

CAN UNSTABLE PREFERENCES PROVIDE A STABLE STANDARD OF WELL-BEING?

ABSTRACT:

How do we determine the well-being of a person when her preferences are not stable across worlds? Suppose, for instance, that you are considering getting married, and that you know that if you get married, then you will prefer being unmarried, and that if you stay unmarried you will prefer being married. The general problem is to find a stable standard of well-being when the standard is set by preferences that are not stable. In this paper, I shall show that the problem is even worse: incoherence threatens if we accept both that preferences determine what is better for us and that desires determine what is good for us. After I have introduced a useful toy model and stated the incoherence argument, I will go on to discuss a couple of unsuccessful theories and see what we can learn from their mistakes. One important lesson is that how you would have felt about a life had you never lead it is irrelevant to the question of how good that life is for you. What counts is how you feel about your life when you are actually leading it. Another lesson is that a life can be better for you even if you would not rank it higher, if you were to lead it.

1. The problem of changing desires

How do we determine the well-being of a person when her preferences are not stable across worlds? To give you a feel of this problem, consider the following examples:

The career choice. 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.¹

Here is another example:

The bachelor's dilemma: 'To wed or not to wed'. You are considering getting married. The problem is, however, that you know that if you get married, then you will prefer being unmarried to being married. If you get married, you will adopt certain perfectionist ideas about marriage and think that your marriage does not live up to the standards. However, if you stay unmarried, you will accept less exacting requirements and prefer being married to being unmarried.

Which life is better for you in these cases? To answer this question we need to find a vantage point from which we can judge which life is better. But the problem is exactly how to identify this vantage point, since what is the better life seems to depend on which life is realized. In the first example, whatever life is chosen you will prefer that life to the alternative life, and, in the second example, whatever life is chosen you will prefer the alternative life to the chosen life. In a nutshell, the problem is to find a stable standard of well-being when the standard is set by preferences that are not stable.

It is important to stress that this is not just something that should worry desire-based theorists. This problem will also afflict *endorsement theories* that define a person's good as the right combination of some kind of objective desirability (moral, religious, intellectual, aesthetic, or athletic excellence or worth) and subjective endorsement, and allow preferences to be tie-breakers when the compared objects are equally desirable (or

¹ Similar examples are presented in Bricker (1980), pp. 381-401, and Gibbard (1992).

incommensurable).² Suppose, for instance, that being married and being unmarried are equally worthy of concern. Now, if preferences are seen as *tie-breakers*, then what is better for you is decided by what you prefer. But then we are back to the problem of how to decide which preference should act as tie-breaker.

I shall argue that the problem is even worse: incoherence threatens if we accept both that preferences determine what is better for us and that desires determine what is good for us. I will begin by explaining how preferences and desires are usually seen as determining well-being, and then show how this leads to incoherence if applied to cases with unstable preference and desires. After that, I shall discuss a couple of unsuccessful solutions and see what we can learn from their mistakes. One of the most important lessons is that how you would have felt about a life had you never lead it is irrelevant to the question of how good that life is for you. What counts is how you feel about the life when you are actually leading it. Another lesson is that we should give up the idea that your preferences over two lives determine which life is better for you. Indeed, I will argue that a life can be better for you even if you would not rank it higher, if you were to lead it.

2. Desires and well-being

A desire-based well-being theory is often assumed to be committed to the following principles:

- (1) x is *good* for S iff S wants x .
- (2) x is *better* for S than y iff S prefers x to y .³

² For some recent endorsement theories, see, for instance, Dworkin (2002), ch 6, Darwall (1999), Kraut (1994), and Parfit (1993), p. 502. I should say that it is not clear that they all would accept that preferences can be tie-breakers.

³ This way of stating the desire-based theory assumes that it is the *objects* of wants and preferences that have value. But the desire-based theory could be formulated in an alternative way. Instead of assigning value to the objects of attitudes it could assign value to the state of affairs that an attitude is satisfied. On this account, it is the state of affairs S wants x and x obtains that have value, not the object x . In this paper,

Since any desire-based theory will allow that things that are not desired or preferred by a person can still have *instrumental* value for that person, (1) and (2) must be understood as talking about intrinsic value and intrinsic wants and preferences.⁴ To avoid cluttering the exposition, I will suppress the qualifier ‘intrinsic’ in the following.

Even with this clarification in mind, (1) can’t be exactly right. It implies that a person’s wants determines what is good for her, but this seems false if wanting *x* is simply defined as preferring *x* to its negation. Suppose you want not to have a headache, understood as your preferring not having a headache to having a headache. Then (1) implies that when this want is satisfied something positively good occurs in your life. It also implies that if you create anti-headache wants in order to satisfy them you make your life better, other things being equal. But if you are like me you take a *neutral* attitude towards not having a headache, and a negative attitude towards having a headache. Remember that we are talking about *intrinsic* attitudes here. Obviously, I can take a positive *instrumental* attitude towards not having a headache since not having a headache might cause me to feel the pleasure of relief and enable me to focus on other intrinsically desired activities in my life.

I shall stick to the object-version, but my own theory could easily be reformulated as a satisfaction-version. The distinction between object- and satisfaction-versions is clearly stated in Rabinowicz and Österberg (1996). For more on the differences between these versions, see **** (1998).

⁴ This is still inadequate, if (1) and (2) quantifies over all possible objects. Surely, a world, or an outcome, can be intrinsically good or better for a person without her having a desire or preference for this world, or outcome. She might even be unable to conceptualize such a complex object. To overcome this inadequacy, we have to distinguish between what has intrinsic value in the *most fundamental* way or *basic* way and what has intrinsic value in virtue of containing something that has intrinsic value in a basic way. Whole possible worlds and outcomes normally have only a derived intrinsic value for a person in virtue of the basic intrinsic valuables they contain. (1) and (2) are therefore most plausibly read as criteria for what has basic intrinsic value. Note that this is not a special problem for desire-theories. Hedonists, for instance, face the same problem. They want to say that an outcome can be intrinsically good or better for a person and that only pleasures are intrinsically good. But since an outcome is not a pleasure, they have to be understood as saying that an outcome can be intrinsically good in virtue of containing pleasures that have basic intrinsic value. For more on the notion of basic value, see Feldman (2005), pp. 379-400.

Therefore, it seems more sensible to say that it is good for you to get what you *favour*, i.e., what you have a positive attitude towards:

(1*) x is good for S iff S favours x .

It is of course incumbent on me now to say something more about the *polarity* or *valence* of attitudes. Very roughly put, to have a positive attitude (a pro-attitude) towards x is to be positively oriented towards x in your actions, emotions, feelings or evaluative responses. So, if you have a positive attitude towards 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) towards x is then to be negatively oriented towards 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. I also assume that an attitude can have zero valence and thus be an attitude of indifference, accompanied by indifference in actions, emotions, feelings, or evaluative responses.⁵

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 cover a lot of different mental states, including urges, whims, appetites, likings, goals, plans, commitments, projects, 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 evaluative response would be defined as seeing something in a *good* light, a negative one as seeing something in a *bad* light, and a neutral one as seeing something in a *neutral* light.⁶ Since my purpose is to discuss a problem that affects the whole family of desire-regarding theories, including endorsement theories, I shall not argue for a particular choice of attitude.

⁵ For a similar account of the polarity of attitudes, see Hurka (2001), pp. 13-14.

⁶ Seeing something in a good light need not be the same as having a *belief* that something is good. Things can present themselves in a good light without being judged to be good.

In the following, I shall use ‘favour’ as a place-holder for a positive attitude, ‘disfavour’ for a negative attitude, and ‘indifference’ for an attitude of indifference. ‘attitude’ will be used to refer to any kind of attitude, including comparative ones, i.e., preferences.

3. Toy model

To avoid dealing with too many difficulties at once, I will work with a highly idealized model. I shall assume that the possible attitudes a person has towards her possible lives can be represented by a grid of the following kind.

		Lives			
		w1	w2	w3	..
Attitudes	w1	$u_{w1,w1}$	$u_{w1,w2}$	$u_{w1,w3}$..
	w2	$u_{w2,w1}$	$u_{w2,w2}$	$u_{w2,w3}$..
	w3	$u_{w3,w1}$	$u_{w3,w2}$	$u_{w3,w3}$..
	:

If you look into a horizontal world row, you’ll find a distribution of numbers that represent the attitudes the person has, in a certain world, towards her various possible lives. For instance, if you look into the w1-row, you’ll find a representation of the attitudes the person has *in w1* towards her life in w1, her life in w2, her life in w3 and so on. A vertical column gives you a representation of all her possible attitudes towards the life in a certain world. So, for instance, if you look into the w1-column, you’ll find a representation of all possible attitudes towards her life in w1.

Positive numbers represent favourings, negative numbers disfavourings, and zero neutral attitudes. A preference, in w_i , for the life in w_j over the life in w_k is represented

by a greater number in w_i, w_j than in w_i, w_k , ($u_{w_i, w_j} > u_{w_i, w_k}$). Indifference, in w_i , between w_j and w_k is represented by assigning the same number to both w_i, w_j and w_i, w_k , ($u_{w_i, w_j} = u_{w_i, w_k}$).

In this model, a case where the comparative preferences concerning two worlds, w_k and w_l , stay fixed across two worlds, w_i and w_j , will be represented by a grid in which $u_{w_i, w_k} > u_{w_i, w_l}$ and $u_{w_j, w_k} > u_{w_j, w_l}$, or $u_{w_i, w_k} < u_{w_i, w_l}$ and $u_{w_j, w_k} < u_{w_j, w_l}$. Here is a simple case:

Case 1

	w1	w2	..
w1	10	5	
w2	20	10	
:			

This grid tells us that, no matter whether w_1 or w_2 is realized, I will prefer my life in w_1 to my life in w_2 . It also tells us that, no matter whether w_1 or w_2 is realized, I will favour my life in w_1 as well as my life in w_2 .

Preference reversal cases will be represented by grids where this kind of invariance does not hold. An example of preference reversal would be:

Case 2

	w1	w2	..
w1	0	-2	
w2	6	8	
:			

This grid tells us that, in world w_1 , I am neutral towards my life in w_1 , disfavours my life in w_2 , and thus prefers my life in w_1 to my life in w_2 . It also tells us that, in w_2 , I favour both my life in w_1 and my life in w_2 but prefer my life in w_2 to my life in w_1 . This is

thus a possible representation of the career choice case in which it holds that, whatever life is chosen, you will prefer the chosen life.

A case of the bachelors' dilemma type would be the following.

Case 3

	w1	w2	..
w1	-2	0	
w2	8	6	
:			

This tells us that in w1 I disfavour my life in w1, see the alternative life in w2 in a neutral light, and thus prefer my life in w2 to my life in w1. It also tells us that, in w2, I favour both my life in w1 and my life in w2 but prefer the former to the latter. So, no matter which is realized, I will prefer the alternative life.

It is not assumed at this stage that we can compare favourings and disfavourings across worlds and say that one possible self favours (disfavors) her life more than another possible self favours (disfavors) her life. Nor is it assumed that that we can compare preference intensities across worlds and say that one possible self's preference is stronger than another possible self's preference. The incoherence argument I will present in the next section does not require any of these controversial measurability assumptions. However, some of the solutions I will discuss later will require stronger assumptions.

Before we move on to the argument for incoherence, I need to clarify some further idealizing assumptions.

(a) When I say that a person has an attitude *in a world* I mean that she has that attitude with the same strength at all times in her life in that world. This will make it possible to sidestep the thorny issue about how to deal with conflicts of attitudes across time.⁷ I shall

⁷ I have addressed this problem elsewhere. See **** (2003).

also assume that the lives we consider have exactly the same duration. This is to avoid deciding on how the duration of a life matters to lifetime well-being.

(b) When I say that a person has an attitude towards *a life* I mean that she has an attitude toward that life as a whole, not just an attitude towards some local aspects of it. This means that I will only evaluate a person's life in terms of her *global* attitudes. Though this restriction is controversial, it enables us to illuminate the desire-theories under discussion in a clear and simple way. It should be noted that this restriction is not wholly implausible. It seems reasonable to give priority to global desires, since they are more comprehensive than local desires about particular states of affairs.⁸ There are two ways in which global desires can be said to be more comprehensive than local ones. Firstly, global desires concern the way particular states of affairs make up bigger wholes, for instance, the way they unfold in time and make up temporal wholes. Secondly, they concern your local desires and their satisfactions and frustrations. Even if many of your local desires are satisfied, you may not be happy about having these desires. For instance, your addictive desires to a certain drug may all be satisfied, but you may strongly desire not to have these addictive desires in the first place.

(c) Since I assume that each cell in the grid has a value I am, in effect, ruling out worlds in which the subject fails to exist or exists but lacks any preferences or desires. I am also ruling out worlds towards which the subject has no attitude. So, I am limiting myself to evaluating the well-being of fully-opiniated preferrers.

(d) Not many desire-theorists accept that any old desire or preference can be relevant for a person's well-being. It is common to count only those that are rational, self-regarding, autonomous and authentic. To accommodate these theories, I shall assume that all desires and preferences in my model are properly 'laundered'. By 'rational preferences' I just mean weak preferences that are transitive and reflexive. For simplicity, I shall also assume that all preferences are connected. I will come back later to the

⁸ Even if the *total* well-being of a life should be seen as a function of the global attitudes towards the whole life and the local attitudes towards parts of it, it is plausible to assume that the *basic* intrinsic value of a whole life is determined only by the global attitudes towards the whole life. For more on the notion of basic value, see footnote 4.

question of whether shifty preferences can be said to be rational in a more demanding sense.

4. An inconsistency

To state the argument for incoherence, we need to formulate the favouring-goodness link and the preference-betterness link in a way that suits our simplified model. The most natural way to formulate the idea that favourings determine goodness would be to say that a life is good for a person just in case she would favour it, were she to lead it. More exactly:

World-Bound Well-Being (WW)

S's life in *w* is good for S iff S favours, in *w*, her life in *w*.

Endorsement theories would not accept this principle as it stands, but they will be inclined to accept it if we restrict the domain of quantification to lives that are objectively desirable or worthy of concern.⁹

It is more difficult to find an appropriate formulation of the idea that preferences determine betterness in the present model since preferences may change across worlds. But if you think that preferences matter in this context, it is tempting to think that they matter when preferences stay fixed and thus accept that if a person would prefer one life to another, no matter which life were realized, then the first life is better for her than the second. More exactly:

⁹ No doubt some endorsement theorists might even find this restricted version of WW unacceptable. It will be rejected by those who think that endorsement is crucial only for the most important parts of a person's well-being and that a person's unendorsed excellence can still have some positive value for her. To accommodate this pluralist endorsement theory, WW has to be qualified so that it talks only about what is '*significantly* good for S'.

Dominance

If S prefers, in both w and w' , her life in w to her life in w' , then her life in w is better for her than her life in w' .

Again, endorsement theories will accept this principle only if the domain is restricted to lives that are equally worthy of concern (or incommensurable).

Too see why these principles generate a contradiction, consider the following case:

Case 4

	w1	w2	..
w1	0	-2	
w2	8	6	
:			

Now, WW entails that

(1) Her life in $w1$ is not good for her (since she does not favour, in $w1$, her life in $w1$).

and

(2) Her life in $w2$ is good for her (since she favours, in $w2$, her life in $w2$).

So

(3) She can't be better off in $w1$ than in $w2$.

So

(4) Her life in $w1$ is not better for her than her life in $w2$.¹⁰

¹⁰ The move from 'not better off' to 'not better for' is uncontroversial in this context. It is more controversial in cases of creation, since some would claim that, even if we cannot make a person better off by creating her, existence can still be better for her than non-existence.

But *Dominance* implies that

(5) Her life in w1 is better for her than her life in w2 (since she prefers, in both w1 and w2, her life in w1 to her life in w2).

So

(6) Her life in w1 is both better and not better for her than her life in w2.

Contradiction!¹¹

5. Idealize!

One obvious response to this argument is to say that the problem will vanish if we only consider fully rational or ideal desires and preferences, the desires and preferences we would have in an epistemically ideal situation. This response assumes not only that the desire-regarding theory should favour ideal desires, which is in itself a controversial assumption, but also that these ideal desires will be insensitive to our actual character traits and personalities. Recall that the desires 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 desires 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 sceptical 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 instrumental reasoning. Obviously, what a person would desire in this sense depends crucially on her actual psychological make-up.

But couldn't the friend of ideal desires respond that if each possible self was fully informed not just about the objects of their attitudes but also about what would happen to

¹¹ A first, but more complicated, version of this argument was given in **** (2006).

his attitudes if these objects were realized, they would no longer disagree in their ideal desires? For instance, if the bachelor knew that he would not favour being married if he were married, then the bachelor would no longer favour being married. He might think: 'What is the point in being married if I won't favour it?'

I think this response will work for some cases. It will work for those cases in which the bachelor's attitude is *conditional on its own persistence*: he favours being married only on the condition that were he to be married, he would still favour it.¹² I guess this is how many people view marriage today. But, of course, one's attitudes towards 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 favour 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 people would favour being married. To take another example which is closer to home, 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.

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.

As a last attempt to save the invariance of ideal attitudes, one could simply define an ideal attitude in a way that guarantees that a person's possible selves would have the same ideal attitudes. An endorsement theorist could, for instance, say that our ideal desires are those we would have if we had full knowledge about the evaluative facts and were exclusively interested in what is objectively desirable. But then ideal desires become an idle wheel. A person's good is simply what is objectively desirable in her life. Since ideal desires are defined as tracking objective desirability, it is trivially true that something is good for a person only if it is endorsed by her ideal desires. Moreover, if this idealization is applied to absolute as well as comparative attitudes, the idealized

¹² This kind of conditionality is discussed in Parfit (1992), p. 151.

preferences can no longer work as tie-breakers. For if two options are equally desirable, then the idealized self will always be indifferent between the options.

6. Relativize!

This is the standard remedy for inconsistencies. If you can derive two propositions that do not square logically, relativize the propositions in some suitable way. One way of doing it is this.

Relativized well-being

Her life in w is good for S , *relative to* w' iff S favours, in w' , her life in w .

Relativized betterness

If S prefers, in w , w' to w'' , then w' is better for S than w'' , *relative to* w .

Using these principles, we can avoid the inconsistency and instead generate the following consistent judgements for Case 4.

Relative to w_1 : w_1 is not good for her, w_2 is not good for her, w_1 is better than w_2 .

Relative to w_2 , w_1 is good for her, w_2 is good for her, w_1 is better than w_2 .

But it is not enough to add some indices in the right places and claim victory. We need to understand what it means to say that relative to a world, w , the life in world w' is good for a person. No matter how the relativization is spelled out, a relativistic theory will not give us the whole story: we want to know what is good for or better for the person *period*, not just what is good or better for the person relative to this or that. Remember that we are trying to find a stable standard of well-being, not a set of different standards. More can be said about this, but I will move on. I hope you agree that relativism should be seen a last resort.

7. Actualize!

Another approach would be to defer to actual preferences.

Actualist well-being

Her life in w is good for S iff S favours, *in the actual world*, her life in w .

Actualist betterness

Her life in w is better for S than her life in w' iff S prefers, *in the actual world*, her life in w to her life in w' .

One problem with this approach is that if 'actual world' is treated as an indexical, we have to give up our search for a stable standard of well-being and accept that whether a life is best might depend on whether or not it is realized. To see this, go back to the career-example. If you were to move to Oxford, then your actual preferences in this scenario would favour your move. Since your actual desires determine the values of outcomes, in this scenario the philosopher's life is better for you than the fiddler's life. On the other hand, if you were to move to Sweden, a different scenario would be realized, and your actual preferences in this scenario would not favour your move to Oxford. So, in this alternative scenario the fiddler's life is better for you. The conclusion is that the philosopher's life is best for you only if you become a philosopher.

This axiological variance is troubling for two reasons. First, it will entail that a life can be better for a person even if it would not be better for her to lead it. Suppose that I prefer a certain counterfactual life to my actual life, but that I would not prefer it if I were to lead it. Then actualism entails that the life is better for me, since it preferred by my actual self, but also that it *would* not be better for me to lead it, since if I were to lead it

my actual self in that alternative scenario would not prefer it. But surely, a life cannot be better for a person if it would not be better for her to lead it.¹³

The other problem with axiological variance is that it short-circuits prudential deliberation.¹⁴ If whether a life is best for you depends on whether it will be realized, then whether it is prudentially right for you to realize it will depend on whether you will realize it. If you become a philosopher, then this life is best for you and thus the right life to choose. On the other hand, if you become a fiddler, the philosopher's life will not be best for you and thus not something that is right for you to choose.¹⁵ This normative variance is troubling since when you use a theory as a guide to action you use the theory in your deliberations about what to do. In particular, you use it to decide which options are right or wrong. On the basis of this deliberation you then make up your mind and decide what to do. But if an action's rightness depends on whether it is performed, then in order to decide whether an action is right you first have to know whether or not you are going to perform it. But there is no point in deliberating about whether to perform an action if either you believe that you will perform it, or you believe that you will not perform it. If you believe that you will perform the action, the issue is settled for you, and there is no point in deliberating about it further. If you believe that you will not perform the action, the action is no longer a serious possibility, i.e. something that is compatible with what you believe (even if it might be something you can do); so again there is no point in deliberating about whether to perform it.¹⁶

¹³ Now it is important to remember that 'better for' is supposed to capture facts about a person's *well-being*, what makes a person better-off. My argument will not work if 'better for S' is read descriptively as 'better according to S' or 'judged to be better by S'.

¹⁴ For a more thorough discussion of this issue, see **** (forthcoming).

¹⁵ This assumes that a right option is one whose outcome *is* at least as good for the person as that of any other options. To avoid shifty prescriptions, the actualist could instead define a right option as one that is *ratifiable* in the sense that if the option were to be realized, the outcome *would* be at least as good as that of any other option. One major problem with this approach is that in many cases there are no ratifiable options (the unmarried person's dilemma is one example), and yet we want to say that there is a right option. I say more about this in **** (2006), pp. 275-76.

¹⁶ A similar argument is spelled out in Carlson (1995), pp. 101-102, and touched upon in Bricker (1980), p. 395. The general idea that the prediction of one's actions crowds out deliberation has widespread support. See, for instance, Goldman (1970), p. 194, and Taylor (1966), p. 174.

The claim here is not that once you have formed the belief (or disbelief) that you are going to do A, then you are no longer able to deliberate about whether to do A. If you give up the belief or the disbelief, you may start deliberating again. The claim is rather that while you are in the grip of the belief or disbelief that you will do A, it is not possible for you to deliberate about whether to do A. Or at least, this is not possible if you are rational. For rational agents, belief or disbelief about what they are going to do excludes wondering about whether to do it.

To avoid axiological invariance, we could adopt a *rigidified* notion of ‘actual world’. The relevant preferences and desires are those that we have here in our concrete world.¹⁷ What is an actual preference in this sense will not vary across worlds, since when we ask whether a counterfactual world matches our actual preferences, ‘actual’ rigidly refers back to our world.

One obvious problem with this view is how to make sense of the well-being of *non-actual* persons. Sometimes we want to compare the possible lives of a non-actual person. For instance, we might want to compare two possible lives for a child we could have conceived but as a matter of fact did not. If only preferences of actual people count, there are no preferences we could use to determine the well-being of the non-actual person. True, we, the actual people, might have preferences concerning the life of a non-actual person, but what determine the well-being of a person must in the end be her own preferences.¹⁸

Another worry is that this rigidified actualism does not provide us with a well-being theory that is sufficiently sensitive to preferences. No matter how drastically different a person’s counterfactual self is in terms of personality and character, it will be the preferences of her actual self that determines the well-being of her counterfactual

¹⁷ A similar approach applied to preference utilitarianism is defended by Wlodek Rabinowicz in Rabinowicz and Österberg (1996).

¹⁸ Wlodek Rabinowicz, in Rabinowicz and Österberg (1996), argues that this problem can be solved. Even though non-actual persons cannot be assigned well-being from the perspective of our world, they can be assigned well-being from the perspective of worlds in which they exist and have attitudes, because from this latter perspective they are actual. I argue in **** (1998) that this shifty actualism has controversial meta-ethical implications.

counterpart. But this means that one of the main virtues of a preference-based theory is lost. It does no longer provide us with a flexible theory that takes into account changes in a person's personality and character when determining her well-being.

8. Think comparatively!

Economists and philosophers oriented towards economics would say that we should forget about absolute values and simply reject *World-Bound Well-Being*. The only sensible option is to be a comparativist and exclusively focus on comparative value (betterness, worseness, equality in value) and let a person's comparative attitudes concerning two worlds determine the comparative values of the worlds. Now, since in the present context the preferences may change across worlds, it is not clear what the necessary and sufficient conditions for comparative value should be according to the comparativist, but he seems at least to be committed to *Dominance*.

One problem is that *Dominance* does not seem to be especially attractive in contexts where the polarity of the attitudes changes across worlds. In Case 4, the comparativist has to say that the life world w_1 is better than the life in w_2 even though the person would be indifferent towards his life in w_1 and would favour his life in w_2 . Comparativism seems all too insensitive to non-comparative attitudes.

The comparativist could respond by arguing that, in Case 4, the person will feel regret in world w_2 , since here she will prefer the alternative life. The basic idea is that it is more important to prevent grousing than to give a person what he would favour.

This is not a convincing reply. As I will argue later, the feeling of regret is an unwanted experience that should be reflected in the global attitudes. More importantly, there are regret-free cases where it seems clearly wrong to go by comparative attitudes. Consider the following case:

Case 5

	w1	w2	..
w1	-20	-20	
w2	20	20	
:			

If we should think comparatively, then surely we have to go by the invariant attitudes of comparative indifference in this case and say that w1 is equally as good as w2 for the person. But this is absurd since she would detest her life in w1 but would favour her life in w2.

It is not even clear that the comparativist has a coherent theory to offer. Repeated applications of *Dominance* generate a circular value-ordering. Suppose we have the following preference profiles over the lives in three worlds (> stands for preference):

- w1: w3 > w1 > w2
- w2: w1 > w2 > w3
- w3: w2 > w3 > w1

Since w1 is preferred to w2 in both w1 and w2, *Dominance* implies that w1 is better for the person than w2. Similarly, since w2 is preferred to w3 in both w2 and w3, w2 is better for her than w3. But since w3 is preferred to w1 in both w3 and w1, we have to say that w3 is better for her than w1 and we end up in a circle. (Note that this betterness circle is not generated by circular preferences. In each world, the person's preferences are transitive.)

Though it is a contested issue whether circular betterness is conceptually impossible, it is definitely not an attractive feature of a well-being theory.¹⁹ It makes it difficult to use the theory as a guide to action since it is not clear how we should define prudential rightness when no action maximizes well-being.

¹⁹ For an argument for circular betterness, see Temkin (1996). For useful criticism of this argument, see Broome (2004), ch. 4.

9. Think vertically!

The idea here is to aggregate the values in each column: a person's well-being in w is some function of values in the w -column. More informally, the value of a person's life in a world w is determined by how well her life in w matches her attitudes in w and her attitudes in other possible worlds. The inconsistency would be avoided since a life in a world w is assigned a unique value on the basis of all the values in the w -column. A positive value means a good life, a negative value a bad life, and zero value a neutral life. A higher value means a better life.

It seems to be a non-starter to claim that all *logically possible* attitudes of a person are relevant to how well-off she is in a particular possible world. There is an infinite number of different logically possible attitudes, and, moreover, they seem to cancel each other out. For any possible favouring of a life in a world, we can find a possible disfavouring of a corresponding strength, and vice versa.²⁰ Some restriction on relevant possible worlds must be imposed. It would perhaps be more reasonable to limit the relevant attitudes to those that are within the reach of some agent, (the welfare-subject herself, perhaps). But even this seems too permissive. Suppose the w_1 and w_2 are available and that the attitudinal profile is the following:

Case 6

	w1	w2	..
w1	0	20	
w2	0	20	
:			

It seems clear that the life in w_2 is better than the life in w_1 , at least if we assume that her attitudes in w_1 and w_2 concerning w_3 and the rest are identical. The person would prefer w_2 to w_1 no matter which world were to be realized, and she would be cold towards her

²⁰ For a similar collapse argument, see Rabinowicz and Österberg (1996), pp. 17-18.

life in w1 (if w1 obtained), but would love her life in w2 (if w2 obtained.) However, suppose there is a third available world w3:

Case 6*

	w1	w2	..
w1	0	20	
w2	0	20	
w3	50	5	
:			

Should the attitudes in w3 have a say about the relative values of w1 and w2? I can't see why, if we assume that the attitudes in w3 are no more rational, informed or autonomous than the attitudes in w1 and w2. More generally, the lives in two worlds should be valued independently of attitudes in other worlds.

Allowing attitudes in one world to affect the value of a life in another world has also some implausible implications for comparisons of worlds that *differ* in attitudes. Go back to Case 5:

	w1	w2	..
w1	-20	-20	
w2	20	20	
:			

According the account I have just sketched, the fact that the life in w2 is strongly disfavoured in w1 counts against the life in w2. But that seems implausible.²¹ If the

²¹ Of course, your attitude in w2 towards your life in w1 will count *positively* towards this life. But if no special weight is given to the attitudes you have in a world towards your life in the same world, this will imply, implausibly, that your lives in w1 and in w2 have the same value, despite the fact that you will hate

attitudes in w1 and w2 are on par with respect to how well-informed, rational, and autonomous they are why give any weight to the attitudes in w1 when deciding on the value of the life in w2? If you were to lead the life in w2, you would love it and have no regret.

10. Think horizontally!

The idea here is to aggregate the values in each row: a person's well-being in w is some function of the values in the w-row. Incoherence is avoided, since each life in a world is assigned unique value on the basis of the row-values for that world. A positive value means a good life, a negative value a bad life, and zero value a neutral life. A higher value means a better life. It is not clear how this function should look and how it should be motivated. I can see three main alternatives.

(1) *Regret*. One option is to say that how well off I am in a world depends not only on what I feel about my life in that world but also how much I regret not living an alternative life. The row-values are then used to define a regret-factor by taking the difference between the value I assign to my actual life and the value I assign to the alternative life. An example:

Case 7

	w1	w2	
w1	5	2	
w2	20	10	

How well off I am in w2 depends on the intensity of my favouring of my life here (10) and the regret-factor (10-20). Even though my life in w2 would be favoured, the regret-factor in w2 tells against my life in w1. How much weight to give to the regret-factor is

your life in w1 and love your life in w2. So, what we end up with is a violation of *World-Bound Well-Being*.

an open question. A simple version would state that the value of a life in a world w = the intensity of the absolute attitude in w towards the life in w + the regret factor. If there is no higher-ranked alternative, the regret-factor is zero. If there is more than one higher-ranked alternative, the regret-factor should be defined in terms of the alternative that is ranked the highest (maximum regret).²² But this simple version will obviously violate World-Bound Well-Being, since a sufficiently great regret-factor will outweigh the intensity of a favouring. A more plausible version would make the regret-factor a *tie-breaker* so that if I love my life to the same degree no matter which world is realized, then the life with the least regret is the better life. This means that if the case is like this

Case 8

	w1	w2	
w1	5	2	
w2	10	5	

the fact that the regret-factor is negative in $w2$ but zero in $w1$ makes $w1$ better for me than $w2$.

This looks like a plausible view, but I doubt that it holds water. In Case 8, $w1$ is better for me than $w2$ given the assumption that $w1$ and $w2$ are the only alternatives considered by my $w1$ -self and $w2$ -self. This evaluation will change if we add a third alternative, $w3$, about which my $w1$ -self and $w2$ -self feel differently:

Case 8*

	w1	w2	w3
w1	5	2	20

²² Obviously, this is only guaranteed to work if the number of lives considered is finite. If the number is infinite, there might not be a top-ranked life, in which case maximum regret is not well-defined.

w2	10	5	5

The regret-factor for w2 it will still be -5, whereas for w1 it will now be -15. This means that if I consider this third alternative, w1 is no longer better for me than w2. It is surely odd to say that whether one life is better for me than another depends on which *other* alternatives I consider. Note that w3 might be some merely *logically possible* life, not accessible to me. This also means that the mere fact that, in w1, I imagine w3 as a blissful life will make w1 come out as worse for me than w2.

One could try to fix this by defining the regret-factor in terms of the highest-ranked *feasible* alternative life. But then the comparative well-being assessment of lives will depend on the set of feasible alternative lives. I doubt that it makes much sense to claim that whether I am intrinsically better off leading a certain life depends on my options in this way. It is true, of course, that whether I *feel* regret will often depend on what I think are my feasible options. It is common to feel the strongest regret about options we know were in our hands.

But this suggests that the regret-factor as it is defined seems redundant. Suppose that I am satisfied with my actual career, but feel deep regret that I was never came round to writing a book that gave proper expression to what I thought of as my best ideas.²³ The fact that I feel regret seems relevant to my well-being. But recall that the attitudes I am focusing on are global, about my life as a whole. To determine my well-being it is not enough to ask what I feel about my career, which is only one aspect of my life; we also need to know what I feel about having the career *while feeling deep regret*. When we know this we seem to have all the information necessary for taking proper account of regret. It is therefore crucial not to misread the numbers in my examples. They are not supposed to represent the amount of some one feature, say, money, or material wealth, we tend to care about; they represent the intensity of an overall attitude towards all relevant features of a life.

²³ Dennis McKerlie uses this example to defend a regret-sensitive view. See McKerlie (2007), p. 50.

(2) *Badness of frustrated comparative preference.* Instead of invoking a regret-factor one might think that it is bad to have frustrated comparative preferences.²⁴ Go back to Case 8:

	w1	w2	..
w1	5	2	
w2	10	5	
:			

One could claim that what makes w2 a worse alternative is not regret, but the fact that in w2 but not in w1 my comparative preference is frustrated. In w2, I prefer w1 to w2 but w1 does not obtain. In general, a comparative preference for x over y is frustrated when y but not x obtains does not obtain. The idea is then that how well-off I am in a world is determined by what I feel about my life in this world and the intensity of my frustrated preferences in this world. This would imply that the mere absence of higher-ranked lives makes my life worse. I don't think this theory works. If the absence of a higher-ranked life does not cause any unwanted feelings of regrets, how can it make things worse? It is true that by leading a lower-ranked life I might miss out on other higher-ranked lives that are good in my own light. But the mere absence of a good does not itself make a bad.

(3) *Badness of frustrated favouring.* Instead of invoking the frustrations of comparative preferences, we could invoke the frustrations of my favourings.²⁵ Note that in Case 8, in w2, my favouring of w2 is satisfied whereas my favouring of w1 is not. If we count the satisfaction of my favouring of w2, shouldn't we also count the frustration of my favouring of w1? More generally, if satisfied favourings make my life better shouldn't frustrated favourings make it worse? If we answer 'yes' to this question, we have a reason to choose w2 over w1, because the intensity of my frustrated favouring in w2 of w1 is stronger than the intensity of my frustrated favouring in w1 of w2.

Again, this theory would have implausible implications. Suppose, again, that I rank the worlds, w1, w2, w3, w4,..., wn, in the stated order, and that I favour all of them.

²⁴ This is suggested in McKerlie (2007), p.

²⁵ This is suggested in McKerlie (forthcoming).

According to the view in question, w_1 , the life that I strongly favour and prefer to all other alternatives, will still be bad in many respects, for it contains the frustrated favourings of w_2, w_3, \dots, w_n . But it seems absurd to say that the frustrations of these favourings make my life worse. They concern alternative lives that I favour less than my actual life.

Of course, by leading the top-ranked life I will miss out on other lives that are still good in my own light. But, again, the absence of a good does not make a bad. What makes a life worse is that I *disfavour* it in certain respects.

11. Think diagonally!

By now it might be fairly obvious what my favoured solution will be. I think we should decide cross-world comparisons by looking at the values in the *diagonal*. To decide whether the life in a world w is better than the life in another world w' for a person we should not focus on her comparative attitudes concerning these lives. We should instead focus on what absolute attitude she *would* have towards the life in w , if w obtained, and compare that attitude with the absolute attitude she *would* have towards the life in w' , if w' obtained. More exactly:

Diagonal well-being

Her life in w is better for S than her life in w' iff

- (i) S would *favour her life in w more*, if w obtained, than she would *favour her life in w'* , if w' obtained,
- (ii) S would *disfavour her life in w less*, if w obtained, than she would *disfavour favour her life in w'* , if w' obtained,
- (iii) S would *favour her life in w* , if w obtained, and she would *disfavour favour her life in w'* , if w' obtained,
- (iv) S would *favour her life in w* , if w obtained, and she would be *indifferent* towards *favour her life in w'* , if w' obtained, or

(v) S would be *indifferent* towards *her life in w*, if *w* obtained, and she would *disfavour* *her life in w'*, if *w'* obtained.

A shorter but slightly misleading formulation of this principle would be: her life in *w* is better for S than her life in *w'* iff S's *w*-self wants *her life in w* more than her *w'*-self wants *her life in w'*.²⁶

Absolute values are then defined in the following way:

Her life in *w* is good for S iff she favours, in *w*, her life in *w*.

Her life in *w* is bad for S iff she disfavours, in *w*, her life in *w*.

Her life in *w* is neutral for S iff she is neutral, in *w*, towards her life in *w*.²⁷

This theory avoids incoherence by sticking to *World-Bound Well-Being* but rejecting *Dominance*. Note also that this principle does not generate axiological variance. Whether the life in *w* is better for a person than the life in *w'* does not depend on whether *w* or *w'* obtains.

One might object to this theory on the grounds that it seems to presuppose that absolute attitudes are primitive and can't be reduced to comparative ones. But this is not so. My theory could be defended even if we defined favouring, disfavouring, and indifference in terms of preference in the following way:

S favours *x* iff S prefers *x* to something she is indifferent towards.

S disfavors *x* iff S prefers *y* to *x* and *y* is something S is indifferent towards.

S is indifferent towards *x* iff S is indifferent between *x* and the negation of *x*.²⁸

²⁶ Bricker (1980), pp. 381-401, seems to suggest a principle similar to mine, but he does not make explicit use of the attitudes of favouring, disfavouring, and indifference.

²⁷ Remember that we are assuming a highly idealized toy model here. These conditions will not do for a less idealized environment in which attitudes change across time. For instance, a life can be good without being favoured at all times. It is enough that the favoured patches make up for the disfavoured ones.

²⁸ Chisholm (1964), pp. 613-625.

Of course, I do have to assume that it makes sense to compare attitudes of different possible selves of the same person. I see no problem in comparing absolute attitudes with different polarity: favourings with disfavourings, favourings with indifferent attitudes, and disfavourings with indifferent attitudes. What could create a problem are comparisons of absolute attitudes that have the same positive or negative polarity. It is here the comparativist may think he has an advantage, since he only needs to make sense of comparisons of preferences. What does it mean to say that one possible self favours x more than another possible self favours y ?

In reply, I would first of all say that comparisons of this kind are commonplace. Think of examples such as ‘Jane loves John more than Jake loves Kath’. Surely, these comparisons make sense, even though we might disagree about how to make sense of them. Secondly, if favourings can be defined in terms of preferences along the lines presented above, then a comparison of favourings boils down to a comparison of preferences. To decide whether my x -self favours x more than my y -self favours y , we should compare my x -self’s preference for x over something he is indifferent towards with my y -self’s preference for y over something he is indifferent towards. Comparisons of favourings will then be comparisons of preference *differences*. The same reasoning can of course be applied to comparisons of disfavourings. I can’t, therefore, see that the comparativist has an advantage, if he assumes that it makes sense to compare preference differences across possible selves of the same person.²⁹ We are in the same boat. We both need to make sense of comparisons of preference differences.

It should be noted that my theory has still something to say even if drop this measurability assumptions. In order to give us guidance on how to compare lives that differ in the valence of the attitudes taken towards them we only need to make the minimal assumptions that lives I would favour are better for me than lives I would disfavour, or would be indifferent towards, and that lives I would be indifferent towards are better for me than lives that I would disfavour.

One striking aspect of my theory is that a life can be better for me even if I would not rank this life higher if I were to lead it. One could claim that this shows that my theory is

²⁹ If the comparativist denies this, his theory will be seriously impoverished, since he will then be unable to compare life-options that involve conflicting preferences of possible selves.

flawed.³⁰ One way to spell this objection out would be to say that a life is better for a person *only if* she would rank it higher, if she were to lead it. However, this is clearly not an acceptable constraint, for it would rule out saying that one life is better than another in all cases where we have a preference reversal of the kind exemplified in the bachelor's dilemma case ('To wed or not to wed'). Recall that in this case, whichever life is realized, I will prefer the alternative life. But, surely, we do not want to say that no life can be better in this kind of case. Take, for instance, Case 3. If I lead the life in w1, I will hate it and see the alternative life in w2 in a neutral light. If I lead the life in w2, I will love it but see the alternative life in w2 in an even better light. Surely, the fact that I would *hate* my life in w1 and would *love* my life in w2 speaks clearly in favour of the latter life.

But perhaps I have overstated the objection. Perhaps what is assumed is only that the fact that a life would not be ranked higher if it were realized speaks against that life to some extent. The problem with my theory, one might therefore argue, is that it does not give any weight to this fact. If my x-self would favour x more than my y-self would favour y, that, on my theory, decides the issue and x is deemed better for me. No weight is given to the fact that my x-self would not rank x higher than y. In response, I would say that the temptation to give weight to this fact is understandable, since we tend to read into this case a feeling of regret or restlessness in leading a life that you would not find optimal. But, as argued earlier, we do not normally feel regret just because we imagine a blissful life we know is merely logically possible and not accessible to us. Furthermore, in those cases we do feel regret or restlessness, this feeling is something that our global attitudes will take into account. The more you care about this negative feeling, the less you favour your life as a whole. Once these feelings have been taken into account by our global attitudes, I can see no reason to give special weight to the fact that a certain life, if realized, would not be seen as optimal.

12. Conclusions

³⁰ This objection was pressed by Luc Bovens and an anonymous referee.

We have thus solved the problem of deciding which life is best for a person whose attitudes are not stable across possible worlds. It is a mistake to look for a single vantage point identified with the attitudes of one of the person's many possible selves. Instead, each of the person's possible selves should have a say, but only about the world they inhabit. In order to decide whether a life *x* is better for her than another life *y*, we should consider her *x*-self's attitudes towards *x* and compare those with her *y*-self's attitudes towards *y*. If her *x*-self wants *x* more than her *y*-self wants *y*, then *x* is better for her than *y*, (at least if we assume that both *x* and *y* are equally objectively desirable.)

Of course, my solution does not address all pressing problems concerning preference change. Most importantly, it does not deal with preference conflicts across time and the creation and satisfaction of new preferences and desires. But I hope to have shown that the theory defended in this paper is one important building block in a complete theory of well-being.³¹

References

- Bricker, P., 1980, 'Prudence', *Journal of Philosophy*, Vol. LXXVII, No 7, pp. 381-401.
- Broome, J., 2004, *Weighing Lives*, OUP, Oxford.
- Bykvist, K., 1998, *Changing Preferences. A Study in Preferentialism*, Acta Universitatis Uppsaliensis.
- Bykvist, K., 2003, 'The Moral Relevance of Past Preferences', in Dyke, H. (ed.), *Time and Ethics: Essays at the Intersection*, Kluwer, pp. 115-136.
- Bykvist, K., 2006, 'Prudence for Changing Selves', *Utilitas*, Vol. 18, No. 3, pp. 264-283.
- Bykvist, K., 2006, 'What are desires good for? Towards a coherent endorsement theory', *Ratio* 14: 286-304.
- Bykvist, K., 'Violations of Normative Invariance. Some Thoughts on Shifty Oughts', *Theoria*, Vol. LXXIII, Part 2, 2007.
- Carlson, E., 1995, *Consequentialism Reconsidered*, Kluwer.

³¹ (Self-identifying reference deleted.)

- Chisholm, R., 1964, 'The Descriptive Element in the Concept of Action', *The Journal of Philosophy*, 61, pp. 613-625.
- Darwall, S., 1999, 'Valuing Activity', in Paul, E., Miller, F., and Paul, J. (eds.), *Human Flourishing*, Cambridge, Cambridge University Press.
- Dworkin, R., 2002, *Sovereign Virtue*, Harvard University Press.
- Feldman, F., 2005, 'Basic Intrinsic Value', in Rønnow-Rasmussen, T., and Zimmerman, M. (eds.), *Recent Work on Intrinsic Value*, Library of Ethics and Applied Philosophy:17, Springer.
- Gibbard, A., 1992, 'Interpersonal comparisons: preference, good, and the intrinsic reward of a life' in Elster J., Hylland, A., (eds.), *Foundations of Social Choice Theory*, Cambridge, Cambridge University Press.
- Goldman, A. I., 1970, *A Theory of Human Action*, Princeton University Press.
- Hurka, T., 2001, *Virtue, Vice, and Value*, Oxford University Press.
- Kraut, R., 1994, 'Desire and the Human Good', *Proceedings and Addresses of the American Philosophical Association*, vol. 68, no. 2: 39-54.
- McKerlie, D., 2007, 'Comments on Bykvist 'Prudence for Changing Selves'', *Utilitas* 19: 47-50.
- Parfit, D., 1992, *Reasons and Persons*, Clarendon Press, Oxford, Appendix I.
- Rabinowicz, W., Österberg, J., 1996, 'Value based on preferences. On two interpretations of Preference Utilitarianism', *Economics and Philosophy* 12: 1-27.
- Taylor, R., 1966, *Action and Purpose*, Prentice-Hall, Inc.
- Temkin, L., 'A continuum argument for intransitivity', *Philosophy and Public Affairs* 25: 175-210.