ference-Based Views of Well-Being

ister Bykvist

e Oxford Handbook of Well-Being and Public Policy

lited by Matthew D. Adler and Marc Fleurbaey

nt Publication Date:
Jun 2016

bject:
Economics and Finance, Health, Education, and Welfare

ine Publication Date:
Nov 2016

l:
10.1093/oxfordhb/9780199325818.013.10

Preference-Based Views of Well-Being

Abstract and Keywords

This chapter deals with preference-based views of well-being, according to which well-being depends exclusively on preferences or desires. The aim is to spell out this dependency in more detail and discuss the pros and cons of these views, seen as substantive theories of well-being. In particular, it is argued that the standard formulations of preference-based views are defective, mainly because they do not pay due attention to the distinction between comparative and monadic attitudes and values. Further, it is argued that in order to find out how well these views can answer the usual complaints levelled against them, it is crucial to distinguish between object preferentialism and satisfaction preferentialism.

Keywords: well-being, good for, better for, preferentialism, preferences, monadic attitudes

11.1 Introduction

YOUR welfare or well-being concerns what is good *for* you, what makes your life worth living. It depends crucially on facts about you and your life. Exactly which facts your well-being depends on is a controversial question. Preference-based views of well-being, or, as I shall call them, forms of *well-being preferentialism*, claim that a person's well-being depends exclusively on his or her desires and preferences. One of the main aims of this chapter is to spell out this dependency in more detail, but I shall also discuss the most important pros and cons of well-being preferentialism.

Here is the overall plan. In section 11.2, I shall explain what it means to say that well-being preferentialism is a substantive view of well-being. In section 11.3, I shall introduce the standard formulation of this theory and list some of its problematic aspects. section 11.4 presents the notions of absolute well-being and monadic attitudes, which are not captured by the standard formulation. section 11.5 clarifies the distinction between two kinds of well-being preferentialism: object preferentialism and satisfaction preferentialism. In section 11.6, I shall examine some standard arguments *for* well-being preferentialism, and

in section 11.7, I shall discuss the standard arguments *against* the theory. Finally, in section 11.8, I shall say something about comparisons of preference strength.

In the end, I shall not take a stand for or against well-being preferentialism. The aim of the chapter is instead to give a detailed characterization of the theory and a list of its most important vices and virtues.

(p. 322) 11.2 Well-Being Preferentialism as a Substantive View about Well-Being

It is important not to conflate well-being preferentialism with other views that assume a strong link between preferences and well-being. First, it should be distinguished from the view that you are always the best judge about what is better for you, for this is a view about the epistemic role of preferences, according to which the best way to find out what is better for you is to consult your preferences. Your preferences provide evidence about what is better for you in the same way as a litmus paper provides evidence of the presence of an acid. This idea is common among those economists who assume that agents are rational and self-interested in the sense that they always prefer what they have good reason to think is better for them.

Second, well-being preferentialism is not the view that we should respect people's preferences. This is a moral principle, most plausibly linked with the idea that we respect people's autonomy by letting them act on their preferences. Whether respecting people's preferences always make them better off is a separate question; you can subscribe to this preference-based view of autonomy and still deny that people's well-being is always promoted by satisfying their preferences. You may think that, in the name of autonomy, we should let people make mistakes and become worse off, at least within certain limits.

Well-being preferentialism is instead a substantive well-being theory that provides us with a list of *what* is good, bad, better, or worse for you, but also, crucially, an explanation of *why* things have these values for you. In this sense, substantive well-being theories are similar to substantive moral theories, such as utilitarianism and Kantianism, since they too do not just list what is right or wrong but also explain what makes actions right or wrong. Roughly put, according to well-being preferentialism, what has value for you is explained by your desires or preferences. This can be contrasted with well-being hedonism, according to which what has value for you is explained by your pleasure or displeasure (Haybron, chapter 12, this Handbook).

11.3 Problems with the Standard Characterization of Well-Being Preferentialism

It is surprisingly difficult to formulate well-being preferentialism in exact and uncontroversial terms. It is common, especially among economists and economics-oriented philosophers, to characterize well-being preferentialism simply in terms of the following two conditions (see, for instance, Broome 1999; Gibbard 1992, 68):

> x is better for you than y iff (if and only if) you prefer x to y, x is equally good for you as y iff you are indifferent between x and y,

(p. 323) where "preference" and "indifference" should be spelled out in terms of choice dispositions (or, possibly, in the case of indifference, the lack thereof), and "x" and "y" pick out either alternative social states, alternative lives of yours, or alternative consumption bundles. This standard characterization has many controversial features, however.

Preference as choice disposition. The attitudes that determine well-being are pretty thin. To prefer x to y is just to be disposed to choose x over y in a choice between the two. To be indifferent between x and y is just to lack a disposition to choose either in a choice between the two (or, alternatively, just to be disposed to do the following: for all z, choose x over z if and only if one chooses y over z, and choose z over x if and only if one chooses z over y). Nothing is said about the reason why you choose one item over another. Nor is anything said about your emotional engagements in these items. But one could claim that the reasons why you choose these items and your emotional engagements in them are relevant to your well-being. For example, one could wonder whether a brute disposition to sit down and twiddle your thumbs for no reasons at all and without any emotional engagement should be relevant to your well-being.

No explanation. The characterization seems incomplete since it does not say that it is *because* you prefer something that it is better for you; it only says that what is better for you coincides with what you prefer. But, as pointed out above, we want a substantive well-being theory to do more than just provide us with a list of what is better or worse for you; we want it to explain what makes things better or worse for you. Indeed, as these conditions stand, they could even be accepted by a well-being hedonist who assumes (mistakenly) that all persons prefer what they find more pleasant or less unpleasant, and are indifferent between what they find equally pleasant or unpleasant. In order to distinguish preferentialism from hedonism we need to say *why* something is better for people.

Object preferentialism. The standard characterization assumes that what is better for you is what you prefer, the *object* of your preference. But a viable alternative is that it is instead the *satisfaction* of your preferences that is better for you. The distinction could be brought out (very roughly) by the following example. Suppose that you prefer drinking champagne to drinking water. One option is to say that it is *drinking champagne* that is better for you than *drinking water*, a view I shall call *object preferentialism*. Another option is to say that it is the compound state of affairs *your preferring champagne and drinking champagne* that is better for you than *your preferring drinking water and drinking water*, a

view I shall call *satisfaction preferentialism* (Rabinowicz and Österberg 1996; Bykvist 1998). This may seem a subtle distinction indeed, but it has great significance. First, as I will show in section 11.6, some arguments for well-being preferentialism are only arguments for the object version. Second, as I will show in section 11.7, many standard objections to well-being preferentialism are only objections to one version, not to both. Finally, it is important to note that, on object preferentialism, what is better or worse for you need not be subjective in nature. For example, having freedom, being physically healthy, and being the first to reach the mountain peak are all items that are better for you, if you prefer these things, but none (p. 324) of these items are subjective in the sense that they themselves necessarily involve your attitudes or pleasures. For example, you can be free without taking pleasure in it or preferring it. Satisfaction preferentialism, on the other hand, will have to say that all things that are better or worse for you must involve something subjective, namely, your preferences. In this sense, satisfaction preferentialism is similar to hedonism, according to which what is better or worse for you necessarily involves pleasant or unpleasant experiences. So, on one construal of the notorious distinction between subjective and objective well-being theories, satisfaction preferentialism is necessarily subjective, whereas object preferentialism is not (it all depends on what people prefer). Of course, when it comes to *what makes* something better or worse for you, both object preferentialism and satisfaction preferentialism are subjective, for they both claim that what explains why something is better or worse for you will necessarily involve something subjective, namely, your preferences. This goes to show that one has to be very careful when using the labels “subjective” and “objective.” The distinction between subjective and objective well-being theories can be drawn either at the level of *what* is better or worse for you or at the level of *what makes* something better or worse for you. (For more on objective theories, see Hurka, chapter 13, this Handbook.)

Restricted comparisons. The standard characterization precludes one whole life from being better for you than another, if you have no preferences over whole lives (perhaps it is too difficult for you to grasp whole lives in all their complexities and details). But it seems possible that one life can be better for you despite the fact that you have preferences only for smaller parts of that life. For example, one life can better for you than another because each moment in the first life is preferred by you to each moment in the second. Conversely, if “x” and “y” in the standard conditions only range over whole social states, whole lives, and complete consumption bundles, then one can wonder how we are to assign well-being to *parts* of these items, such as days or moments in your life or aspects of a social state or a bundle. But we clearly do talk about one day or moment being better for you than another, or one aspect of a state or a bundle being better for you than another.

Only comparative well-being. The standard characterization only states what is better, worse, or equally good for you. As it stands, it does not say anything about what is *good, bad*, or *neutral* for you, but one would have thought that a complete well-being theory should give us an account of both comparative and absolute well-being. After all, we often talk about what is good, bad, or neutral for people, not just what is better or worse for them, and we seem to attach special significance to absolute well-being. For example, whether I should take sympathetic displeasure or pleasure in your life depends on whether

it is bad for you or good for you. If your life is bad for you, I should take displeasure in it, and the worse your life is, the more displeasure I should take in it. If your life is good for you, I should take pleasure in it, and the better your life is, the more pleasure I should take in it. Furthermore, whether we have reason to create or refrain from creating people will depend on whether their lives would be absolutely good or bad for them (Blackorby, Bossert, and Donaldson 2005, 21–26). For example, the fact that your future child's life would be bad for her provides a reason to refrain (p. 325) from creating her, whereas the fact that her life would be good for her provides a reason to create her (or at least not a reason to refrain from creating her).

Preferences rather than monadic attitudes. A feature related to the previous one is that, on the standard account of well-being preferentialism, only comparative attitudes—attitudes that take two objects, such as a preference for x over y—are relevant, and monadic attitudes—attitudes that take only one object, such as a desire for x—are not explicitly mentioned. But, as many philosophers tend to think, the right formulation of well-being preferentialism should be formulated in terms of your desires. Indeed, this is why they often call it the *desire-based* account of well-being. On this alternative account, one starts with conditions on absolute well-being: x is good for you if and only if you desire x (i.e., have some positive attitude toward x), x is bad for you if and only if you have an aversion toward x (i.e., have some negative attitude toward x). Then one adds the claims that the stronger you desire something, the better it is for you, and that the stronger aversion you have toward something, the worse it is for you.

No distinction between final and instrumental well-being and preferences. As it stands, the standard characterization does not distinguish between what has *final* value for you and what has *instrumental* value for you.[1] This is a very intuitive distinction. Most people will say that informing them about healthy food options is good for them because it promotes their health, and health is good for them. If you ask them why health is good for them, many will say that it is good for them because it leads to pleasure, and this is good for them. If you ask why pleasure is good for them, they will often be at a loss and say that it is *just* good for them. Roughly put, then, something has final value for you if it has value for you in virtue of "what it is in itself". Something has instrumental value for you if it has value for you in virtue of the value for you of its effects. This distinction is crucial if we want to distinguish between different well-being theories, for all plausible such theories will agree, at least to a large extent, on what has instrumental value for you, for example, health, income, and education. What they disagree about is what has final value for you.

In order to make clear what the final values are for a well-being preferentialist, we need to distinguish in an analogous way between final and instrumental *preferences*. Roughly put you have a final preference for something if you prefer it for its own sake and not for the sake of its effects, whereas you have an instrumental preference for something if you prefer it for the sake of its effects. We can now say that, according to the standard account

of well-being preferentialism, x is *finally* better for you than y just in case you *finally* prefer x to y.

(p. 326) These notions of final betterness and preference can either be taken as basic or can be defined in terms of *all things being equal* betterness and preferences. Roughly, x better for you than y, all other things being equal, just in case if two alternatives differ only with respect to whether x or y holds in them, the alternative with x is better than the one with y. Analogously, x is preferred to y, all other things being equal, just in case if two alternatives differ only with respect to whether x or y holds in them, the one with x is preferred to the one with y.[2] I will not take a stand on whether final value and final preferences should be taken as basic or not. In the following, I will suppress the qualifier "final" if nothing important hangs on it.

11.4 Absolute Well-Being and Monadic Attitudes

Generally speaking, an account of absolute well-being (suitable for object preferentialists) would look something like this:

> x is good for you iff you favor (i.e., have a positive attitude toward) x. x is bad for you iff you disfavor (i.e., have a negative attitude toward) x. x is neutral for you iff you are neutral toward x.[3]

The *polarity* or *valence* of attitudes can then be understood in the following way (Hurka 2001, 13–14). Very roughly put, to have a positive attitude (a pro-attitude) toward x is to be positively oriented toward x in your actions, emotions, feelings, or evaluative responses. So if you have a positive attitude toward x, you tend to be motivated to bring it about, be glad and happy when you think it obtains, have pleasant thoughts about it, or see it in a good light. To have a negative attitude (a con-attitude) toward x is then to be negatively oriented toward x in your actions, emotions, feelings, or evaluative responses. You tend to be motivated to avoid it, be sad and unhappy when you think it obtains, have unpleasant thoughts about it, or see it in a bad light.

This is indeed very rough, and there are different ways to spell out the polarity of attitudes in more detail. Since the term "attitude" or "desire" can be stretched to (p. 327) cover a lot of different mental states, including urges, whims, appetites, likings, goals, plans, commitments, projects, emotional responses, and evaluative responses, the exact details of an account of polarity depend crucially on which of these attitudes we have in mind. For instance, the polarity of evaluative responses would arguably give most weight to the evaluative light in which we see things, so that a positive emotional response would be defined as seeing something in a *good* light, a negative one as seeing something in a *bad*

light.[4] Since my purpose is to discuss a problem that affects the whole family of well-being preferentialist theories, I shall not argue for a particular choice of attitude. I also assume that an attitude can have zero valence and thus be an attitude of neutrality, accompanied by neutrality in actions, emotions, feelings, or evaluative responses. To facilitate the discussion, I shall often use “favor” as a placeholder for a positive attitude, “disfavor” for a negative attitude, and “neutral” for an attitude of neutrality.

One might wonder how monadic attitudes, such as favorings and disfavoring, relate to comparative attitudes, such as preferences. In particular, one might worry that we have to assume that there are no conceptual links between them. This worry is not justified, for we can state the following links (Chisholm 1964, 613–25; Bykvist 2010, 23):

> You favor x iff there is a state of affairs toward which you are indifferent and you prefer x to it. You disfavor x iff there is a state of affairs toward which you are indifferent and you prefer it to x. You are indifferent toward x iff you are indifferent between x and not-x. You are neutral toward x iff you are indifferent between x and some y toward which you are indifferent.[5]

Some clarifications are in order.

First, to make these links plausible we need be careful about exactly what we mean by “favor,” “disfavor,” and “preference.” For example, if by “favor” we mean “seeing something in a good light,” then “preference” cannot just be understood as a choice disposition; rather, it has to be understood as “seeing something in a *better* light.” Furthermore, depending on what kind of comparative attitudes we have in mind and what kind of objects “x” and “y” pick out, preferring x to y and being indifferent between x and y must be understood in different ways. For example, if preferences are seen as involving choice dispositions (rather than just as seeing something in a better light) and x and y are parts of complete alternatives, then it is reasonable to assume that, in the context of defining what has value for you, preferring x to y (or being indifferent between them) is to prefer x to y (or to be indifferent between them), *all other things being equal*. That you prefer feeling pain over feeling no pain when *only* feeling pain will (p. 328) save your life does not show that pain is not bad for you, for, other things being equal, you do not prefer feeling pain over feeling no pain.

Second, the reason why we need to distinguish between being neutral toward something and being indifferent toward something is that you can disfavor something, say, having a headache, and be neutral toward not having a headache. If being neutral toward not having a headache were the same as being indifferent toward not having a headache, this would be impossible, for you are not indifferent toward not having a headache (i.e., you are not indifferent between not having a headache and having one); you prefer not having a headache to having one.

Third, these links also enable us to give a neat explanation of *strength* or *degree* of favoring and disfavoring (for a person, at a time, in a world). We can simply say that the degree of your favoring for x is greater than your degree of your favoring for y iff you favor both but prefer x to y. Similarly, the degree of your disfavoring of x is greater than your degree of your disfavoring of y iff you disfavor both but prefer y to x.

11.5 Object Preferentialism and Satisfaction Preferentialism and Fundamental Well-Being Components

Object preferentialists and satisfaction preferentialists agree that what determines a person's well-being has something to do with the match or correspondence between his or her preferences and the world, but they disagree about how the match between desires and the world determine her well-being.[6]

To make the distinction clear we need to introduce the notion of a *fundamental well-being component*, which captures a state of affairs that is finally good, bad, or neutral for you *in the most fundamental way*, and which therefore determines the final value for you of everything else, including lives, situations, outcomes, and worlds. Hedonists too need to invoke this notion, for they want to say that situations, lives, and outcomes can have final value for you despite the fact that these items are not themselves experiences. Hedonists think that your feeling pleasure (to a certain degree and at a certain time) is a state of affairs that is finally good in the most fundamental way. Situations, lives, and outcomes are finally good for you only in virtue of containing finally good well-being components, that is, instances of your feeling pleasure.

(p. 329) To see what satisfaction preferentialism and object preferentialism identify as fundamental well-being components, consider the following kinds of attitude/world correspondences:

1. 1. You favor (to a certain degree and at a certain time) p; and p obtains.
2. 2. You disfavor (to a certain degree and at a certain time) q; and q obtains.
3. 3. You are neutral (at a certain time) toward r; and r obtains.

Satisfaction preferentialists think that the fundamental well-being components are all and only the compound states of affairs of the forms (a), (b), and (c), where (a)-states are good for you, (b)-states are bad for you, and (c)-states are neutral for you. In contrast, object preferentialists think that the components are all and only the *objects* of these kinds of attitudes: p, q, and r, where p is good for you, q bad for you, and r neutral for you, assuming that you have no other attitudes toward these states of affairs. When more than

one attitude is directed toward the same state of affairs, the value for you of the state of affairs is some function of the degrees of all the attitudes directed toward the state.

Satisfaction preferentialists and object preferentialists do not just differ on what they put on the list of what is good, bad, or neutral for you; they also give different explanations of why things are good, bad, or neutral for you. According to satisfaction preferentialism, (a)-states are good well-being components because they are correspondences between favorings and the world—*satisfactions*, as they are often called; (b)-states are bad well-being components because they are correspondences between disfavorings and the world; (c)-states are a neutral components because they are correspondences between attitudes of neutrality and the world.

In contrast, object preferentialism claims that a good well-being component for you is good for you because it is an object of a favoring of yours, a bad one is bad for you because it is an object of a disfavoring of yours, and a neutral one is neutral for you because it is an object of an attitude of neutrality of yours. Note that object preferentialism need not deny that (a)-, (b)-, and (c)-states are, respectively, good, bad, and neutral for you. But unless these states are themselves objects of your attitudes, they have value for you only in virtue of *containing* well-being components.

Satisfaction preferentialists have a straightforward account of comparative well-being:

> A is a better well-being component for you than B just in case
>
> 1. 1. A is a good well-being component for you and B is a bad one for you,
> 2. 2. A is a good component for you and B is a neutral one for you,
> 3. 3. A is a good component for you and B is a less good one for you (because it involves a weaker favoring), or
> 4. 4. A is a bad component for you and B is a less bad one for you (because it involves a stronger disfavoring).

(p. 330) Object-preferentialism can accept this, if well-being components are seen as objects of favorings, disfavorings, and neutral attitudes. But they need not accept it, for they can insist, with the standard account of well-being preferentialism presented earlier, that x is a better well-being component for you than y just in case and because you *prefer* x to y. Satisfaction preferentialism, on the other hand, must deny that what you prefer and what is better for you are related in this way. As we will see in section 11.7.6, these different ways of understanding comparative well-being have important ramifications.

It is important to note that, on this construal of the two versions of preferentialism, a whole life can have value for you without you having an attitude toward your whole life. It is enough that you have attitudes toward *parts* of the life, for these parts will then have fundamental value for you, either on their own, if we assume the object version, or in conjunction with the attitudes in question, if we assume the satisfaction version. These

fundamental values will then determine the value for you of your whole life, for it is by containing these fundamental values that your life gets to have overall value for you.[7]

This completes the characterizations of well-being preferentialism. For ease of exposition, I will often use “preference” or “desire” for both comparative and monadic attitudes. I will use “satisfied” and “frustrated” in the following ways. To have a *satisfied* preference or desire is just to get what you prefer or desire. To have a *frustrated* preference or desire is just to fail to get what you prefer or desire. Note that the fact that you fail to get what you prefer or desire need not be bad for you. It all depends on what attitude you take toward the absence of what you prefer or desire. It is only if you disfavor the absence that it (or it in combination with your disfavoring) is bad for you. Finally, I will also assume that, in the way suggested above, we can assess whole alternatives (lives, outcomes, or consumption bundles) as well as parts of them.

11.6 Arguments for Well-Being Preferentialism

One may think that well-being preferentialism is more or less obviously true, if one reasons in this way. Your well-being has to do with what is in your self-interest, and what (p. 331) is in your *self* -interest is what *you* take an interest in. But what you take an interest in is just what you prefer. So, your well-being has to do with what you prefer. This argument is all too quick, however. Being in your self-interest and taking an interest are different things. It is in your self-interest to do some exercise, but you may not take an interest in doing any exercise.

A more interesting argument would be to say that well-being preferentialism provides the best explanation of the perspectival or relational nature of well-being, the fact that it is about what is good, bad, better, or worse *for* a person (Sumner 1996). Many well-being theories can capture the fact what is good for you must in some sense be part of your life, for they would insist that it has to be *your* pleasures, *your* achievements, or *your* knowledge. But well-being preferentialism seems to be able to go further and explain why these things are not just good, but good *for* you. They are good *for* you because they are favored *by* you; the perspectival nature of well-being—x is good for you—is thus explained by the perspectival nature of preferences—x is favored by you.

This is an interesting argument, but note that it is only an argument for object preferentialism. According to satisfaction preferentialism, what is good for you is not good for you because you favor it; it is good for you because it consists in a correspondence between the world and your favoring.

It is doubtful that this argument has much force, however. It is one thing to say that the notion of well-being is perspectival or relational in the sense that it about what is good *for* people. It is quite another thing to say that what explains why something is good for

people must itself be perspectival or relational in a sense that goes beyond the uncontroversial idea that what is good or better for you is some part of *your* life.

Another argument for well-being preferentialism is that it satisfies the plausible-sounding principle that says that things you do not care about cannot be good for you (be good well-being components). Since to favor something is to care about it, preferentialism can agree that if you do not care about, it cannot be good for you.

Again, note that, as stated, this is only an argument for *object* preferentialism, since satisfaction preferentialism implies that the fact that you get what you favor can be good for you even though you do not have any preference for this fact. It is also doubtful how convincing this argument is. Is it really true that what you do not *in fact* care about cannot be good for you? Isn't it more plausible to say that what you *could* not care about cannot be good for you? After all, things that would leave you cold, no matter what, seem odd candidates for good well-being components. But this new principle can be satisfied by satisfaction preferentialism, since it is not impossible to care about the fact you get what you favor. Indeed, not even objective well-being theories are excluded, since we could care about objective features of our lives, such as various perfections of our human nature (Hurka, chapter 13, this Handbook).

Even if it is not always true that what is good for you must be something you actually care about, it seems plausible to say that there must be some link between your lifetime well-being and what actually matters to you, what you actually care about, and what you are actually concerned with. A theory that claimed that it is only the perfection of your human nature that matters for your lifetime well-being seems to be forced to say (p. 332) that a life full of perfections can still be good for you even if you care very little about these perfections (you care just enough for the achievements). But how can a life that leaves you lukewarm be good for you?

Now hedonism may seem to be a serious contender here, since we obviously do care about our pleasures and pains, but examples with pleasures based on false beliefs, such as the *experience machine* (Haybron, chapter 12, this Handbook; Nozick 1974, 42–44), suggest that what we care about and how things feel to us can come apart. Here well-being preferentialism seems to have a clear advantage over hedonism. Well-being preferentialism can claim that there are things that are bad for you even if these things are never believed to obtain. A life with false pleasures will be bad for you, since you disfavor so many aspects of this life even though you do not believe that they are aspects of your actual life. For instance, it is bad for you to be deceived and slandered behind your back, since you strongly disfavor to be treated in these ways. And it is bad for you no matter whether you are aware that you are slandered.

Note that both object preferentialists and satisfaction preferentialists can subscribe to this verdict. While the object preferentialist will say that it is the deception and slandering of you that is bad for you, the satisfaction preferentialist would say that what is bad for you is

the compound states of affairs consisting of the slandering and your disfavoring of it. This argument for well-being preferentialism is thus an argument for both versions of well-being preferentialism.

Moreover, by contrast with hedonism, well-being preferentialism can take into account our preferences about how our pleasures and pains should be traded off against each other. For example, preferences seem to play a crucial role in deciding whether a painful experience followed by a pleasurable experience is on the whole something good for us. Suppose that the intensity of the excruciating pain from running a marathon was greater than the intensity of the pleasure of winning the race. Would that show that this was overall bad for you, if you nevertheless thought the pain was worth it and thus took a positive attitude toward this sequence of events? Hedonists seem forced to say yes, but well-being preferentialists can say no.

Another related virtue of well-being preferentialism is that it can take into account the preferences we have for the *overall shape of our lives*. We often do care about the order in which things happen and how events unfold in time. Some people prefer to have great variation in activities over time, whereas others want stability and continuity. Some prefer their lives to have a narrative structure with challenges and resolutions, whereas others do not care much about whether their lives resemble good stories. Both hedonism and more objective theories seem unable to give any importance to these “shape” preferences. What matters for hedonism is that the balance of pleasure over pain is positive, and, for more objective theories, that the life has the objectively right shape, which is decided by the theory, not by people’s preferences.

These considerations seem to show that well-being preferentialism is neither too subjective nor too objective. It is of course still true that it is more subjective than objective theories in that *only* preferences matter. Now, as we will see, this exclusive (p. 333) dependency on people’s preferences about their lives is also something that makes for many of the major problems with well-being preferentialism.

11.7 Problems with Well-Being Preferentialism

11.7.1 The Explanation of Value-For Goes in the Wrong Direction

One might object to well-being preferentialism because it seems to get the order of explanation wrong: Things are not good for you because you favor them; they are (or should be) favored because they are good for you (Sen and Williams 1990, 12–13; Scanlon 1993, 25; Griffin 1993, 48–52).

The objection assumes that what is good for you is the *object* of my favoring. So, strictly speaking, it is only an objection against object preferentialism. Satisfaction preferentialism

will avoid the charge of reversing the order of explanation, since it will not say that things are good for you because, or partly because, you favor them. If the compound state of affairs *your favoring p, and p* is good for you, then it is good for you no matter whether you favor this compound state of affairs.[8]

11.7.2 Irrational, Uninformed Preferences

Another problem for well-being preferentialism is that many preferences are based on false beliefs and faulty reasoning and it seems strange to say that satisfying these preferences makes you better off. Here are some popular examples (for similar examples, see Sobel 1994):

1. 1. I have a choice between drinking a gray liquid and an orange one. I prefer to drink the orange one because I think it will be tastier. But, as a matter of fact, it contains deadly poison.
2. 2. I have a choice between two treatments for cancer: A and B. I prefer B in the belief that only this treatment will cure me. But, as a matter of fact, it will not. A would have cured me.
3. 3. I prefer the life of a philosopher to the life of a tennis player. I choose to pursue a career in philosophy but realize after a while that philosophy bores me to death.

(p. 334) Satisfying these preferences does not seem to make me better off, but well-being preferentialists seem to be forced to say that it does. One obvious remedy would be to count only informed and rational preferences, that is, actual preferences that are based on true beliefs and correct reasoning (Brandt 1998, 113; Hare 1981, 101). But one might wonder if this rationality constraint is necessary once we distinguish between *final* and *instrumental* preferences. Now, if you have a final preference for x over y, then you cannot be mistaken about what x is in itself. If you are mistaken about this, then, strictly speaking, you do not prefer x in virtue of what it is in itself; you prefer x in virtue of what you *think* x is in itself. So a well-being preferentialism that only counted final preferences would not need to adopt a rationality constraint. Let us apply this idea to the cases at hand.

Consider (1). Is it true that I have a final preference for drinking the orange liquid? No, my reason for drinking it is that I mistakenly think it will be tastier. My final preference is about drinking something tasty. Since I believe that the orange liquid is tasty, I form an instrumental preference to drink this liquid, but satisfying this preference will not make me better off.

Consider (2). Again, my preference for treatment B is an instrumental preference formed on the basis of a final preference to stay alive and a belief that B will cure me, so, again, satisfying this instrumental preference will not make me better off.

Consider (3). A preference for a certain career path is plausibly seen as an instrumental preference. I have a final preference for a job with certain general features: adventurous,

multifaceted, dynamic, flexible, demanding, and so on. And I mistakenly think that a philosopher's job has all these features to a high degree. Choosing the career as a philosopher will therefore not make me better off.

This shows, I think, that there is no need to impose a rationality constraint to deal with cases such as (1) to (3).

But can we not imagine cases where the satisfaction of our *final* preferences seems irrelevant to our well-being? Consider the example from (Sen 1987, 45–46):

> A person who had a life of misfortune, with very little opportunities, and rather little hope, may be more easily reconciled to deprivations than others reared in more fortunate and affluent circumstances…. The hopeless beggar, the precarious landless labourer, the dominated housewife, the hardened unemployed or the over exhausted coolie may all take pleasures in small mercies.

Do we really want to say that these unfortunate people, who have been forced to take pleasures in small mercies, have good lives because they get exactly what they prefer for its own sake?

To avoid this implication, one could claim that we should only count *ideal* final preferences, the final preferences that we *would* have if we knew all relevant facts and reasoned rationally (Harsanyi 1990; Sobel 1994). If these unfortunate people knew how their final preferences were formed, they would no longer hold them. For example, if the dominated housewife knew that she formed her final preference to please her husband (p. 335) just as a way of coping with the submissive role assigned to her by society, she would no longer hold this preference.

However, it is doubtful whether a constraint of this kind is of much help. First, it is not impossible that the housewife would in fact endorse her preference to please even if she were to know about how it was formed. After all, if she is deeply convinced that she deserves no better life, this conviction need not be abandoned once all empirical facts about her preference to please are out in the open. Of course, one could make the constraint stronger by demanding that ideal preferences are those that one would have if one knew all relevant facts including *evaluative* facts, such as the fact that one deserves a better life.

But this move will not deal with cases where the ideal self knows the evaluative facts but does not care much about them. For example, suppose that the dominated housewife would not care much about the fact that she deserves a better life. Indeed, we can suppose that she has already been told by close friends that she should not put up with her submissive role. Even if she acknowledges that this is true, she can still fail to be moved simply because of fatigue and apathy.

To deal with this, we could qualify ideal preferences further and demand that they are the preferences we would have if we had full knowledge about empirical and evaluative facts and were exclusively interested in what is objectively desirable. But then ideal preferences become idle preferences. A person's good is simply what is objectively desirable.

A more general worry with only counting ideal preferences is that actual preferences are completely ignored. I would certainly have been a very different person if I had known all empirical and evaluative facts concerning my actual preferences. Suppose I prefer to drink cheap wine, play football, and listen to 1950s rock-and-roll. But perhaps I would not have had any of these preferences if I had been fully ideal and known all relevant evaluative facts. Does this show that it is not good for me to indulge in these cheap but innocent activities? (For more on this problem, see Railton 1986, 16; Egonsson 2007, 113–25.)

It should also be noted that the term "ideal preference" is a bit of a misnomer, since it suggests that it picks out a special kind of preference in the same way "rational preference" picks out preferences that are rationally based. But to say that you have an ideal preference is just shorthand for saying that you *would* have a certain preference if you had all relevant information and reasoned carefully. You need not actually have this preference. The move to ideal preferences seem therefore not to be available to satisfaction preferentialists, since they assign value only to the correspondences between the world and desires *in the same world.*

11.7.3 Adaptive Preferences

All well-being preferentialists seem to be committed to what we could call the Stoic Principle: Your well-being can be increased either by making the world conform to your preferences, or by making your preferences conform to the world.

(p. 336) Many critics of well-being preferentialism find this principle problematic. Recall the examples about the unfortunate people who adjust to oppressive circumstances by taking pleasure in small mercies. Furthermore, as Rawls (1990, 181) points out, "a bare person," someone who is willing to adapt to any circumstances in order to satisfy her preferences, does not seem to be able to live a life that is good for her.

To find an acceptable solution, we need to get a bit clearer about what exactly makes adaptive preferences so problematic. It cannot be the mere fact that the preferences are adaptive. Often it seems perfectly fine to say that your well-being is increased when you adapt your preferences to circumstances that cannot be changed. Suppose, for instance, that you desperately want to become a professional opera singer, but you simply do not have the voice for it. Suppose, further, that once you accept your limitations you abandon your opera ambitions and go for the more modest goal of singing in a local amateur choir. Why should it not be good for you to satisfy this more modest ambition?

Much more important is the fact that adaptive preferences often do not seem to be about things that are *worthy of concern*. Satisfying preferences that concern things that are not worthy of concern does not seem to make us (much) better off. For instance, a person whose main aim is to count the blades of grass on public lawns seems to have preferences that are seriously misdirected (Rawls 1971, 432). The strength of this preference does not seem to match the value of the preferred object. I am not saying that there is no value in counting the blades of grass. Perhaps there is some excellence involved, endurance, for instance, so that the achievement merits an entry in the Guinness Book of World Records. But to make grass-counting the main aim is to care too much about something that has only minor value. Similarly, someone who takes great pleasure in small mercies seems to take all too great an interest in something that is not worthy of great concern. Finally, what makes a bare person such an odd figure is not that he is willing to change his preferences, but that he is willing to change his preferences no matter whether his new preferences will be for something more valuable or something less valuable. Replacing one's old aims and convictions with new ones is appropriate when the new aims and convictions are concerned with things of greater value. Likewise, abandoning loyalties and attachments is perfectly acceptable when they concern people who are not worthy of our concern.

If this diagnosis is right, it shows that it is not enough to adopt a rationality constraint. There is no guarantee that rational and informed desires will match up with worthwhile activities, for an informed and rational grass-counter or a bare person is not an impossibility.

We can avoid this problem if we say that what makes a person well off is not simply that he gets what he would favor. It is also important that his favorings are about things that are *worthy of concern*. This is of course to reject well-being preferentialism in its purest form, and instead adopt a hybrid theory, an "endorsement theory," as it is now often called (Dworkin 2002, ch. 6; Darwall 1999; Parfit 1992, 502). But note that, on the hybrid view, preferences are not idle wheels; it is still true that nothing can be good for a person if it is not in fact favored by him (or does not consist in a correspondence between the world and a favoring of his). So, this theory is still radically different from (p. 337) a pluralist theory that would accord value to worthwhile activities even if they were not endorsed.

It should be noted that, to a certain extent, these objective values are already taken into account by pure well-being preferentialism, for, normally, we do not just favor to perform certain actions and be a certain kind of person, we favor to do things *well* and be *good* (Raz 1986, 305–7, 317). For instance, if you favor being a parent, you normally favor being a *good* parent. It is strange to say, "Yes, I favor being a parent, but I do not mind being a very bad one." Similarly, if you favor being a friend, you favor being a *good* friend; if you favor being an athlete, you favor being a *good* athlete, and so on. Now, since your favoring to be a good x can only be satisfied if you really are a good x, it is not enough that you believe that you are a good x, for you may be wrong.

11.7.4 Self-Sacrifice

Another popular objection to well-being preferentialism is to say that it cannot accept the possibility of willingly sacrificing yourself for others (Sumner 1996, 135). To willingly sacrifice yourself is to do something you want to do. But if you do what you want to do, then, according to well-being preferentialism, it must be good for you. So it cannot be a sacrifice.

This objection does not threaten well-being preferentialism as such, for we could consider a version that incorporates a *personal restriction*, according to which *only* final preferences that concern things that are essentially about the person count (Overvold 1980, 117–18).[9] This version will exclude *other-regarding* desires, since they are not about things that essentially involve the person. So, if your final desire is *that Jane, Bob, and Henry live* and you satisfy this desire by throwing yourself on the hand grenade that is threatening to kill them, you are not thereby made better off.[10]

However, if we change the example so that your final desire is that *you save your friends*, this move will no longer work. For then it is not ruled out that satisfying this desire will be good for you, since it seems to essentially involve you. Does this show that your action is not a self-sacrifice? No, even if you satisfy one of your present desires by sacrificing your life, you still fail to satisfy many *other* desires, namely, all the future desires that would have been satisfied if you had not killed yourself now. Since your lifetime well-being depends on all your desires, past, present, and future, there is no problem imagining that you do what you most want to do *now*, and still the resulting life is not best for you on the whole.

(p. 338) 11.7.5 Other-Regarding Preferences

It is common to object to well-being preferentialism by saying that it is too permissive in what it counts as being part of a person’s well-being. Suppose that for several years I wanted a democratic, nonracist government to be installed in South Africa. Suppose further that I wanted this not because of how I would benefit but because I thought it would be just. This kind of government has now been installed in South Africa, but it does not seem true to say that I am better off just because this preference of mine has been satisfied. In general, the satisfaction of our *disinterested* desires does not seem to make us better off (Sumner 1996, 134).

But, again, I do not think that this threatens well-being preferentialism as such. At most, it shows that an unrestricted preferentialism should be rejected. A restricted version that incorporates a personal restriction will exclude this preference, since it is not about a state of affairs that essentially involves the preferrer. However, if we change the example and assume that, for decades, I have been working for a political change in South Africa, motivated by a final desire to play an active part in the bringing about of this change, then it no longer seems strange to say that the satisfaction of this desire makes me better off.

To bring about this political change is to realize one of my most important life projects, and thus to make my life successful in an area that concerns me the most.

What is important to note here is that aims and ambitions are desires that concern the way something is brought about. To aim at something is to desire to do something or take part in the bringing about of some state of affairs. My aim cannot be identified with a desire that something is brought about, for example, that a book is written. My aim must be identified with a desire that something is brought about *by me*, for example, that a book is written by me (Fischer 1993, 17). This agent-relative feature is present in the real example of Bertrand Russell, since he did not just want it to be the case that nuclear weapons are disarmed, he wanted *to work for* disarmament. This feature also explains why other people cannot lead our lives and realize our aim and goals. Since the attainment of my goals requires that I act, it is conceptually impossible for other people to realize my aims without my assistance.

11.7.6 Changing Preferences

11.7.6.1 Change across Times

It is a common fact that what we prefer need not occur at the same time as the preference itself. I can prefer *now* that I meet my friends *tomorrow*, for example. This diachronic nature of our preferences can create problems when the preferences change from one time to another, as the following two examples show (Hare 1981, 101–6; Bykvist 2003; Velleman 1993):

1. (p. 339) 1. You now endorse your current life of adventurous sports, but you will later look back upon it with very strong regret. “I completely wasted my youth,” you will come to think.
2. 2. When you were young, what you most favored was to have a future career as poet. Now when you are older you have lost this preference. Indeed, you are now completely indifferent toward poetry. You favor being a teacher instead.

These cases present a challenge for well-being preferentialism. In case (1), is your current life less good for you because of your future disfavoring of your current life? In case (2), is your future less good for you because it fails to contain something you strongly favored in the past? To answer yes to both questions seems counterintuitive, for how can my well-being during some period be affected by things outside this period?

The challenge is especially difficult for object preferentialism, since only this version of well-being preferentialism will have to say that the *object* of preferences has value. It is true that satisfaction preferentialists will say that in case (1) it is bad for you that your current life will be disfavored, and that in case (2) it is good for you that your current life was favored. But it is open to them to deny that these values should be located at the time of your current life, the object of your future or past attitudes. They could instead say that these well-being components occur at the whole interval stretching from the time of the

attitude to the time of the object of the attitude, or they could say that these components lack a determinate temporal location.

One could argue that the problem of changing preferences is only a pseudoproblem: we will not get any cross-temporal conflicts of preferences if we only consider *unconditional* preferences. Go back to example (2) about the young poet. Is it reasonable to assume that when you were young you had an unconditional preference for becoming a poet when you grew up, that is, a preference for becoming a poet that was not conditional on anything? Isn't it more reasonable to assume that you wanted to become a poet only *given* that this is what you would later want? In other words, your preference to become a poet is more reasonably understood as *conditional on its own persistence* (Parfit 1992, 151). If so, then there will not be a conflict of unconditional preferences because you, when you are older, no longer want to become a poet, and so the circumstance on which your past preference was conditional does not obtain.

This move might work if we focus on example (2). But it will not work if we consider example (1), since here it is more plausible to assume that your regret later in life is unconditional. You regret the fact that you indulged in adventurous sports, and this regret is not conditional on anything, especially not on your earlier attitudes. To take another example, my desire now to be an honest and healthy person in the future is not conditional on my desiring it then. I want now that I am honest and healthy even in the future scenario in which I have become dishonest and lazy.

One could argue that the problem is evaded if we claim that only *rational* or *ideal* preferences should count, and add that we will not have any intertemporal conflicts between rational preferences, since rational preferences cannot change over time: If you rationally prefer p at one time, you will always rationally prefer p. When we have (p. 340) a conflict of preferences, as in the examples above, at least one preference must be irrational and should thus be disregarded.

Is it true that rational preferences cannot change over time? Well, it all depends on how we define "rational." It is false on Harsanyi's famous account, according to which rational preferences are those preferences we would have if we were fully rational and were informed. He claims that the condition of full rationality is one in which the preferrer has all the relevant factual information and reasons with the greatest possible care (Harsanyi 1990, 55). This account does not rule out an intertemporal change in what we rationally want. For if your psychological makeup differs radically over time, there is no guarantee that satisfying Harsanyi's condition would lead you to hold the very same preferences over time.

The problem of cross-time intervention arises because of two facts: (i) the preferences you have at a certain time can be about what happens at other times, and (ii) preferences change across time: you can now prefer that x happens tomorrow rather than y, but when tomorrow comes the preference is reversed. One radical solution would be to focus on the

first fact and ban all *diachronic* preferences; only *synchronic* preferences should count. Something important is missing in this picture, however.

First, synchronism implies that no past preferences for the future will have a say. But we do want to count some past preferences. For example, suppose that I want to publish a book I have been working on for 30 years but that I happen to die just before the manuscript is sent to the press. If you care about me and my overall well-being, it seems that you have reason to send the manuscript to the press even when I am dead.

Second, synchronism gives no weight to our preferences for how things should *unfold in time*. But, as suggested in section 11.6, we do not just want things to happen at specific times; we also want them to happen in a certain order. For instance, I want to work hard before I receive some gratitude, I prefer an intimate relationship that starts poorly and ends well to one that starts well and ends poorly, and I want to have a period of moderate pleasure (or perhaps even some moderate pain) before I go through some heights of ecstasy. Reversing the order of events may make them less attractive. The synchronist's *snapshot view*, that is, the exclusive focus on what happens at particular moments of time, prevents her from taking into account preferences about temporal wholes. But clearly this would be to exclude too much, since almost all of our projects, commitments, and plans concern getting things done at the right time and in the right order.

A more inclusive approach to the problem of cross-time intervention would be to say that overall well-being is determined by both synchronic and diachronic preferences, but we should give less weight to diachronic preferences when they do not coincide with synchronic ones (Bykvist 2003). Velleman (1993, 352) suggests a possible justification for this:

> An essential and significant feature of persons is that they are creatures that naturally live their lives from the successive view-points of individual moments, as well as from a comprehensive diachronic point of view.

(p. 341) By counting all synchronic preferences we take seriously the fact that we live our lives from the successive viewpoints of individual moments. However, we should give less weight to diachronic preferences when they conflict with synchronic ones, since from the diachronic perspective, it is important our life shows unity and continuity of purpose, and thus that preferences at one time link up to preferences at other times. Of course, it may be difficult to claim that this is a pure form of well-being preferentialism, since on this view unity and continuity of purpose seem to have some independent positive impact on overall well-being. It is also unclear exactly how much more weight we should give to synchronic preferences.

11.7.6.2 Change across Worlds

Not only can preferences change across times, they can also change across *worlds*, as the following two examples show (Bricker 1980; Gibbard 1992; Bykvist 2010):

1. 1. Suppose that if you get married, you will prefer being unmarried to being married, and if you stay unmarried, you will prefer being married to being unmarried.
2. 2. Suppose that you are a philosopher who has been offered a job: a teaching position in Oxford. You must now choose between moving to Oxford and moving to Sweden, where you will become a professional folk fiddler. Moreover, suppose that if you choose to take up the position in Oxford, you will come to prefer this life to being a fiddler in Sweden—playing intricate polska tunes on the fiddle would not be for you! If you choose to live in Sweden, however, then you will come to prefer living in Sweden as a fiddler to living in Oxford as an academic philosopher.

Which life is better for you in these examples? It is pretty clear how to go about answering this question if we assume satisfaction preferentialism. Look into each life and see which parts of the life you favor and which you disfavor and then for each life weigh the strengths of all the favorings against the strengths of all the disfavorings. Suppose, for instance, that you would disfavor every aspect of your life, no matter whether you got married or not, but that you would disfavor every aspect of your life *less* if you were to get married. Then we can say that getting married is better for you than not getting married, since it is the least bad option for you.[11] Similarly, assume that you would favor every aspect of your life, no matter whether you went to Oxford or not, but that you would favor every aspect of your life *more* if you stayed in Sweden and became a professional folk fiddler. Then staying in Sweden is better for you than going to Oxford, since it is the "most good" option for you.

(p. 342) In contrast, object preferentialism has a much harder time answering this question, for, according to the standard version, whether one option is better for you than another depends on whether or not you prefer it, but the problem is that in the examples above whether or not you prefer your life depends on which life you choose to live. In the first example, whatever life you choose, you will prefer the alternative life to the chosen life. So you will end up with a life that is worse for you than the alternative, no matter which life you choose. In the second example, whatever life you choose you will prefer the chosen life to the alternative life. So you will end up with a life that is better for you than the alternative, no matter which life you choose. We thus get the perplexing conclusion that whether a life is better for you than another depends on which life you end up living (Harsanyi 1953; Bykvist 2010).

The problem is even worse, for an inconsistency threatens object preferentialism, if we assume the following plausible principles (Bykvist 2010):

1. 1. It would be bad for you to lead a life if you would disfavor every aspect of it.
2. 2. It would be good for you to lead a life if you would favor every aspect of it.

3. 3. A life that would be bad for you is worse for you than a life that would be good for you, and this holds no matter whether the good or the bad life were to obtain.

Here is how the inconsistency is generated. Suppose that if you were to get married, you would disfavor every aspect of your life but that you would disfavor it less than not getting married and thus prefer getting married to not getting married. If you did not get married, you would instead favor every aspect of your life but favor it less than getting married and thus prefer getting married to not getting married. Object preferentialism would still have to say that getting married would be better for you, since you would prefer getting married to not getting married, no matter whether you got married or not. But (a), (b), and (c) together entail that it would be worse for you to get married, no matter whether you got married or not. Something will have to give.

Note that satisfaction preferentialism is not threatened by inconsistency. Getting married would be bad for you since you would disfavor every aspect of this option, and not getting married would be good for you since you would favor every aspect of this alternative option. These absolute values then determine the comparative values, so that the good life would be better for you than the bad one, no matter which life you were to lead.

The easiest way out for object preferentialism would be to (i) assume that how good a life is for you depends only on how you would feel about your life, if you were to lead it (so an actual favoring or disfavoring of a merely hypothetical life cannot affect the value of this hypothetical life), and (ii) give up the condition that x is better for you than y iff you prefer x to y. But note that this is to accept that x can be better for you than y even though you would prefer y to x, if x obtained. But this may be a price worth paying.

A more evasive response to this argument is to say that if we only consider fully rational or ideal desires and preferences, the desires and preferences we would have in an epistemically ideal situation, these cases will never occur. The reason why the preferences are contingent in the examples above is that at least one of the person's contingent selves lacks some crucial information about the alternatives.

(p. 343) This response assumes not only that well-being preferentialism should count only ideal preferences, which, as we have seen, is itself a controversial assumption, but also that these ideal preferences will be insensitive to our actual character traits and personalities. Recall that the preferences we are thinking of may concern life options that, if realized, would have drastic effects on the personality, character traits, and belief system of the person. In order to defend this claim it has to be shown that the specification of the ideal epistemic situation will somehow guarantee that the resulting ideal preferences do not vary with even the most drastic change in the personality and the belief system of the person. This is a tall order, and there are plenty of reasons to be skeptical about this. It will not do to say that an ideal epistemic situation is one in which the person has all the relevant factual information and makes no mistakes in reasoning.

Obviously, what a person would desire in a situation like this depends crucially on his actual psychological makeup.

But couldn't the friend of ideal desires respond that if each contingent self was fully informed not just about the objects of their attitudes but also about what would happen to his attitudes if these objects were realized, these selves would no longer disagree in their ideal desires? For instance, if the bachelor knew that he would not favor being married if he were married, then the bachelor would no longer favor being married. He might think: "What is the point in being married if I won't favor it?"

This response will work for some of the cases. It will work for those cases in which the bachelor's attitude is conditional on its own persistence: he favors being married only on the condition that were he to be married, he would still favor it. I guess this is how many people view marriage today. But, of course, one's attitudes toward marriage might be based on *personal ideals*, and it is a characteristic (if not defining) feature of ideals that they are not conditional on their own persistence. I might favor being married because my religious or perfectionist ideals tell me that matrimony is sacred, and therefore has a value that does not depend on whether I would favor being married.

This response has therefore only limited success: it will only take care of cases in which the attitudes are conditional on their own persistence. But we still have cases in which the attitudes are expressive of personal ideals, and there is no guarantee that these attitudes must converge, even if they were properly idealized.

11.8 Comparisons of Preference Strength

The discussions about changing preferences assume that we can meaningfully compare strengths of your attitudes *across worlds and times*, across your "different possible selves," as we may call them. But this is a controversial assumption. Indeed, this problem is analogous to the notorious problem of interpersonal comparisons of well-being. This is not the place to delve deeply into this problem. Instead, I shall focus on some points that are especially relevant to well-being preferentialism.

(p. 344) As I pointed out in section 11.4, comparing the strengths of the absolute attitudes you have at a certain time in a certain world is pretty straightforward if there are certain conceptual links between absolute attitudes and comparative ones. What is not clear is how to compare favoring and disfavoring strength *across* your different possible selves.

If favorings can be linked to preferences along the lines presented in section 11.4, then a comparison of strength of favorings is equivalent to a comparison of strengths of preferences. To decide whether the degree to which one of your selves favors x is greater than the degree to which another of your selves favors y, we should compare the first

self’s preference for x over something she is indifferent toward to the second self’s preference for y over something she is indifferent toward. If the first self’s preference is stronger, then her degree of favoring is greater. Comparisons of favorings can then be grounded in comparisons of preferences. Exactly the same reasoning can be applied to comparisons of degrees of disfavorings.

But one may complain here that we have just pushed the problem one step further, for what does it mean to compare strength of preferences across different possible selves? The answer crucially depends on what we mean by preference. If preference is seen as some kind of emotional engagement that comes with a characteristic qualitative feel to it, such as “feeling stronger attraction to” or “having more pleasant thoughts about,” the comparisons of preferences boils down to comparisons of qualitative feels, which is a version of the old question of how we gain knowledge about the inner feelings of other selves. It is generally assumed that there is an answer to this question. Few people would deny that, typically, one person’s pain of having all her teeth pulled out is greater than another person’s pain from a pinprick.

If, on the other hand, preference is seen as some kind of choice disposition, the problem is more difficult, for how do we compare choice dispositions across different selves? To see the problem here, let us start with comparisons of preference strengths for one self: you, here and now. It seems meaningful to compare your different preferences over alternative lives, for it seems sensible to say that your preference for life A over life B is stronger than your preference for life C over life D, if you would prefer (i.e., be disposed to choose) a lottery that gives you a fifty-fifty chance of living A or living D to a lottery that gives you a fifty-fifty chance of living B or living C.

If you choose the former lottery, it seems that we can conclude that, according to you, the 50% chance of living A rather than B outweighs the 50% chance of living D rather than C. (For more on this, see Mongin and Pivato, chapter 24, this Handbook.)

But what does it mean to say that one self’s preference for life A over life B is stronger than *another* self’s preference for life C over life D?

One suggestion that has been discussed in relation to interpersonal comparisons of well-being is to invoke the notion of an *extended* preference (Adler, chapter 17, this Handbook). Not only can you have preferences for leading a certain kind of life; you can also have preferences for extended alternatives: leading a life, with someone’s objective features (her career, health, and income) and subjective features (her tastes, values, and preferences). So, to form an extended preference concerning someone, you need to imagine not just being in her shoes, but being in her shoes with her feet!

(p. 345) The notion of extended preferences can be applied to preference comparisons across possible selves of the same person, for your actual self can have extended preferences concerning various hypothetical selves. We could then say that your actual

self's preference for life A over life B is stronger than your hypothetical self's preference for life C over life D, if both your actual self and your hypothetical selves would prefer (i.e., be disposed to choose)

1. 1. a lottery that gives you a fifty-fifty chance of living A, with your actual self's features, or living D, with your hypothetical self's features to
2. 2. a lottery that gives you a fifty-fifty chance of living B, with your actual self's features or living C, with your hypothetical self's features.

One obvious problem with this proposal is that it will not give a verdict if your actual and hypothetical selves *disagree* in their preferences over these lotteries. It seems likely that this will happen. After all, if we can disagree about which life to lead, we can also disagree about lotteries concerning extended options. For example, if your actual self is a committed violinist, while your hypothetical self is a committed city banker, your actual self will typically give more weight to extended options involving violin playing and your counterfactual self will typically give more weight to extended options involving city banking (Broome 1995, 55). Recall that, as pointed out in section 11.7.6.1, our preferences for lives can express personal ideals and thus be unconditional.

One possible reply is to concede that there will sometimes be "gaps" in the comparisons of preference strength—sometimes we can neither say that one preference is stronger than the other, nor that it is weaker, nor that it is of the same strength—but hope that the extent of such gaps will be limited. Alternatively, one might claim that this problem suggests that we should invoke a different notion of preference, one that is defined in terms of some kind of qualitative feel. I will not take a stand here, but just point out these two options. In any case, it is important to remember that the problem of comparing attitude strengths is not just a problem for well-being preferentialists; it is a problem for all well-being theories that give *some* weight to how much people (or different selves) care about their lives, which seems to be something all reasonable theories should do.

References

Blackorby, Charles, Walter Bossert, and David Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press.

- Find it@SUB

Brandt, Richard. 1979. *A Theory of the Good and the Right*. Oxford: Oxford University Press.

- Find it@SUB

Bricker, Philip. 1980. "Prudence." *Journal of Philosophy* 77: 381–401.

- Find it@SUB

Broome, John. 1999a. *Ethics Out of Economics*. Oxford: Oxford University Press.

- Find it@SUB

Broome, John. 1999b. *Weighing Goods*. Oxford: Oxford University Press.

- Find it@SUB

Broome, John. 2004. *Weighing Lives*. Oxford: Oxford University Press.

- Find it@SUB

Bykvist, Krister. 1998. “Changing Preferences: A Study in Preferentialism.” Acta Universitatis Uppsaliensis.

- Find it@SUB

(p. 346) Bykvist, Krister. 2003. “The Moral Relevance of Past Preferences.” Heather Dyke, ed., *Time and Ethics: Essays at the Intersection*, 115–36. Boston: Kluwer.

- Find it@SUB

Bykvist, Krister. 2010. “Can Unstable Preferences Provide a Stable Standard of Well-Being?” *Economics and Philosophy* 26: 1–26.

- Find it@SUB

Chisholm, Roderick. 1964. “The Descriptive Element in the Concept of Action.” *Journal of Philosophy* 61: 613–25.

- Find it@SUB

Darwall, Stephen. 1999. “Valuing Activity.” Ellen Paul, Fred Miller, and Jeffrey Paul, eds., *Human Flourishing*, 176–96. Cambridge: Cambridge University Press.

- Find it@SUB

Dworkin, Ronald. 2002. *Sovereign Virtue*. Cambridge, MA: Harvard University Press.

- Find it@SUB

Elster, Jon, and John E. Roemer, eds. 1993. *Interpersonal Comparisons of Well-Being*. Cambridge: Cambridge University Press.

- Find it@SUB

Egonsson, Dan. 2007. *Preference and Information*. Aldershot: Ashgate.

- Find it@SUB

Fischer, Martin, ed. 1993. *The Metaphysics of Death*. Stanford, CA: Stanford University Press.

- Find it@SUB

Gibbard, Allan. 1992. “Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of a Life.” Jon Elster and Anand Hylland, eds., *Foundations of Social Choice Theory*, 165–93. Cambridge: Cambridge University Press.

- Find it@SUB

Griffin, James. 1993. “Against the Taste Model.” Jon Elster and John E. Roemer, eds., *Interpersonal Comparisons of Well-Being*, 45–69. Cambridge: Cambridge University Press.

- Find it@SUB

Hare, Richard M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.

- Find it@SUB

Harsanyi, John. 1953. “Welfare Economics of Variable Tastes.” *Review of Economic Studies* 21: 204–13.

- Find it@SUB

Harsanyi, John. 1990. “Morality and the Theory of Rational Behaviour.” Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond*, 39–62. Cambridge: Cambridge University Press.

- Find it@SUB

Hurka, Thomas. 2001. *Virtue, Vice, and Value*. Oxford: Oxford University Press.

- Find it@SUB

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

- Find it@SUB

Overvold, M. 1980. “Self-Interest and the Concept of Self-Sacrifice.” *Canadian Journal of Philosophy* 10 (1): 105–18.

- Find it@SUB

Parfit, Derek. 1992. *Reasons and Persons*. Oxford: Clarendon Press.

- Find it@SUB

Rabinowicz, Wlodek, and Jan Österberg. 1996. “Value Based on Preferences: On Two Interpretations of Preference Utilitarianism.” *Economics and Philosophy* 12: 1–27.

- Find it@SUB

Railton, Peter. 1986. “Facts and Values.” *Philosophical Topics* 14: 5–31.

- Find it@SUB

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

- Find it@SUB

Rawls, John. 1990. “Social Unity and Primary Goods.” Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond*, 159–85. Cambridge: Cambridge University Press

- Find it@SUB

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

- Find it@SUB

Scanlon, Thomas. 1993. “Moral Basis of Interpersonal comparisons.” Jon Elster and John E. Roemer, eds., *Interpersonal Comparison of Well-Being*, 17–44. Cambridge: Cambridge University Press.

- Find it@SUB

Sen, Amartya. 1987. *On Ethics and Economics*. Oxford: Blackwell.

- Find it@SUB

Sen, Amartya, and Bernard Williams, eds. 1990. *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

- Find it@SUB

Sobel, David. 1994. “Full Information Accounts of Well-Being.” *Ethics* 104: 784–810.

- Find it@SUB

Sumner, Wayne. 1996. *Welfare, Happiness, and Ethics*. Oxford: Oxford University Press.

- Find it@SUB

Velleman, David. 1993. “Well-Being and Time.” Martin Fischer, ed., *The Metaphysics of Death*, 327–57. Stanford, CA: Stanford University Press.

- Find it@SUB

Notes:

([1]) One reason for this omission is that, as pointed out above, the standard account typically assesses only *complete* alternatives (social states, whole lives, consumption bundles). The distinction between final and instrumental value is difficult to draw if one only compares complete alternatives, since it does not make much sense to say that a complete alternative has value for you in virtue of its effects. A complete alternative for you already incorporates *all* relevant effects on your well-being. Of course, your life can have effects on other people's lives, but this does not seem to be relevant to *your* well-being.

([2]) This is rough since strictly speaking it might be impossible for two alternatives to differ *only* in respect of whether x or y holds in them. For example, they will surely differ in terms of what necessarily follows from x and y, respectively. A more accurate definition would have to be phrased in terms of "maximally similar" alternatives. Note also that, as it stands, this account rules out "organic unities" with respect to final value and preference. It rules out that A may be finally better for you than B and still A and C is not better for you than B and C, for all C. Similarly, it rules out that you may finally prefer A to B and still not prefer A and C to B and C, for all C. Whether there are organic unitities of these kinds is a contested issue, of course. But here is one possible example: it is finally better for you to feel more pleasure than less, but not better for you to feel more pleasure when it is taken in worthless activities.

([3]) To avoid needless complications, I here ignore the possibility that some alternatives are neither good, nor bad, nor neutral for you.

([4]) Seeing something in a good light is not the same as having a *belief* that something is good. Things can present themselves in a good light without being judged to be good.

([5]) I assume here that you are indifferent *between* all the things you are indifferent *toward*. If this does not hold, what is favored is not guaranteed to be preferred to what is disfavored.

([6]) It is easy to miss this distinction, if one thinks that object preferentialism says that x is better for you than y, *on condition that you prefer x to y*, for this means that the combination of x *and* your preference for x over y is better for you than the combination of y *and* your preference for x over y, which looks like a form of satisfaction preferentialism.

([7]) When aggregating the value of a whole life, note that it is possible that two well-being components can create a new one when combined. I may finally favor A and finally favor B and not just favor the whole consisiting of A and B just because I favor its parts. The combination of A and B can create *holistic* effects that I finally favor. So here there are three well-being components: A, B, and the whole A-and-B (or, as satisfaction preferentialists would prefer to individuate them: A combined with a favoring of A, B combined with a favoring of B, and A-and-B combined with a favoring of A-and-B). For

example, I may finally favor pleasure to degree 2 and finally favor pleasure to degree 10 but not just favor, derivatively, the combination of first feeling pleasure to degree 2 and then feeling pleasure to degree 10, for I may *finally* favor a combination of pleasures that goes from less intense to more intense pleasures, because of its “upward shape.”

([8]) Satisfaction preferentialism will of course hold that certain things are *proper parts* of what is good for you because you favor them; it is because you favor p that p is part of the good compound state of affairs *your favoring p, and p*. But this does seem to be a minor worry (if a worry at all). Any view that thinks that a life is good partly because you favor parts of it has to say this about the value of lives.

([9]) Note that Overvold treats his constraint as a *sufficient* condition for well-being, whereas I treat it as a necessary condition.

([10]) One might worry that this restriction excludes too much. Perhaps we want to say that you are made better off when your preference for your children being happy is satisfied. It is true that this would be excluded if your preference was understood *de re*: you prefer that *Jane* is happy (assuming that Jane is your daughter). But if you prefer (*de dicto*) that *whoever is your* child is happy, then your preference is not ruled out by the restriction, for this preference is about something that essentially involves you: that there is someone who is *your* child and who is happy.

([11]) Note that from the fact that you would disfavor your married life less than you would disfavor your unmarried life it does not follow that you *actually* prefer your married life to your unmarried life, not even if we assume that disfavorings are defined in terms of preferences, as explained in section 11.4. To say that you would disfavor a life A is to say that, *if you were to lead A*, you would then prefer a life that you would be indifferent toward to A.

Krister Bykvist

Philosophy, Stockholm University

Krister Bykvist is Professor of Practical Philosophy at the Department of Philosophy, Stockholm University.

- Oxford University Press