Feldman (2002*a*), 32
Feldman (2002*b*), 1, 31, 32, 45
Finnis (1980), 1, 29
Finnis (1983), 29
Fishburn (1968), 19
Fishburn (1981), 19
Fisher (1918), 17
Fotion (1979), 124
Friedman and Savage (1948), 13
Friedman and Savage (1952), 13, 16, 25, 26
Gallie (1954), 45
Gauthier (1986), 86
Gert and Culver (1976), 124
Gert and Culver (1979), 127
Gibbard (1990), 86, 89
Gibbard (1998), 72
Glover (1977), 57
Goldstein (1985), 45
Goldsworthy (1992), 1, 30
Goodin (1986), 73
Goodin (1990), 27
Goodin (1991), 125
Goodin (1993), 125
Goodin (2002), 125
Griffin (1986), 1, 2, 29, 42, 44, 45, 57, 71, 86
Griffin (1996), 1, 2, 86
Griffin (2000), 1, 86
Hagen (1984), 23
Hagen (1994), 23
Hammond (1990), 143

Schoemaker (1982), 19, 143
Schwarz and Strack (1999), 64
Scoccia (1990), 127
Seel (1997), 1
Sen (1971), 26
Sen (1973), 26
Sen (1977), 72
Sen (1980), 42
Sen (1993), 26
Sen (1994), 26
Sidgwick (1907), 42–45, 52, 86
Silverstein (2000), 30, 40
Smart (1973), 29, 72
Smith (1987), 70
Smith (1988), 70
Smith (1994), 70, 83, 90, 91, 98, 137
Smythe (1972), 70
Sobel and Copp (2001), 70
Sobel (1994), 86, 94–96
Sobel (1998), 57
Sobel (1999), 90, 137
Sobel (2001), 87
Sober (1992), 28
Stigler (1950*a*), 18
Stigler (1950*b*), 18
Strack *et al*. (1988), 64
Sturgeon (1982), 89
Sumner (1992*a*), 1
Sumner (1992*b*), 59
Sumner (1995), 1, 3, 4
Sumner (1996), 1, 3, 4, 10, 11, 29, 33, 35, 43, 56–67, 78
Sumner (2000), 1, 45, 56
Thomson (1987), 29
VanDeVeer (1980), 124, 127
VanDeVeer (1986), 124–127
Velleman (1988), 86, 89, 93
Weale (1978), 125
Weymark (1991), 144
Weymark (1995), 142
White (1958), 43
Wiggins (1987), 1
Williams (1980), 68, 90, 91
Wright (1963), 45
Zapf (2000), 60

Index