Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
doi: 10.1111/phpr.12334

Extended Preferences and Interpersonal Comparisons of Well-being

HILARY GREAVES
Oxford University

HARVEY LEDERMAN
University of Pittsburgh

An important objection to preference-satisfaction theories of well-being is that these theories cannot make sense of interpersonal comparisons of well-being. A tradition dating back to Harsanyi (1953) attempts to respond to this objection by appeal to so-called *extended* preferences: very roughly, preferences over situations *whose description includes agents' preferences*. This paper examines the prospects for defending the preference-satisfaction theory via this extended preferences program. We argue that making conceptual sense of extended preferences is less problematic than others have supposed, but that even so extended preferences do not provide a promising way for the preference satisfaction theorist to make interpersonal well-being comparisons. Our main objection takes the form of a trilemma: depending on how the theory based on extended preferences is developed, either (a) the result will be inconsistent with ordinary preference-satisfaction theory, or (b) it will fail to recover sufficiently rich interpersonal well-being comparisons, or (c) it will take on a number of other arguably odd and undesirable commitments.

1. The Preference-Satisfaction Theory of Well-being

What constitutes well-being? That is, what makes a person's life go well *for* that person? The basic menu of candidate theories is familiar. Hedonists hold that well-being consists in pleasurable experiential states. Attitude-satisfaction theorists hold that well-being consists in the satisfaction of one's (actual or somehow-idealised) desire-like attitudes. 'Objective list' theorists hold that well-being consists in the possession of various items (perhaps including: understanding, accomplishment, deep interpersonal relationships, freedom, and aesthetic beauty) whose place on the list is not due to their being desired or causing pleasure; while the list in question can also *include* pleasure and desire-satisfaction, the typical objective list theorist denies that those are the *only* things on the list.

The attitude-satisfaction theory is motivated by the idea that if a person gets what she wants, her life goes better. Different versions of this theory are based on different desire-like attitudes. In philosophy, the most familiar is the desire satisfaction theory, which holds that the attitude relevant to the determination of well-being is desire itself, a binary relation between an agent and some other entity, the object of the agent's desire.

According to the desire satisfaction theory, an agent's life goes better to the extent that her desires are satisfied.

Simple versions of the desire satisfaction theory are, however, unpromising. According to perhaps the simplest of these theories, for example, a person's well-being is determined by the number of her desires which are satisfied. This theory fails because it does not take account of differences in the strengths of different desires. A person's strong desire for a good job should not be counted on a par with her weak desire for a new bicycle: if the person gets a good job (but no bicycle), her life goes better than if she gets a new bicycle (but no job). But the simple theory fails to deliver this result: in each case the agent has satisfied one desire, so it falsely predicts that the different outcomes yield equivalent improvements to her life. The same objection applies to close variants of the simple theory; for example, it also applies to a theory which takes a person's well-being to be determined by the percentage of her desires which are satisfied.

The obvious response for the attitude-satisfaction theorist in the face of these objections is to amend the theory to incorporate a notion corresponding to strength of desire. The most popular way of doing this—indeed the only one which appears to have been worked out in any detail—is to abandon the desire-satisfaction theory, and move to a preference-satisfaction theory. A preference-satisfaction theory of well-being holds that the attitudes relevant to the determination of well-being are *preferences* (as opposed to desires), where a preference is a ternary relation between a subject and two further entities, one of which the agent prefers to the other. According to the preference-satisfaction theory, an agent's life goes better to the extent that her preferences are satisfied. This theory explains the phenomena we earlier described in terms of strength of desire in terms of an entity's position in the agent's preference ordering. Whereas before we spoke of agent having a stronger desire for a good job (while still having no bicycle) and a weaker desire for a bicycle (while not having a good job), now we speak simply of her preferring having a good job (but no bicycle) to having a bicycle (but not having a good job). For the bulk of the remainder of the paper, we will assume that the attitude-satisfaction theorist will respond to this problem about strength of desire by adopting a preference-satisfaction theory. In the closing section of the paper, we will briefly reconsider whether it was this move—from a desire-satisfaction theory to a preference-satisfaction theory—which led the attitude-satisfaction theorist astray.

A simple form of the preference-satisfaction theory is the unrestricted actual-preference theory. According to this theory, one state of affairs is better than another for a given individual just in case that individual *actually, all things considered* prefers the first to the second.

This simple version of the preference-satisfaction theory is subject to three standard objections. First: some of a person's preferences (for example, those driving a person to donate to charity) are for things which don't intuitively contribute to her own well-being, so that an *unrestricted* preference-satisfaction theory counts too many things as relevant to her well-being. Second: whereas it is highly plausible that the better-off-than relation satisfies certain logical constraints—for example, it cannot be that a person in situation A has greater well-being than she would in situation B, that she has greater well-being in situation B than in situation C while nevertheless having greater well-being in C than in

A^1—it is at least somewhat plausible that many people's preferences do not satisfy these logical constraints: for example, at first sight it seems possible that a person might prefer ice cream to pie, pie to cake, and cake to ice cream. Third: some of a person's actual preferences are defective *even by the agent's own lights*, because they are the product of misinformation. For example, if Ahmed wants to catch the 7pm train and mistakenly believes that for this it is necessary to leave the house at 6pm, he will prefer leaving the house at 6pm to leaving the house at 6.30 even though he otherwise prefers being at home for longer; but if in fact leaving at 6.30 would suffice for catching the train, the preference-satisfaction theorist will not want it to follow that leaving at 6 is *better for Ahmed*.

It is straightforward, however, for the preference-satisfaction theorist to answer these three objections, by refining her theory. The requisite refinements are clear: the theory must count only *self-regarding* preferences, and must appeal, not to actual preferences, but to (something like) the preferences that the person in question would have under suitably specified conditions of rationality (at least eliminating the kind of 'cycle' we saw above with pie, cake and ice cream) and full information (knowledge of basic worldly facts such as the time it takes to get to the train station). It is a somewhat subtle matter precisely how the line is to be drawn between preferences which are 'self-interested' and those which are not, and precisely what should count as conditions of 'rationality' and 'full information'. For example, some theorists require the idealised preferences to agree with the agent's actual, unidealised, fundamental values, while others permit idealisation to result in a change of fundamental value, so long as the new preferences are those that the agent would have, were she to be idealised in the relevant way. For the most part, we will simply assume that this kind of subtlety has already been dealt with, and attempt to remain neutral on how it has been resolved. But we will assume that 'full information and rationality' does not involve 'substantive' conditions of rationality that can be justified only by appeal to prior claims about what is objectively better or worse for the individual in question (e.g. that a life spent counting blades of grass is objectively worthless). This is required if the resulting theory is to be truly a preference-satisfaction theory of well-being, as opposed to one in which much or all of the work is ultimately done by some other account of objective betterness, A^1 itself independent of matters of preference.

Let us grant that these refinements can be used to give an adequate response to the initial objections. Even so, the theory faces an important further objection: the *problem of interpersonal comparisons*. Roughly, the problem is that there are at least some positive facts concerning comparisons of one person's well-being to another person's, but it is unclear whether a preference-satisfaction theorist can make sense of these facts. This problem for the preference-satisfaction theory is the subject of the present paper. We will argue that one popular approach to recovering interpersonal comparisons within a preference-satisfaction theory of well-being, the 'extended preferences program', does not ultimately lead to an attractive way of saving the preference-satisfaction theory.

The remainder of the paper is structured as follows. Section 2 states the problem of interpersonal comparisons in more detail, outlines why this issue might be thought particularly problematic for a preference-satisfaction theory of well-being, and describes why extended preferences might seem to be a promising way of resolving

[1] *Pace* Rachels (1998) and Temkin (1987; 2014).

the problem. Section 3 takes up the question of whether we can make sense of extended preferences, and of what kind of attitude they could be; our conclusion here will be that these issues are less problematic for the extended preferences program (henceforth, 'EP program') than other authors have argued, so that *thus far* the EP program is in good shape. Section 4 sketches a simple formal framework, based on this conceptual understanding of what extended preferences are, which we will then use in the remainder of the paper.

We then begin to approach our main objection to the program, with a brief, preparatory section. Section 5 considers a principle endorsed by Harsanyi, which we will call the 'Principle of Full Coincidence', according to which all individuals ultimately have the same extended preferences. We argue that the principle is false, at least if 'fully informed and rational' preferences are understood in the usual, relatively minimal, ways. This means that a full formulation of the EP program faces a *prima facie* problem of how to aggregate diverse extended preferences to generate a single well-being ordering; for the purpose of this paper, however, we set this problem aside (we deal with it in detail in a companion paper, Greaves & Lederman (2016)).

Our main challenge to the EP program takes the form of a trilemma, and occupies the next four sections (sections 6–9). We argue that the extended preferences program must either (a) accept a well-being ordering which fails to respect individuals' preferences about their own lives, (b) admit massive incomparability in interpersonal well-being comparisons, or (c) endorse a deeply unattractive form of holism and ungroundedness in the connection between preferences and betterness. Since the preference-satisfaction theorist should hope to avoid each horn of this trilemma, we suggest, in the conclusion (section 10) that the would-be preference satisfaction theorist should abandon extended preferences *qua* foundation for interpersonal well-being comparisons. In closing we sketch two alternative approaches to this problem which we believe may hold more promise than the use of extended preferences.

2. Interpersonal Comparisons and Extended Preferences

2.1. The Problem of Interpersonal Comparisons

In more detail, then: it is a datum that there are at least some positive facts concerning interpersonal comparisons, of each of two forms. Interpersonal *level* comparisons are facts of the form: state of affairs x is better for person S than state of affairs y is for person T. (In the simplest examples, x and y are identical. For instance, the actual state of affairs is better for either of us than it was for the average slave in ancient Rome.) Interpersonal *unit* comparisons are facts of the form: the amount by which state of affairs x is better for person S than state of affairs y is greater than the amount by which state of affairs v is better for person T than state of affairs w. (If Donald donates \$5 to the Against Malaria Foundation, thereby causing an additional child in a malarial region to be provided with a mosquito net, the reduction in his well-being that is occasioned by the sacrifice of the \$5 is far less in magnitude than the expected gain in the child's well-being resulting from her access to protection from malaria; his firm belief that this is so is of course the reason for his donation.) A theory of well-being that denies that any such facts exist is implausible. It is also implausible in a particularly pernicious way: any sane approach to large-scale public policy analysis requires appeal to some sensible 'social

welfare function', but any such function (utilitarian, prioritarian, egalitarian, maximin or anything else) in turn presupposes the existence of some interpersonal well-being comparisons of at least one of the above two forms.[2]

Why suspect that the preference-satisfaction theory will have difficulty in recovering such comparisons? Because the facts in which the preference-satisfaction theory seeks to ground all well-being facts consist of one *apparently entirely separate* ranking of states of affairs for each person. Emily's preferences are given by one ordering of the set of all possible states of affairs (or lotteries thereon, that is, probability distributions over states of affairs); Faruk's preferences are given by a second, potentially radically different, ordering of states of affairs (or lotteries thereon). It is clear how a profile of such preference orderings can ground *intra*personal level comparisons, and arguably clear (given an appeal to lotteries, together with the machinery of decision theory) how it might ground intrapersonal unit comparisons.[3] But these two people's preferences do not exhibit any obvious *interpersonal* structure: anything that is structurally of the right type to ground interpersonal comparisons.[4] In contrast, an hedonist (respectively, an objective-list) theory of well-being can ground interpersonal comparisons by considerations of, for example, similarity of brain states (respectively, objective circumstances more generally).

2.2. Extended Preferences: A First Pass

The basic idea of the EP program is to meet this challenge by claiming that the *objects* of the preferences which determine well-being are richer than we have hitherto supposed. We begin with a toy example. Suppose that Makena and Laurence are in a restaurant, each deciding whether to order meat or fish. Suppose that Makena prefers meat to fish, while Laurence prefers fish to meat. These are what we might call the individuals' *ordinary* preferences; certainly, individuals have such ordinary preferences. But, proponents of the EP program suggest, Makena and Laurence *also* each have preferences over the following *four* alternatives: being Makena, having Makena's preferences, and eating meat; being Makena, having Makena's preferences, and eating fish; being Laurence, having Laurence's preferences, and eating meat; being Laurence, having Laurence's preferences, and eating fish. In discussions of the EP program, alternatives of this sort are called 'extended alternatives', and individuals' preferences over extended alternatives are called their 'extended preferences'. Although we will be arguing that extended preferences are just preferences, so that from our perspective this terminological distinction (with its suggestion that extended preferences are different in some important way from ordinary preferences) is somewhat misleading, we will continue to use the standard terminology.

---

[2] The so-called 'new welfare economics' notoriously tries to issue policy guidance without recourse to interpersonal well-being comparisons, via its 'potential Pareto improvement' criterion. As is increasingly recognised, however, this approach effectively amounts to smuggling in interpersonal comparisons via a (highly implausible) hypothesis that a marginal dollar corresponds to the same marginal change in well-being for each person, regardless of existing wealth level.

[3] The identification of ratios of well-being differences with ratios of von Neumann-Morgenstern utility differences is 'Bernouilli's hypothesis'. This hypothesis has been contested, notably by Sen (1976; 1977) and Weymark (1991; 2005). For a defence of the identification, see e.g. Broome (2008); Greaves (2016).

[4] The point is simply that the interpersonal structure is not *obviously* there; in section 10, we discuss briefly a 'structuralist' approach which generates interpersonal comparisons by first considering the structure of individuals' preference-orderings.

From a formal perspective, the claim that people have extended preferences is important because, unlike an ordinary preference ordering, an extended preference ordering has sufficient structure to ground interpersonal well-being comparisons. An extended-preference for being Makena and eating meat over being Laurence and eating fish is already equivalent, mathematically, to a betterness relation which makes interpersonal level comparisons (e.g. between being Makena and eating meat on the one hand, and being Laurence and eating fish on the other); similarly, extended preferences that rank *lotteries* over extended alternatives will (again via the standard machinery of von Neumann-Morgenstern decision theory) yield utility functions that can serve as the ground of interpersonal unit comparisons.[5] From a conceptual perspective, too, extended preferences seem promising for preference satisfaction theory: if one was happy using individuals' non-extended preferences to ground intrapersonal well-being comparisons, then there seems no reason why one would not for the same reasons be happy with using extended preferences to ground interpersonal comparisions; the extended preferences approach, for example, would equally promise to ground well-being facts in naturalistically unmysterious facts about preferences, and would equally establish an intimate relationship between what is preferred and what is better. The use of extended preferences is thus both natural and well-motivated from the standpoint of a theorist who seeks to ground interpersonal well-being comparisons solely in facts about (*some* sort of) preferences. In addition to Harsanyi himself, Arrow (1963; 1978) and Adler (2012; 2014; 2016a) have also used extended preferences for this purpose.

It is not merely that it would be convenient for the preference-satisfaction theory if people did have extended preferences. The evidence that we have extended preferences is, on the face of it, as strong as the evidence that we have ordinary preferences. It is just as acceptable in English to say "John would prefer being Queen Victoria to being Spartacus" as it is to say "John would prefer to ride your bike than to ride his". Admittedly, there are some linguistic differences between descriptions of the two preferences: we might more normally say that John *prefers* riding your bike, but only that he *would prefer* being Queen Victoria. But the fact that John is not habitually Queen Victoria, whereas he does (presumably) habitually ride his bike is sufficient to explain this difference: the present tense "John prefers riding my bike to riding his" carries an implicature that John has ridden each of these bikes in the past. Thus although uttering "John prefers being Queen Victoria to being a Roman slave" would be *inappropriate* unless as a matter of fact John had somehow managed the metaphysically impossible trick of being Queen Victoria in the past, it would not obviously be false.

It seems, then, that the only putatively significant conceptual difference between John's two preferences is that his preference for being Queen Victoria as opposed to Spartacus cannot be directly manifested in corresponding choice dispositions, since it concerns states of affairs that the subject could not bring about. Although there is a technical use of the word 'preference' in the economics literature, where the word is simply stipulated to be synonymous with 'choice-disposition', we take the preference-satisfaction theorist to be concerned with the attitude of preference which features in

[5] See n. 3.

folk psychology, and not with this technical usage, which has no simple correspondence with desire-like mental states. And it is obvious that for this more ordinary notion of preference there are at least some preferences that are not equivalent to choice dispositions (Arrow (1978, 224) uses the example of someone who is sick and would prefer to be well, even though there is no way to bring this about). So although preferences over who one is do not have the connection to choice that some preferences do, there is no threat from this quarter that preferences over who one is would thereby not be preferences.

3. Conceptual Foundations

3.1. What Extended Preferences are Not

Notwithstanding this *prima facie* case, Matthew Adler has recently called into question whether the notion of extended preferences ultimately makes sense. What exactly is it for Makena to prefer *being Makena and eating meat* to *being Laurence and eating fish*?

The reason this question is supposed to point to a problem for the EP theory will become evident when we turn to some answers that have hitherto been suggested. We will now survey three of these answers, none of which is entirely satisfactory, before suggesting our own resolution (or dissolution) of the puzzle.

First suggestion: for Makena to prefer *being Makena and eating meat* to *being Laurence and eating fish* is for Makena to prefer a state of affairs in which *Makena is identical to Makena and eats meat* to a state of affairs in which *Makena is identical to Laurence and eats fish*.

As Adler points out (2012, 198–199), this suggestion is obviously problematic. Since Makena and Laurence are distinct individuals, with (moreover) incompatible essential properties, the proposition that Makena is identical to Laurence is metaphysically impossible. Further we may suppose that it is *obvious* to both Laurence and Makena that it is metaphysically impossible. There is thus no prospect of arguing that Laurence and Makena can coherently have preferences regarding such in-fact-impossible propositions from their impoverished epistemic standpoint: even those who countenance the use of impossible situations as the content of attitudes should not countenance *blatantly* impossible ones (Lewis (2004)). At least arguably, if asked for her preferences regarding such propositions as that Makena is Laurence and eats fish, Makena should reply that she is unable to make sufficient sense of this proposition to place it in any preference ordering; (it is too clear to her that) there is nothing it could be like for Makena to be identical to Laurence.[6]

Second suggestion: Makena's so-called 'extended preference' is not really a *preference* at all. Rather, Makena *believes* that Makena eating meat is better off than Laurence eating fish.

[6] Adler also considers a variant of this first suggestion, according to which the subject's essential properties are 'screened off' so as to avoid this problem. Besides the *ad hoc* nature of the modification, however, this variant also fails, for the reasons Adler gives (2012, 206); the basic problem arises because even a subject's essential properties can be relevant to her well-being.

This is essentially the proposal that Adler himself adopts (2012, 210–211; 2014, 144).[7] From the point of view of the EP program *qua* program for solving the problem of interpersonal comparisons within a preference-satisfaction theory of well-being, though, to accept this second suggestion would be to give up the game. The game was to ground interpersonal well-being comparisons in *preferences* (of some sort). If 'extended preferences' are not really preferences at all, then they cannot serve this purpose (and nor has anything else been said to indicate that the content of Makena's interpersonal-comparison belief is in any way grounded in any preferences).[8]

Voorhoeve (2014) makes a third suggestion. Consider a subject suffering from severe amnesia, so that she has temporarily forgotten virtually everything about herself. In particular, she has forgotten whether she is Makena, Laurence or anyone else, and even whether she is male or female. It apparently makes sense to ask such a subject (even if she is *in fact* female, and even if she is in fact *essentially* female) what she would prefer to learn: would she prefer to learn that she is Makena and about to eat meat, or that she is Laurence and about to eat fish? According to the third suggestion, then, for Makena to prefer being Makena and eating meat to being Laurence and eating fish is for her to be such that *if she were* behind such a veil of ignorance, she would prefer to learn the first fact about herself than the second.

According to Voorhoeve, while this third suggestion is somewhat promising, its success or failure hinges on controversial matters in the metaphysics of personal identity. If it is to serve its purpose at all, the 'veil of ignorance' involved in this construction obviously cannot strip the subject behind the veil of all of her *preferences*. But suppose that Makena's preferences are radically different from Laurence's, and suppose further that a strong degree of psychological similarity (including similarity of preferences) is a necessary condition for two person-stages to count as stages in the life of the same person. Then it might already be too-obviously impossible for Makena, even given the impoverished self-knowledge that is permitted her behind this veil, to turn out to be identical to

---

[7] Adler's proposal differs from the one we discuss here, in two inessential respects. First, Adler does not use the word 'believe', as we have done; he consistently says that the interpersonal preferences are grounded in 'judgements'. But Adler's use of 'judgement' and its cognates seems to us equivalent to the standard uses of 'belief' and its cognates, so the difference appears to be insignificant. Even if the terms are intended to have different meanings, the point in the main text would still apply, since whatever judgements are, they are not preferences. Second, Adler does not think that all extended preferences are grounded in judgements (beliefs); he thinks that only 'preferences' between pairs of alternatives with different centres are in fact judgements. But the difference is irrelevant to the point in the main text: the EP program will fail if *any* extended 'preferences' used in determining well-being comparisons turn out not to be preferences.

[8] The argument in the main text assumes that preferences are not reducible to beliefs. But there is independent reason to think that the preference-satisfaction theorist cannot hold that preferences are reducible to beliefs (for discussion of a different problem with this identification see e.g. Lewis (1988; 1996), Price (1989), Byrne & Hájek (1997)). For according to the preference-satisfaction theory, for x to be better for Bethan than y is for Bethan to prefer x to y. But if in addition preferences were reducible to beliefs, then for Bethan to prefer x to y would in turn be for Bethan to believe that x was better than y. So, for x to be better for Bethan than y would be for Bethan to believe that x is better for her than y. There may well be informative partial definitions of this kind; for example, it may be that to be cool is to be thought to be cool (see for discussion, Fine (2012)). But plausibly being good is quite different from being cool; there must be more to *goodness* than merely being thought to be good. Since the reduction of preferences to beliefs leads to what we think is an unattractive consequence in the context of the preference-satisfaction theory, for the remainder of the paper, we will assume from now on that preferences cannot be reduced to beliefs.

Laurence. Thus Voorhoeve concludes that whether or not extended preferences are coherent depends on the extent to which psychological similarity is necessary for personal identity.

There are several problems with Voorhoeve's argument against this view.[9] And in fact, our own proposal will to some degree vindicate the suggested biconditional relationship between extended preferences on the one hand, and preferences over news items behind a suitably specified veil of ignorance on the other. But in our view, this is at best a contorted way of making the point: we will argue that the appeal to the veil of ignorance is, first, unnecessary, and, second, does not provide the most illuminating account of what extended preferences are. Even so, ideas which are hinted at, but not then adequately developed, in the veil of ignorance approach, will turn out to be key to a cleaner solution to the problem. We present these ideas in the next few sections.

3.2. Self-locating Beliefs and Self-locating Preferences

Let us take a step back, to consider a related issue which arises in the case of belief. In an important discussion of belief, John Perry (1979) describes (among others) a case in which he is in the supermarket, aware that someone is spilling sugar but unaware that it is he himself who is doing so. As Perry points out, it is not his realisation that *John Perry* is making a mess, but his realisation that *he himself* is making a mess, that eventually leads him to straighten up the sugar sack in his shopping cart. Similarly, David Lewis (1979) discusses (among others) the case of mad Heimson who has a delusional belief that he is Hume. Lewis suggests that Heimson's delusion is not characterised by the belief that *Heimson* is Hume (Heimson may know that Heimson isn't Hume), but rather the belief that *he himself* is Hume. Sometimes, the Latin terminology *de re* and *de se* is applied to this distinction: Heimson believes *de re* that Heimson isn't Hume, but he nevertheless believes *de se* that he himself is Hume. But the Latin terminology just gives a name to the problem: the question is how we are to make sense of the contrast between the aspects of Heimson's psychology captured in the first and the second of these descriptions. For it is clear that we must, extended preferences aside, make sense of this distinction.

As for beliefs, so also for preferences. For one thing, it is easy to construct examples for preferences which are analogous to those above for belief; indeed, Lewis himself constructs analogous cases for desire (Lewis, 1979, 528–531). For another, given the intimate links between cognitive and conative attitudes, it is implausible that the two kinds of attitudes could behave differently when it comes to this kind of basic distinction. If Heimson can believe that he is Hume without believing that Heimson is Hume, he can also prefer being Hume to not being Hume without thereby having a preference that Heimson be Hume as opposed to not.

The key point then is that the arguably-puzzling aspect of extended preferences—what it could mean for Makena to prefer (or disprefer) *being Makena* or *being Laurence*—is clearly of the same character as the arguably-puzzling aspect of the cases discussed by Lewis and Perry. If we can make sense of Perry's and Heimson's *de se* beliefs in a way

[9] First, the controversy over theories of personal identity notwithstanding, it is implausible that such *very strong* psychological similarity is *necessary* for personal identity. Second, Voorhoeve conflates the having of preferences with knowing that one has them: it does not follow, from the fact that the individual behind the veil must *have* preferences, that she must *know* her own preferences.

that distinguishes them from their *de re* beliefs, then we should, via the same machinery (whatever it is), also be able to make sense of extended preferences. The conclusion of this argument is that just as beliefs *de se* are, nevertheless, beliefs, so too extended preferences are just preferences.

But, one might ask, don't we still need the veil of ignorance? Heimson has *de se* beliefs which differ from his *de re* beliefs because in some sense he is mistaken about who he is. One might thus worry that extended preferences, like *de se* beliefs, require that subjects not know who they are. This would be problematic since the extended preference theorist wishes to claim that *everyone* has extended preferences, regardless of whether they know who they are. But the worry rests on a mistake: although arguably Makena cannot (coherently) believe that she will be cured of her illness while also knowing that she will remain sick, clearly Makena can prefer to be cured of her illness even if she knows she will remain sick. More generally, unlike the case of beliefs, it is not at all incoherent to have preferences for things one knows not to obtain. So Makena, like the rest of us, can have extended preferences—preferences over being other people—even *while knowing she is in fact Makena*. There is no need for any veil of ignorance, since preferences (extended or not) can be held over alternatives one knows not to obtain.

4. A Framework for Extended Preferences

4.1. Extended Preferences and Centred Worlds

That is the more general and basic point: We can be confident, even prior to settling on a treatment of cases such as Perry's and Lewis's, that whatever account explains the coherence of the relevant beliefs will also suffice for explaining the coherence of *de se* extended preferences. Moreover, this account must not merely vindicate the coherence of extended preferences; it must explain how extended preferences are in fact preferences. There is thus no danger that extended preferences will ultimately have to be rejected as incoherent, and also no danger that they must be reinterpreted as really being beliefs.

It will be useful, however—even at the cost of taking on some inessential and controversial commitments—to proceed at a level of lesser abstraction. A simple and well-developed formal treatment of Lewis's and Perry's examples takes the relevant beliefs to differ in content, and uses the notion of a 'centred world' to explicate this difference in content. We will now recall this general framework, and then show how to apply it to the case of extended preferences.[10]

We start with a standard benchmark framework for representing belief. On this standard account, the contents of people's beliefs in the simplest cases are *propositions*. For

[10] One of us doubts that the kind of self-locating ignorance described by Lewis and Perry is distinct from the general phenomenon of identity confusion—in which an agent assigns nonzero credence to the claim that one thing is two—and suspects that generalizations of the centred world framework are not the best way of treating identity confusion itself. But the aim here is simply to present a framework for representing extended preferences, which makes good on our claim that there is no conceptual challenge from this quarter to the coherence of extended preferences. Many of the accounts of Perry and Lewis's cases which treat them as instances of the more general phenomenon of identity confusion could plausibly be adapted to provide treatments of extended preferences as well (for some examples, see the survey in McKay & Nelson (2014)).

a person to believe that Hume was born in Edinburgh is for that person to stand in the belief relation to the proposition that Hume was born in Edinburgh. We adopt a simple formal treatment of propositions, where they are identified with sets of possible worlds:[11] the proposition that Hume was born in Edinburgh is the set of possible worlds in which Hume was born in Edinburgh. To believe that Hume was born in Edinburgh is thus to stand in the belief relation to the set of worlds in which Hume was born in Edinburgh.

To make sense of Lewis's and Perry's examples, we enrich this framework. Let a *centred world* be a pair (x, i), where x is a possible world (as we will say, the world on which the centred world (x, i) is 'based') and i is an individual (the 'centre' of the centred world (x, i)). In the enriched framework we identify *properties* with sets of *centred* worlds, and posit that the contents of beliefs are in general *properties*. For Perry to believe *de se* that he is making a mess is for Perry to stand in the belief relation to the property containing all centred worlds (x, i) such that individual i is making a mess in possible world x;[12] for Heimson to beleive *de se* that he is Hume is for Heimson to stand in the belief relation to the property containing all centred worlds (x, i) such that i is Hume. As promised, on the centred-worlds approach, the content of Heimson's *de se* belief that he was born in Edinburgh is different from the content of his *de re* belief that Hume was born in Edinburgh: the former is the set of centred worlds where the centre was born in Edinburgh; the latter is the set of worlds where Hume was born in Edinburgh.

Many beliefs, of course—the belief that Hume was born in Edinburgh among them—do not contain any element of 'self-location'. In our expanded framework, we can easily distinguish between beliefs that are self-locating and those that are not. Propositions are identified with special properties: specifically, those properties P such that for all centred worlds $(x, i) \in P$ and all individuals j, the centred world (x, j) is also in P. (Thus, as desired, a proposition, unlike an arbitrarily chosen property, does not 'distinguish between' any two centred worlds that agree on the the base possible world but disagree on which individual is the centre.) As we have already seen, for example: for Heimson to believe that Hume was born in Edinburgh is for Heimson to stand in the belief relation to the set of all centred worlds (x, i) such that in x, Hume was born in Edinburgh (regardless of the identity of i). A non-self-locating belief is then a belief whose content is a proposition; it is just that propositions are now seen to be a special case of a more general notion of the content of beliefs (that is, a special kind of property).

It is straightforward to develop an analogous framework, using centred worlds, to represent extended preferences. Once again we start from a standard benchmark framework, where 'ordinary' preferences are described by a binary relation on possible worlds. In this usual framework, there is no obvious way to model Makena's preferences between

---

[11] We need not *identify* propositions with sets of worlds; we could take them to be abstract objects which are isomorphic to sets of possible worlds. But for simplicity, we will speak as if propositions are identical to sets of worlds in the main text. It is of course a highly controversial matter what propositions are. We do not intend to endorse this theory of propositions, but are just using it to give a toy model which shows that one can make sense of extended preferences.

[12] More generally, as this example hints, we need to take centred worlds to consist (at least) of triples (x,i,t), where t is a time. We ignore this additional complexity for simplicity of exposition, since the time index plays no role in this paper.

being Makena and eating meat and *being Laurence and eating fish*. To describe preferences of this kind, we enrich the standard representation, and take preferences instead to be described by a binary relation on *centred* worlds. For Makena to have an (extended) preference for being Makena and eating meat as opposed to being Laurence and eating fish, in this framework, is simply for Makena's preferences to rank the centred world where the centre is Makena and everything is as it is except that Makena eats meat over the centred world where the centre is Laurence and everything is as it is except that Laurence eats fish.[13]

What is the relationship between 'ordinary' and 'extended' preferences in this model? The main point is the one we made earlier: there is only one notion of preference represented; 'ordinary' and 'extended' simply pick out different subspecies of those preferences. More precisely, as in the case of belief, we can think of extended preferences (the preferences that are defined over centred worlds, roughly analogous to *de se* beliefs) as being the fundamental phenomenon and 'ordinary' preferences (defined over worlds, roughly analogous to belief in propositions) as arising from them. Although there is a rough similarity between the case of belief and that of preference, the details in the preference case are somewhat disanalogous to the case of belief. In the case of belief, we thought of belief-contents as most fundamentally being sets of centred worlds, and 'ordinary' (non-self-locating) beliefs as arising in the special case in which the belief fails to distinguish between any two centred worlds that have the same base but different centres. In the case of preference, it is more plausible that 'ordinary' preferences are not *indifferent* to who is the centre of a given world, but rather that 'ordinary' preferences hold the centre fixed. For example, when we say that Makena prefers eating meat to eating fish, what we usually mean is that Makena prefers being herself and eating meat (and everything else being as it in fact is) to being herself and eating fish (while everything else is as it in fact is). In the formalism, she ordinary-prefers eating meat to eating fish just in case her (extended) preferences rank the centred world (Makena eats meat but everything else is as it in fact is, Makena) above the centred world (Makena eats fish but everything else is as it in fact is, Makena); to hold this preference, she certainly need not be indifferent between e.g. (the actual world, Makena) and (the actual world, Laurence). Ordinary preferences are thus just a special case of extended preferences, in an analogous (though not isomorphic) way to the way in which ordinary (i.e.

[13] As usual we take preferences to be most fundamentally relations on the (centred) worlds themselves, not on sets of these (centred) worlds. Typically in English people's preferences are described using sentences such as 'Makena prefers eating fish to eating meat'. Since this kind of expression appears to describe a relation between two *properties*, the surface grammar might seem to suggest that preferences should be taken to be relations between sets of centred worlds. We propose the following toy analysis of the relationship between the surface grammar and our model. When people utter sentences such as 'Makena prefers eating meat to eating fish', the qualification 'other things being equal' is automatically supplied, where this qualification is interpreted by a contextually supplied mapping from pairs of properties A, B to a symmetric relation $R \subseteq A \times B$. Thus 'Makena prefers eating meat to eating fish' is true just in case for all $\alpha \in A$ and $\beta \in B$, if $\beta R \alpha$ then $\alpha \succeq_M \beta$, where A is the set of centred worlds where the centre eats meat, and B is the set of worlds where the centre eats fish. This is not intended as a general theory of preference ascriptions, but simply as one way of relating the English expressions to the formalism we employ. In the sequel, we will often move without comment from an English statement about an agent's preferences to the claim that her preference-relation has a particular form. Given the complexities just mentioned, more would have to be said to justify this kind of inference, but we will assume in what follows that the cases could be set up in more detail to ensure the appropriate relationship between the English claims and the formal representation.

propositional) beliefs are a special case of the enriched notion of *de se* belief described in the centred-worlds framework.[14]

From the point of view of the centred-worlds framework, we can give a simple diagnosis of the source of the putative problems with extended preferences. The only reason that there appeared to be any conceptual problem with extended preferences in the first place was because we temporarily forgot the possibility of appealing to properties, as opposed to propositions: it was when we began casting about for *un*centred propositions (such as the impossible uncentred proposition that Makena is Laurence and eats fish) that our problems began.

4.2. A Simple Formalism for Extended Preferences

In the remainder of the paper, it will often be convenient to have a simple concrete model of extended preferences; we introduce this (semi-)formally here. This formalism has a number of further assumptions over and above the ones built into the centred world framework by itself, but as far as we are aware nothing in the sequel depends essentially either on the details of the centred worlds framework of the previous subsection, or on the extra assumptions to be introduced in this one.

Let W be a set of *uncentred alternatives* (possible worlds) and N a set of individuals (possible centres). We assume for simplicity that both of these sets are finite. An element of W specifies all features of the world, including what every individual's preferences are; it simply does not specify which individual is the centre. We hold fixed the set of individuals over all of our alternatives; that is, we assume for simplicity that every person in N—everyone we are considering—exists in every world. The set of *extended alternatives* can then be identified with the Cartesian product $X := W \times N$ (as a mnemonic, think of X as extended). In our toy example, this is an eight-element set, containing an item corresponding to 'being Makena/Laurence, while Makena eats meat/fish and Laurence eats meat/fish' for each way of resolving the three two-way choice points in this clause.

As we have said, each uncentred world specifies the preferences of every agent over *extended alternatives*. Formally, this aspect of the models can be made explicit using a function $E : N \to \mathcal{P}(X \times X)^W$, which associates each individual to a second function. That second function in turn maps each (uncentred) world to the agent's preference relation at that world over extended alternatives (a set of ordered pairs each of which has its elements

[14] One might be concerned that the framework we have introduced allows for distinctions which it shouldn't allow. It is natural to think that *who one is,* whether one is Makena or some other character Mary, does not on its own affect one's well-being. Various qualitative aspects of a person's life, which may be associated with who one is, obviously affect well-being, but the mere fact of being Makena as opposed to Mary plausibly does not. And yet our framework allows for the possibility that there are differences in well-being between alternatives which differ merely in who the centre is. Our framework can be easily restricted by stipulation to prevent it from drawing this distinction, but still one might prefer a framework which is built from the outset so that it does not draw these distinctions. For example, instead of thinking in terms of worlds in the first instance, we could have taken the basic relata of preferences to be very specific properties (where properties are no longer identified with sets of centred worlds, but taken formally to be simply 'points'). That would be in line with the way in which Harsanyi often speaks, for example, where the relata of preferences are objects such as 'eating meat'. We could then distinguish between 'ordinary' alternatives—very specific properties which describe everything except agents' preferences—and 'extended alternatives', elements of the cross-product of 'ordinary' alternatives and preference relations over ordinary alternatives. In the remainder of the paper we will continue to speak in terms of the framework in the main text, but we think it is also worth considering the alternative; indeed one of us prefers it. In any event, none of the arguments in this paper turns on differences between these two models.

drawn from X).[15] We will assume that this second function is in fact constant: that is, given an individual, the individual has the same (extended) preferences at every world we will consider. With this simplifying assumption made, the complex function $E : N \to \mathcal{P}(X \times X)^W$ now reduces to $E : N \to \mathcal{P}(X \times X)$, a function which takes individuals to a binary relation on extended alternatives. We will usually write the value of this function as $\succeq_i$, for the extended preferences of i; the *overall* betterness relation will be denoted $\succeq$. We will use the standard notation $\succ_i$ and $\succ$ to represent the asymmetric parts of the relations derived from $\succeq_i$ and $\succeq$ respectively: $x \succ_i y$ just in case $x \succeq_i y$ and $\neg y \succeq_i x$ (and similarly for $\succ$). If $x \succ_i y$ (resp. $x \succ y$) we will say that x is *strictly* preferred to (resp. better than) y; if $x \succeq_i y$ (resp. $x \succeq y$) we say that x is *weakly* preferred to (resp. better than) y. Finally, we will use Latin letters late in the alphabet (e.g. w, x, y) for variables over uncentered worlds (elements of W), and Greek letters early in the alphabet (e.g. α, β, γ) for elements of X.

The simplifying assumption just introduced—that each individual's preferences are constant on worlds—greatly simplifies the mathematics. It is also a conceptual simplification, to avoid a problem that as far as we are aware has not been solved. Preference-satisfaction theory in general faces a problem about fundamental changes in preference. People do seem to undergo fundamental changes of preference in the course of their lives, so that whether a given situation satisfies their preferences seems to depend on whether we consider their preferences now or at some other time. The problem is that arguably the notion of well-being which is most important for the purposes of moral theory is that of *lifetime* well-being, that is, a measure of how well a person's life goes for her *as a whole*. But it is unclear, in the face of fundamental changes of preference, how we are to understand this: should one point of time be taken as the point from which the value of our whole lives is assessed? This seems an unappealing way of assessing lifetime well-being, but other proposals we are aware of are equally so. In advance of finding a solution to this problem for preference-satisfaction theory more generally, it is reasonable to abstract from the possibility of preference-change in the case of extended preferences, as well. Our formal simplification does just that.[16]

---

[15] An anonymous referee raised the question of whether there is some vicious circularity here, since extended preferences are defined over alternatives that already include those preferences in their descriptions. But the 'circularity' is unproblematic. The point can be illustrated by analogy to a more familiar class of formal model, models of belief in the tradition of Hintikka and Kripke. These models are structures consisting of a set of possible worlds, together with a binary accessibility relation on them. The worlds are interpreted as specifying everything which happens in them, *including what the agents believe*. The accessibility relation, like our function E, might thus seem to generate a kind of circularity, specifying beliefs over worlds which already specify what the agents in them believe. But it does not. The accessibility relation (like our function E) does not 'add' some specification of beliefs (preferences): it simply makes one aspect of what is true at these worlds explicit in the model.

[16] A quite different way of justifying the assumption is as follows. The EP theorist might take individuals to be 'preference-bound': if they were to change preferences, the theory should no longer consider them the same individual. This move is unavailable in the impoverished setting of ordinary preference satisfaction theory, because it would mean that our preferences now don't concern the individuals we might become. But in extended preference theory, an individual has preferences over worlds with *different* centres, and those different centres might well be the individual him- or herself after he or she has undergone some fundamental change of preference. Thus in the extended preferences setting, we can assimilate preference-change to variation in the population N of individuals. N is a set of 'preference-slices' of individuals: equivalence classes of time-slices of individuals generated by the equivalence relation obtained by intersecting the equivalence relations 'being the same person' and 'having identical preferences'. It is worth noting that if extended preference theory were to succeed in providing interpersonal well-being comparisons, then this could lead to substantial progress on the question described in the main text, of how to make sense of fundamental preference-change within preference satisfaction theory. For given 'interpersonal' well-being comparisons, we would also know how to compare the well-being of distinct 'preference-slices'. Thus we could choose any one of a number of ways of aggregating the well-being of an individual's preference-slices over his or her whole life.

At least at the outset, we will not be taking for granted any formal constraints on the set of rational preferences. But as usual, *if we did* have sufficient formal constraints (the usual von Neumann-Morgenstern axioms) on preferences over *lotteries* on *X*, we could recover *utility functions on* (rather than merely *orderings of*) the set of extended alternatives. The possibility of using these extra constraints to generate utility functions is important to have in the background; where possible, though, we will set it aside to simplify the exposition.

5. The Principle of Full Coincidence

5.1. Content of the Principle of Full Coincidence

Given that we can make sense of individuals' extended preferences, we now face the question of how (interpersonally comparable) well-being is determined by the facts about individuals' extended preferences. In sections 2–4, we bracketed the focus on self-interested, rational and fully-informed preferences. But now these qualifications become important: the question is not simply how individuals' *actual* extended preferences determine well-being, but how their self-interested, rational, fully informed extended preferences do. For the remainder of the paper, our primary focus will be on self-interested, rational, fully-informed preferences.

The literature on extended preferences began with a very simple, elegant theory of the relationship between preferences and well-being. The founders of the extended preferences program endorsed the following principle:

> **(PFC) The Principle of Full Coincidence:** All individuals have the same extended preferences as one another.

If it were true, this principle would make the question of how individuals' extended preferences combine to determine facts about well-being extremely simple. For given PFC, the extended preference theorist could simply postulate that one extended alternative is better than another (possibly different-centred) extended alternative if and only if the universally shared extended-preference relation ranks the first above the second. If, on the other hand, different individuals in general have different extended preferences (that is, if the rational and fully informed counterparts of their actual extended preferences can differ), then there is no such universally shared extended-preference relation, so this move is unavailable. In that case, a more complicated account is required, invoking an 'aggregation rule'. (We return to this briefly at the end of section 5.2.)

As we have said, the founding fathers of extended-preference theory, Harsanyi and Arrow, took it that PFC was true (Harsanyi (1977, section 4.4); Arrow (1978, 225)). But is this principle plausible? *Pace* Harsanyi and Arrow, we find that the initial prognosis is unpromising. The remainder of this section rejects the positive arguments that have been offered to date for PFC. We then (in sections 6–7) turn to a different principle, the Principle of Acceptance, which is entailed by but weaker than PFC. Sections 8–9 will investigate a new line of thought that could provide grounds supporting both principles; we will argue there, however, that this brings with it very high costs.

5.2. Harsanyi's Argument for the Principle of Full Coincidence

Harsanyi himself (1977, 58–59) notoriously argued for PFC by appealing to a form of psychological determinism: granted that distinct individuals have distinct ordinary preferences, Harsanyi argued that such differences are always traceable to causal factors. One's eventual preferences, on this account, are in principle predictable by a Laplacean Demon who possesses full information of all details of one's genetic inheritance and environmental history; there are 'psychological laws' mapping such inputs to the preferences they determine. If two individuals appear to have distinct extended preferences, therefore—if, say, one prefers an extended alternative consisting of a life of adventure and a taste for adventure to a life of quiet contemplation and a taste for quiet contemplation, while the other has the reverse preference—this is, according to Harsanyi, only because we have failed to describe the objects of preferences at a sufficiently fine-grained level; we have not included the causal history leading to those preferences. Both individuals, he insists, would prefer to have the complex consisting of the life of adventure, the taste for adventure *and to have inherited the causal factors generating those tastes* than the analogous 'quiet life' complex (or perhaps vice versa). (A similar argument is proposed by Kolm (2002, 165–167).)[17]

This last step, though, is a mistake. Whether or not psychological determinism is true, and granting the point that myriad factors have functioned as causal inputs to the preferences (ordinary and extended) that any given individual now has, there is no reason to think that including those causal factors in the description of the extended alternatives will wash out the differences that do now exist between individuals who have in fact been subject to different causal factors.

To borrow an example from Broome's illuminating critique of Harsanyi's argument (Broome (1998)): various causal factors have been responsible for the fact that Olga places a higher value on intellectual achievement and understanding than on earnings, whereas Neil places a higher value on high earnings than on intellectual achievement, so that their respective preference orderings rank the life of an academic and that of (say) a big-city banker in opposite ways. But those same causal factors have further been responsible for the fact that Olga also prefers the life of an academic *who values academia and who has been subject to the causal factors generating a preference for academia* to that of a city banker *who values high earnings and who has been subject to the causal factors generating a preference for high earnings*, while Neil has the opposite preference between these finer-grained alternatives. Olga simply (as things have in fact turned out) does not much value even the complex of having a high salary *and valuing a high salary*; Neil (as things have in fact turned out) simply does not much value even the complex of leading an academic life *and valuing an academic life*. That is, even when the causal influences on (or even determinants of) preferences are included among the objects of preference, the causal factors they have

[17] An anonymous referee pressed the point that while Harsanyi does often write explicitly of preferences, as in our reconstruction here, in other places he seems to be thinking more in terms of a mental-state theory of well-being (framing his discussion in terms of 'satisfaction', and using that term in a way that is more suggestive of felt experiences of satisfaction than of preference-satisfaction). This includes some of Harsanyi's discussions of the influences of causal history: see, in particular, Harsanyi (1982). We agree with this point, as a matter of Harsanyi scholarship. We focus on the 'preferences' version of Harsanyi's argument simply because only that version is relevant to the project discussed in this paper.

in fact been subject to retain their distinct role as causal influences on the preferences they in fact have, and these preferences remain different for different individuals. Nor, barring appeal to a substantive non-preference-based theory of well-being, is there any obvious way for an appeal to 'full information and rationality' to erase such differences. If the relevant information and rationality conditions are as we have been conceiving them to this point (but cf. section 9), PFC is false, even as applied to self-interested, fully informed and rational versions of ordinary people's preferences.

What, then, is to be done? Given that PFC fails, the EP theorist cannot simply identify the overall well-being ordering with the shared preference ordering of all individuals; she must instead have some way of aggregating extended preferences into a single well-being ordering. This is easier said than done, for reasons related to Arrow's celebrated impossibility theorem (Arrow (1963)); the challenge is discussed in detail by Adler (2016b) and by ourselves in other work (Greaves and Lederman (2016)). For the remainder of this paper, however, we will suppose that a reasonable aggregation rule can be found, and develop a different kind of problem for the EP program. To that end, we turn now to a second principle, which is weaker than PFC, and which we will argue is of fundamental importance to the program.

6. The Principle of Acceptance

The extended preferences program is intended as way of rescuing the preference-satisfaction theory from the objection that it cannot make sense of interpersonal comparisons of well-being. But it will succeed in rescuing the preference-satisfaction theory only if it makes sense of interpersonal well-being comparisons in a way that is consistent with the preference-satisfaction theory itself.

The requirement that the EP theory be consistent with ordinary preference satisfaction theory leads to important requirements on a satisfactory extended preference theory. Ordinary preference satisfaction theory is committed to the following principle:

> **(OPS) Ordinary Preference Satisfaction:** For all $x, y \in W$, and all $i \in N$, $(x, i) \succeq_i (y, i)$ if and only if $(x, i) \succeq (y, i)$.

This principle says that if i has a preference between any pair of alternatives which have i as their common centre—that is, i has an *ordinary* preference between possible worlds x and y—then that preference is respected by the overall betterness ordering. It further says that the overall betterness ordering does not 'invent' any betterness facts for same-centred alternatives that are not derived from the ordinary preferences of the centre in question. To see why ordinary preference-satisfaction theory is committed to this principle, consider the example of Makena and Laurence once again. Given OPS, if (and only if) Makena (ordinary-)prefers meat to fish, then Makena eating meat will be better off than Makena eating fish—just as the ordinary preference-satisfaction theorist says. But if the EP theory violates OPS, then the resulting theory will countenance either cases in which Makena (ordinary-)prefers eating fish to eating meat, but eating fish is not better for Makena than eating meat, or cases where eating fish is better for Makena in spite of the fact that she does not prefer it. Each of these possibilities conflicts with the guiding

idea of preference-satisfaction theory: that something is better for an individual precisely to the degree to which it satisfies her own ordinary preferences (or at least, the fully informed and rational versions of them). So, if the EP theory is to save the preference-satisfaction theory from the problem of interpersonal well-being comparisons, then it must satisfy OPS.

From here, though, we can argue that if the EP program is to succeed, it must impose a further rationality constraint on individuals' extended preferences. The most plausible candidate for such a constraint is a principle will call 'the principle of acceptance'. This principle says that if an individual i has a preference over two i-centred alternatives, then all rational preferences agree with i's preferences over these alternatives. That is:

> **(PA) The Principle of Acceptance:** For all $w, z \in W$, and all $i, j \in N$, if $(w, i) \succeq_i (z, i)$ then $(w, i) \succeq_j (z, i)$.[18]

We will now present an argument that the EP theorist should accept this constraint on rational preferences.

Consider a population which finds itself in the following situation. There is some pair of possible worlds w and z in W such that two individuals i and j both strictly (ordinary-)prefer w to z (that is, $(w, i) \succ_i (z, i)$ and $(w, j) \succ_j (z, j)$). Thus, by OPS, we must have $(w, i) \succ (z, i)$ and $(w, j) \succ (z, j)$. But now suppose further that every $k \in N$ (including i and j themselves) strictly prefers being i in situation z to being j in situation w, *and* strictly prefers being j in situation z to being i in situation w (that is, $(z, i) \succ_k (w, j)$ and $(z, j) \succ_k (w, i)$). If i and j have *transitive strict* preferences, it follows from this specification that the extended preferences of i and j, respectively, disagree with the ordinary preferences of j and i respectively: $(z, j) \succ_i (w, j)$, while $(z, i) \succ_j (w, i)$. This disagreement is perfectly consistent; indeed, it seems that such disagreements are observed in the world as it is, just as in the case of the academic Olga and the banker Neil described earlier. Further, nothing we have said about 'full information and rationality' rules out the whole pattern of (idealised) extended preferences postulated in this paragraph.

Trouble, however, is in the wings. For the following principle seems central to the ideas which motivate the EP program:

---

[18] Some authors instead discuss a stronger principle, which replaces the conditional here with a biconditional:

> **(SPA) (Strong) Principle of Acceptance:** For all $w, z \in W$, and all $i, j \in N$, $(w, i) \succeq_i (z, i)$ if and only if $(w, i) \succeq_j (z, i)$.

Harsanyi's own 'principle of acceptance' (1977, 'Axiom 2' on page 54) is our PA, rather than SPA (*pace* many citations of him, e.g. Adler (2016a, 482)). But Sen's 'identity axiom' (1970, 156) is, in our terminology, SPA. Pattanaik's discussion (1968, 1159) is somewhat ambiguous between the two, but arguably suggests SPA more than PA.

As far as we can tell, none of these authors had any strong reason for preferring one principle to the other. There are two differences between the principles: first, if i strictly prefers one i-centred alternative to another, then our PA allows (where SPA does not) other rational agents to be indifferent between the two; second, if i considers two i-centred alternatives incomparable, our PA allows (while SPA does not) other rational agents to have weak or even strict preferences over those alternatives. We will continue to use 'PA' for the weaker principle (and we will discuss this principle to the exclusion of SPA in the main text). We will be arguing that even this weaker principle is unmotivated and indeed false on one construal of the EP program; these arguments apply equally to the stronger principle.

Strict Unanimity: If for some centred worlds α, $\beta \in X$, every $k \in N$ is such that $\alpha \succ_k \beta$, then $\alpha \succ \beta$.

The preference-satisfaction theory aims to make sense of the idea that facts about well-being arise from facts about preferences; while a theory could violate Strict Unanimity and nevertheless agree (technically) that well-being facts are *determined by* preferences, the spirit of the theory seems further to require that the relationship between preferences and well-being facts have positive valence, in the sense that whenever unanimity happens to obtain, betterness facts should respect unanimous preferences.

It follows from Strict Unanimity that in the case under discussion, $(z, i) \succ (w, j)$, while $(z, j) \succ (w, i)$. But, as we saw above, it follows from OPS that $(w, i) \succ (z, i)$ and $(w, j) \succ (z, j)$. Putting these facts together, we see that the overall well-being 'ordering' is forced under these conditions to exhibit an intolerable violation of transitivity. In fact it is cyclic: $(w, i) \succ (z, i)$, $(z, i) \succ (w, j)$, $(w, j) \succ (z, j)$ even though $(z, j) \succ (w, i)$.

Three assumptions—OPS, Strict Unanimity, and the possibility that all individuals' preferences could be chosen freely—give rise to a disastrous result. The most obvious way to resist this argument is to deny that the aggregation rule must be well-behaved no matter how individuals' preferences are specified. The argument uses two assumptions about the domain of preference profiles on which the aggregation rule has to be well-behaved, in the sense of respecting both OPS and Strict Unanimity on this domain. First, we had to allow a situation in which every member of the group agreed in their extended preferences on two pairs of alternatives, that is, for all k, $(z, i) \succ_k (w, j)$ and $(z, j) \succ_k (w, i)$. This assumption seems unproblematic: surely, the aggregation rule must be well-behaved on a domain that allows for consensus of this kind. Although we have argued that PFC is not in general true, that does not mean that rational preferences *never* agree. There seems little promise to denying that it is possible that all extended preferences could agree as they do in the case above.

The second assumption, however, might seem more controversial. In the situation above, if i and j's strict extended preferences are transitive, then they each disagree with the other's *ordinary* preferences—for example $(z, j) \succ_i (w, j)$ even though $(w, j) \succ_j (z, j)$. But one might doubt whether an agents' rational extended preferences could disagree with others' ordinary preferences in this way, and accordingly attempt to respond to the argument by citing a general constraint on rational preferences that rules out such disagreement. PA would clearly be sufficient to eliminate the problem: given PA the situation just considered—which depended on i and j's disagreement over pairs of same-centre alternatives—cannot arise. But it is also hard to think of principles interestingly different from PA which would be both conceptually well-motivated and sufficiently general to rule out all cases analogous to the one just described.[19] If the EP theorist hopes to preserve ordinary preference-satisfaction theory, there is significant pressure from this argument to accept PA.

But to say that the EP theorist should hope that PA is true is not to say that PA is true. In the following three sections, we will present a trilemma for the EP program. In section 7, we argue that PA is false on one natural understanding of the conditions of 'rationality and full information'. This is the first horn of the trilemma: if the EP theorist

[19] One related principle requires agreement on i-centred alternatives if i has a strict preference between them, but doesn't require indifference if i is indifferent. Our argument in section 7 will also show that this principle (which we might call the 'Strict Principle of Acceptance') is false on the usual construal of preferences.

sticks to these notions of rationality and full information, she must give up on PA, so the EP theory of well-being will fail to coincide with the verdicts about well-being given by preference-satisfaction theory.

In sections 8 and 9, we present the second and third horns horn of the trilemma. These arise if the EP theorist appeals, respectively, to weaker or stronger further idealisations, in the form of some close connection between the subject's idealised preferences and her beliefs about betterness. Such connections enable the EP theorist to argue for PA (and, in the 'strong' case, also PFC), but, we will argue, at too high a cost.

7. Failed Arguments for the Principle of Acceptance

7.1. Harsanyi and Sen on PA

Before turning to our own argument *against* the principle of acceptance, we consider, and reject, others' arguments in favor of it.[20] Harsanyi himself writes that the principle of acceptance

> is, of course, merely the familiar *principle of consumers' sovereignty*, often discussed in the literature of welfare economics: The interests of each individual must be defined fundamentally in terms of his own personal preferences and not in terms of what somebody else thinks is 'good for him'. (Harsanyi, 1977, 52; emphasis in original)

In the present context, however, it is unclear how this is supposed to amount to an argument for the *truth* of PA. We have already noted that given the EP program's intention to define well-being in terms of extended preferences, PA *had better turn out to be* true, *on pain of* the EP theory failing to be consistent with ordinary preference-satisfaction theory. Insofar as this is Harsanyi's observation, then, of course we agree; but, equally clearly, this is (wishful thinking aside) no argument for the claim that PA *is* true. Similar remarks apply to Sen's assertion that PA (in Sen's nomenclature, 'the identity axiom') 'can be justified on ethical grounds, as an important part of the exercise of extended sympathy' (Sen, 1970, 156).

Picking up on the language of 'consumers' sovereignty', we might alternatively take Harsanyi to be appealing to a principle of anti-paternalism. But it is equally unclear how this could work: it is perfectly consistent, for instance, for Quinn fully to respect the appropriateness of Petra's determining her restaurant choices on the basis of her own personal (ordinary) preferences, while simultaneously being such that for his own part, *Quinn* would rather be Petra and (have Petra's ordinary preferences) and eat fish than be Petra (and have Petra's ordinary preferences) and eat meat.

7.2. Pattanaik's Argument

A second attempt at an argument for PA is offered by Pattanaik (1968, 1159):

> The introspective utility of individual 2 from [the extended alternative $(x, 1)$] must be the same as that of individual 1 since, to experience the [extended] alternative $[(x, 1)]$ at all, individual 2 has to transform himself through imagination into individual 1.

[20] Recall (n. 18) that some of the authors we discuss in this subsection and the next in fact used the name 'principle of acceptance' for a slightly stronger principle than ours (SPA).

But this argument is, if anything, even more obscure than Harsanyi's. In particular, it is unclear what Pattanaik means by 'introspective utility'. For the notion of 'utility' to be the one that is relevant in the present context, it must be 'utility' *in the sense of representation of preferences*. But in that case, the passage quoted above can be translated, without significant loss of sense, as

> Individual 2 must prefer the extended alternative (x, 1) to the extended alternative (y, 1) if and only if individual 1 does since, to experience the extended alternatives (x, 1) and (y, 1) at all, individual 2 has to transform himself through imagination into individual 1

—and it is very unclear how this is supposed to follow. Individual 2, we grant, 'has to transform himself through imagination into individual 1' in the sense that he must exercise empathy, in order to 'experience person-1-centred extended alternatives' in the sense of adequately grasping what those extended alternatives in question are like; the latter in turn is required if 2's extended preferences regarding 1-centred alternatives are to qualify as being suitably informed. But it does not follow that individual 2's preferences regarding 1-centred extended alternatives must coincide with those of individual 1: here as elsewhere, for all that has been said so far, 1 and 2 might perfectly well succeed in imagining the same things as one another, but react (in terms of their preference orderings) very differently to those same things.

7.3. Against the Principle of Acceptance

It is not just that we know of no good arguments for PA on the usual understanding of rational, fully-informed preferences; there is a good argument against PA on this understanding of preferences. In fact, the same considerations that show that PFC is false, if preferences are not idealised in some way beyond those we have so far considered, *also* suffice to show that under the same conditions, PA is false. Let us return to the earlier example: Olga prefers the life of an academic, while Neil prefers the life of a banker. We noted above, when discussing PFC, that this difference in preferences can remain even when *having the preferences of* an academic (respectively, of a banker) are included as part of 'the life of an academic' (respectively, a banker)—that is, even when the topic is extended rather than ordinary preferences. We noted also that this difference is naturally understood as a difference in values: Olga, for example, extended-prefers being an academic with an academic's preferences over being a banker with a banker's preferences because, for example, Olga values the achievements of an academic's life more than Olga values those of a banker's. But precisely the same line of thought then suggests that Olga, with the preferences and values she actually has, might also prefer living the life of an academic ('objectively speaking', that is, holding down an academic career and so forth) *while holding the ordinary preferences of a banker* over living the life of a banker with those same ordinary-preferences: Olga might (actually) have sufficiently little regard for bankers' ordinary preferences that *as she actually is* she prefers that, on the condition that those ordinary preferences were hers, they be frustrated. (Of course, having frustrated preferences is likely to count in *some* sense as a negative; but this sense might be fully captured by her (different) extended-preference for the life of an academic while holding an academic's preferences over the life of an academic while holding a banker's preferences.) We conclude that like PFC, PA is false given the relatively weak notions of full

information and rationality we have assumed so far. While—as before—the example is most vivid when we think in terms of actual preferences, it is plausible that this pattern of preferences could persist for rational and fully informed versions of Olga and Neil as well, so long as the conditions of idealisation don't involve substantive changes of fundamental values. Furthermore, while we have presented the case as an argument against PA, the argument is in fact more general. Olga and Neil instantiate exactly the pattern of preferences exhibited by i and j in the argument from section 6. The example thus presents a challenge not only to PA but plausibly to any way of avoiding that argument by denying that rational preferences can exhibit disagreements on same-centred alternatives of the kind described there.

This is the first horn of our trilemma: accept the relatively weak notions of full information and rationality that we have assumed so far, and give up on PA (or any related principle designed to block the earlier argument by denying the possibility of that pattern of rational preferences). Since we have argued that the EP program is consistent with the ordinary preference-satisfaction theory of well-being only if PA (or something close to it) is true, and since the aim of the EP program was to recover interpersonal comparisons *within a preference-satisfaction theory of well-being*, this horn leads to the failure of that program. Sections 8 and 9 explore the second and third horns.

8. Full Information and Rationality I: The Weak Preferences-Betterness Principle

So far, we have been working with relatively minimal, weak notions of rationality and full information. But perhaps they have been *too* weak. We argued earlier, and we still maintain, that it would be against the spirit of preference-satisfaction theory to invoke any notion of betterness *that is itself independent of preferences*, and then a 'substantive' notion of rationality that simply counts preferences as irrational if they fail to match these prior betterness facts. That was the thought behind the insistence on a 'purely structural' as opposed to a 'substantive' notion of rationality. However, no violence need be done to the spirit of preference-satisfaction theory if we merely require consistency between an individual's preferences and *her own beliefs about* betterness, if and where she has any such beliefs; while not operating solely within the domain of preferences, this broader type of consistency condition does seem to be 'purely structural' in the relevant sense. And, so far, while we have countenanced the idea that the EP theory might supply facts about betterness, we have not invoked any conditions to the effect that a fully informed individual knows these facts, and that a fully rational individual conforms her preferences to known betterness facts. The present section and the next investigate the extent to which developing this line of thought might lead to a more plausible version of the EP program. There are two cases to consider; the remainder of this section examines the first case.

8.1. The Weak Preferences-Betterness Connection

First, then, consider the following constraint on the relationship between the extended preferences of a fully informed and rational individual i and betterness-for-the-individual facts:

> **(WPB) Weak preferences-betterness connection:** For all extended alternatives α,β and all individuals i, if $\alpha \succeq \beta$, then $\alpha \succeq_i \beta$.

This *weak* preferences-betterness principle allows that a fully informed and rational individual might prefer α to β while α and β are objectively incomparable.[21] (In section 9, below, we consider the strengthening that rules this (too) out, and requires preference relations fully to coincide with the betterness relation.)

Why accept any such principle? *One* type of rationale would be based on acceptance of a *non*-preference-based, substantive theory of the good, together with evaluatively loaded notions of full information and of rationality. If, for example, it is the case (independently of and prior to any facts about anyone's preferences) that other things equal, a life with more education is objectively better for the individual than one with less, then a fully informed individual would be aware of this evaluative fact, hence (in particular) would *believe* that a life with more education is better for the individual than one with less. But it is presumably irrational to believe that α is better for its centre than β is for its centre while failing to prefer α to β, so such a rational and fully-informed agent would prefer α to β.

Rationales of precisely *this* type are of course unavailable to a preference-satisfaction theorist, since these rationales appeal to accounts of objective betterness-for-the-individual that are inconsistent with preference-satisfaction theory. But an account that substitutes preference-satisfaction theory itself in the above line of reasoning might yield relevantly similar conclusions. The idea, then, is that even if preferences are (somehow) constitutive of or serve as the ground for betterness facts, it is *still* the case that a fully informed individual will be aware of the betterness facts: if she is fully informed, then she is aware both of everyone's preferences and of the relation between preferences and betterness, and able to compute which betterness facts are implied by the given profile of preferences. Further, again, presumably it is irrational to believe that α is weakly better than β while failing to weakly prefer α to β. Therefore, the argument concludes, even a preference-satisfaction theorist should accept WPB.

8.2. A New Argument for the Principle of Acceptance

But if so, then the preference-satisfaction theorist—contra the negative prognosis of section 7—does after all have a valid argument for PA:

> **(OPS):** For all possible worlds $x, y \in W$, and all $i \in N$, $(x, i) \succeq_i (y, i)$ if and only if $(x, i) \succeq (y, i)$.
>
> **(WPB):** For all extended alternatives $\alpha, \beta \in X$ and all individuals $i \in N$, if $\alpha \succeq \beta$, then $\alpha \succeq_i \beta$.

Therefore,

> **(PA):** For all individuals $i, j \in N$ and all possible worlds $x, y \in W$, if $(x, j) \succeq_j (y, j)$, then $(x, j) \succeq_i (y, j)$. (From OPS, WPB)

---

[21] It also allows that if α is strictly better than β, a rational and fully informed individual might nevertheless be indifferent between them. Let us call the principle which requires that strict betterness entail strict preference the 'Strict preferences-betterness connection', by analogy to the Strict Principle of Acceptance (see n. 19). (This is to be distinguished from the 'strong' preferences-betterness connection, which will be introduced later on.) The argument scheme in section 8.2 which uses WPB and OPS to derive PA could be used with the Strict preferences-betterness connection instead of (or in addition to) WPB to derive the Strict Principle of Acceptance. But adding the Strict principles to the theory would not avoid the problems we present for it, so to simplify the exposition we will not mention them again in the main text.

The availability of this argument is good news for the EP theorist, since, as we argued earlier, something like PA seems to be required if extended-preference theory is to avoid acyclicity, while holding on to OPS. Unfortunately, there is also bad news.

The bad news is that WPB faces a dilemma. Either the theory of rationality which motivates it allows the widespread disagreement observed in people's actual preferences to persist in rational and fully-informed preferences, or it does not. We will argue in the present section that in the first case—if it allows widespread disagreement in rational and fully informed preferences—it leads to massive incomparability in interpersonal well-being. We will argue in the next section, that in the second case—if it prohibits such disagreement—it (along with theories like it) leads to a problematic form of holism and ungroundedness in the relationship between preferences and well-being.

First, however, let us illustrate what is at stake in the question of whether widespread disagreement in rational and fully-informed preferences persists despite WPB. Recall the example in which Olga and Neil *actually* have opposite strict extended preferences for the life of an academic as compared to that of a banker. There are two ways this example could turn out, if WPB is true. First, it could be that both Olga's and Neil's preferences are rational (and that the extended alternatives in question are objectively incomparable). Alternatively, it could be that at least one is irrational (and possibly both), since they hold a strict preference which disagrees with the betterness facts. Each of these situations is consistent with WPB, and it is not entirely clear which of them is more likely in a setting in which WPB is true. In the remainder of this section, we focus on the implications of the first possibility, assuming for the sake of argument that this is the one that obtains in the present setting. In Section 9 we explore a different setting, in which a phenomenon much like the second possibility occurs.

If widespread disagreement in fully informed and rational preferences persists despite WPB, then the result, given WPB, must be massive interpersonal incomparability. The reason for this is that the only aggregation rules which are consistent with WPB deem two alternatives incomparable whenever there is disagreement in rational, fully informed preferences between them. To see this, consider arbitrary extended alternatives α, β, and suppose that α and β are comparable. Then either α is weakly better than β, or vice versa (or both); without loss of generality, assume the former. Then, by WPB, *all* fully informed and rational preferences must weakly prefer α to β. But in that case no two individuals have opposite self-interested, fully-informed and rational strict preferences regarding α and β. Contraposing: if two individuals do have opposite strict preferences regarding α and β, then α and β are incomparable. But nothing we have said so far prevents rational strict preferences from varying on many pairs of elements drawn from the set of extended alternatives.

Given that Olga and Neil's rational and fully informed preferences disagree in the way their actual preferences do, WPB implies that their lives are incomparable. This result on its own is not obviously problematic: *this* example might well be a case of incomparability. But it is plausible that what is true of Olga's and Neil's lives in this example is also, on the present approach, true of just about *any* pair of different-centred extended alternatives. For whatever it was that permitted Olga's and Neil's rational, fully informed preferences to disagree on this pair of different-centred alternatives will plausibly allow many people's rational fully-informed preferences to disagree on a wide range of different-centred alternatives. More generally: since it (together with OPS) implies PA, WPB forces (and rationalises) a high degree of interpersonal agreement on extended-

preference rankings of *same*-centred alternatives, but there is still nothing to force interpersonal agreement on any given pair of *different*-centred alternatives. And if the 'constituency'—the set of individuals whose idealised extended preferences form the input to the rule which aggregates extended preferences—is anything like as large as the citizenship of a medium-sized country, there will almost always be at least one pair of individuals with opposite strict preferences regarding the given pair of extended alternatives. It only takes one person to be such that their rational, fully informed preferences regard education as a bane, for instance, for it to follow that a life with greater education is neither better, nor even equally as good as, a life that involves less education but which is otherwise the same. Similarly for material consumption, hedonic pleasure, achievement, health and so forth.

If the theory that postulates WPB allows for disagreement among ration and fully informed preferences in cases such as that of Olga an Neil, then, it seems to lead to massive incomparability on different-centred alternatives. This, then, is the second horn of the trilemma: accept WPB as a way of saving PA, but be led by WPB into an unacceptable degree of incomparability in the well-being of different-centred alternatives. Like the first horn, this second horn also amounts to the failure of the EP program, since the aim of that program was to *recover many positive interpersonal comparisons* within a preference-satisfaction approach to well-being.

The natural response for the EP theorist is to hope that the theory which includes WPB in fact rules out disagreement between the rational, fully informed preferences of individuals such as Olga and Neil. Thus, an EP theorist might be moved to endorse stronger rationality constraints on preferences, above and beyond WPB, which ensure that disagreements of the problematic kind just described do not arise. For example, she might endorse an obvious strengthening of WPB, converting the governing conditional of that principle into a biconditional. This move will initially seem to help, since it will facilitate an argument for PFC, and in the presence of that principle the problem of aggregation and the related problem of incomparability evaporate. But, we will argue in the next section, although the theory with PFC avoids these problems, it does not ultimately lead to a satisfactory form of EP theory.

9. Full Information and Rationality II: The Strong Preferences-Betterness Principle

9.1. A New Argument for the Principle of Full Coincidence

Recall that the *weak* preferences-betterness principle, WPB, allowed that a fully informed and rational individual might have a capricious strict preference for x over y when x and y are objectively incomparable. Matters look somewhat different for the EP program if we deny that this is possible, and instead accept the following stronger principle (again, for rational and fully informed preferences):

> **(SPB) Strong preferences-betterness connection:** For all extended alternatives α,β and all individuals i, $\alpha \succeq_i \beta$ iff $\alpha \succeq \beta$.[22]

[22] Note that this principle also rules out the possibility, discussed in n. 21, that a rational agent be indifferent between two alternatives if the betterness ordering deems one strictly better than the other.

PFC follows trivially from SPB: if every individual's extended preferences must match the betterness relation, then every individual must have the same extended preferences as every other individual. As we discussed, WPB was consistent with both Olga's and Neil's divergent preferences both being rational and fully-informed. SPB is not. For given SPB, at least one of Olga and Neil must be irrational. The principle entails that if Olga's life is in fact better, then Neil is irrational in preferring his own life, while if Neil's life is better than Olga's, Olga is irrational in preferring her own life, and if neither is better than the other, then *both* are irrational. Since SPB entails PFC, once SPB is in place, we can simply identify the overall well-being ordering with the unique rational preference relation.

This suggests the following version of the EP program. The *actual* extended preferences of (actual) individuals vary, but that is possible only because actual individuals are irrational or imperfectly informed. For each individual, there are facts about what her extended preferences *would* be under conditions of full information and rationality; the facts about which extended alternatives are better-for-the-individual than which others are grounded in these hypothetical extended preferences. Because a fully informed and rational individual (i) would *know* what everyone's idealised extended preferences are (including her own), (ii) would know that betterness facts are grounded in extended preferences, and what the relation between betterness and preferences is and (iii) would not have preferences that failed to match what she knows (and hence believes) to be the betterness facts, we have a guarantee, in advance of knowing what any individual's idealised extended preferences would be, that every individual's idealised extended preferences are the same as every other individual's idealised extended preferences. As a result, PFC holds, and the extended-preference theorist is after all able to make the simple move noted in section 5, viz. that of postulating that the betterness facts are given by the unanimously shared extended-preference relation, whatever it it may be.

9.2. Problems for the EP Theory with SPB

Unlike on the first horn of the trilemma, this version of the EP program produces something recognisable as a preference-satisfaction theory of well-being. Unlike on the second horn, we are now assuming that the idealisations which generate *rational and fully informed* preferences are sufficient for a high degree of convergence even on interpersonal matters, so that massive incomparability need not result. This third form of the EP theory, however, has some particularly bizarre further consequences that we think few preference satisfaction theorists will want to swallow. At least the first two of these problems appear to us to be quite general; modifications of them apply to any theory which includes WPB, on the condition that the theory rules out disagreement in the rational preferences of people such as Olga and Neil (the 'second possibility' mentioned earlier).

First, the notions of full information and rationality required by any theory which endorses even the weaker WPB lead to quite a dramatic divorce between the preferences it invokes and those typically called on in preference-satisfaction theory. When we introduced the idea of full information using the example of a train schedule, 'full information' was supposed to include knowledge that in a clear sense directly concerns, and is directly relevant to the assessment of, the objects of the preferences in question, taken by themselves. Similarly, 'rationality' consisted in a modest ironing-out of inconsistencies in the agent's beliefs and preferences concerning those objects taken by

themselves. By 'taken by themselves' here, we mean: *independently of the preferences of any other individual concerning those same objects*. In the current form of the EP program, by contrast, each individual's idealised preferences are taken to be those that in some sense she would have if (she were otherwise fully informed and rational, but in addition) she *knew everyone else's idealised preferences* (together with the truth about which theory of well-being is correct). This means that we are no longer dealing with a 'linear' theory in which we can (even in principle) first work out, *by looking at each individual separately*, what that individual's preferences among extended alternatives would be under conditions of full information and rationality—say, by considering a process of seeking reflective equilibrium in light of all the *non-preference* facts and under ideal deliberative conditions—and then move on to determine the betterness facts on the basis of the already-fixed preference profile. We have, rather, an holistic theory, in which the facts about what *each* individual's extended preferences would be under ideal conditions is determined simultaneously, on the basis of the actual psychological states of *all* individuals. (According to the theory, and on one standard semantics for counterfactuals, those facts are given by the goings-on in the closest full-agreement possible world to the actual one, where a 'full-agreement' world is one in which all individuals have identical extended preferences; full information and rationality of all individuals are of course also required.) Relatedly, the extended preferences that Robert would have under conditions of full information and rationality *if Sarah did not exist* can be different from those he would have under conditions of full information and rationality given that Sarah does exist, and this not for any reasons relating to concern for Sarah, but merely because Robert cannot count as being fully informed and rational unless his extended preferences agree with those that Sarah would have if she were similarly fully informed and rational. None of this is literally incoherent, but we suspect that this feature of the view will make it unattractive to most preference-satisfaction theorists.

Second, and relatedly, the form of 'full-information' in the theory (as in the one which includes only WPB, but somehow rules out enough disagreement to avoid massive incomparability) will make it unpalatable to those preference-satisfaction theorists who wish to maintain a relationship between individuals' actual preferences and the preferences called on in preference-satisfaction theory. To recall, in the introduction we noted a distinction between preference-satisfaction theorists who view the full-information condition as merely a way of ensuring that the preferences used in preference-satisfaction theory respect the (actual) fundamental values of the agent in question, and those who believe that idealisation may alter even the fundamental values of the agent. The former kind of preference-satisfaction theorist should reject a theory based on SPB. For the form of idealisation we are considering now requires so much deviation from actual agents' psychology that it is implausible that it would *not* sometimes result in a change in fundamental value. The resulting theory thus loses out on what this form of preference-satisfaction theory sees as one of the fundamental motivations for the theory: that an agent is better-off to the extent that *her own* fundamental preferences are satisfied.

Third, the resulting form of preference-satisfaction theory is stated as a biconditional, but in general preference-satisfaction theories of well-being are intended also to be *analyses* of well-being. It is not just that self-interested, rational, fully-informed preferences coincide with facts about well-being; in some sense these facts about preferences also *explain* the facts about well-being. But if we take the view currently on offer as an analysis, it suffers from a particularly egregious form of ungroundedness. To illustrate the problem,

consider an analogous analysis of mathematical truth. According to this theory: for 1 + 1 to equal 2 is for a rational agent to believe that they are equal. We then ask what it is for an agent to be rational, and part of the answer is that he or she has correct beliefs about all mathematical truths. The resulting 'analysis' of mathematical truth is hard to understand: 1 + 1=2 because an agent who has true beliefs about this equation believes that they are equal. The analysis of well-being on offer here faces something like the same circularity: Zeyad's preferences are rational insofar as they agree with betterness, and one outcome is better than another insofar as it agrees with rational preferences (including Zeyad's).

To sum up: there is, after all, a consistent position which combines an EP theory of interpersonal comparisons with an (ordinary-)preference-satisfaction theory of well-being, namely, the one that includes SPB. This may be the best that the extended-preferences theorist can do. But the position seems highly unattractive; we do not think that it will hold much appeal for the would-be preference-satisfaction theorist of well-being.

10. Conclusion

Whatever its other merits and problems, a preference-satisfaction theory of well-being faces a *prima facie* problem in recovering interpersonal comparisons of well-being. If it cannot recover them, it is false, since there obviously are some.

The EP program seeks to recover interpersonal well-being comparisons within a preference-satisfaction theory of well-being by appeal to individuals' preferences over so-called *extended* alternatives. We have argued that there is no obvious conceptual difficulty in understanding what extended preferences are: extended preferences really are preferences; they need not involve preferences over obviously impossible contents, or be understood as beliefs, or be explicated using the machinery of a veil of ignorance. The extended preferences of any given individual define a standard of interpersonal comparisons; if PFC were true, there would be a unique such standard. But against PFC, we have observed, following Broome, that distinct individuals' extended preference orderings often respect the differing values that the individuals in question have, and nothing in the appeal to 'structural idealisation' (unlike a notion of *substantive* idealisation) can wash out these differences.

We then argued that the EP theory faces a trilemma. There is a direct argument from ordinary preference-satisfaction theory in the context of the EP program for PA; if the EP program is to derive interpersonal comparisons of well-being which respect the ordinary preferences of each individual, then very plausibly PA must hold. But at least on the most natural way of understanding the conditions of full information and rationality that are appealed to in preference-satisfaction theories of well-being, PA appears to be false. That was the first horn of the trilemma: giving up on PA, and hence being forced to give up on the preference-satisfaction theory of well-being.

The second horn of the trilemma arises if the EP theorist embraces WPB and hence is able to argue for PA, but still is unable to rule out widespread disagreement among rational, fully informed preferences. In that case, WPB will give rise to an unacceptable degree of incomparability over different centred alternatives. This degree of incomparability once again amounts essentially to the failure of the program.

The third horn of the trilemma involves accepting a stronger connection between betterness facts and ideal preferences. This stronger connection (SPB) facilitates an argument for PFC, but, we have argued, comes at the cost of odd forms of holism and

ungroundedness, plus an arguably undesirable degree of divorce between actual preferences (even actual fundamental values) and the preferences that are relevant to well-being. If these features are, as we ourselves judge, unacceptable, the conclusion has to be that extended preferences cannot serve as the ground for interpersonal comparisons within a preference-satisfaction approach to well-being.

This conclusion, of course, does not show that preference-satisfaction theories specifically—or attitude-satisfaction theories more generally—are false. There *are* other ways in which the preference-satisfaction theorist might seek to make interpersonal well-being comparisons. We close by sketching two of these approaches briefly. Given what we have argued is the failure of the EP program, we recommend that the preference satisfaction theorist pursue one of these alternative options.

On a first, 'structuralist', approach, interpersonal well-being comparisons are grounded in information already available in each individual's preference ordering. The general idea may be best explained by an example of such a proposal. Suppose that the number of ordinary alternatives is finite. Then, for each individual i and each ordinary alternative x, there is a natural number $n(i,x)$, representing the position of alternative x in i's preference ordering. The structuralist might then define interpersonal well-being comparisons as follows. First, level comparisons: state of affairs x is as good for person i as state of affairs y is for person j just in case $n(i,x)=n(j,y)$. Second, unit comparisons: the ratio of the difference between x and y for i to the difference between v and w for j is given by $\frac{n(i,x)-n(i,y)}{n(j,v)-n(j,w)}$.[23]

A second, 'primitivist' approach—which represents the alternative way of thinking about strength of desire which we flagged in the introduction—has been even less explored. On this view, there are primitive facts about *preference strength*, of at least two kinds. First, there are primitive, interpersonally comparable facts about the strength of S's desire for x; second, there are primitive, interpersonally comparable facts about the *degree* to which S prefers x to y.[24] Given facts about strength of desire, we can perform level comparisons. Given facts about degree of preference, we can perform unit comparisons.[25] Much more needs to be said about the foundations and feasibility of this approach, but to our knowledge it has not been much explored, and it seems to us to have as good a chance as structuralist approaches for saving the preference-satisfaction theory.

If the preference-satisfaction theory is unable to explain or explain away the obvious facts there are about interpersonal comparisons of well-being, the theory should be abandoned. Those who wish to defend the theory, then, must explain how it can make sense

---

[23] See e.g. Jeffrey (1971, 655), Hammond (1991, 216), Rawls (1999, 283–284), Griffin (1986), Cotton-Barratt *et al*. (2014), Hausman (1995). Cf. also Isbell (1959) and Schick (1971).

[24] Perhaps the more exotic aspect of this theory is the first, monadic notion of desire strength. But in support of the existence of such interpersonally comparable units of desire strength, a proponent of the approach might note that we do ordinarily say such things as that one child wants an extra scoop of ice cream, or that one player wants a victory, *more than another does*. There is a worry that these English monadic 'desire comparisons' may covertly describe dyadic preference-strength comparisons; for example, the first player prefers winning over losing to a greater degree than the second prefers winning over losing. But perhaps this worry can be answered.

[25] In fact, there are well-known results describing the conditions under which orderings of levels and orderings of differences in strength can be used to construct utility functions, just as preferences on lotteries do in more standard decision theory. (See Alt (1936; 1971), Krantz *et al*. (1971, 151), and references in the latter.) If the primitive, interpersonally comparable facts about preference and desire strength satisfy these extra conditions, then one could also speak of an interpersonal utility function.

of interpersonal comparisons. We have argued that the EP program is not a promising way of doing this. But it remains to be seen whether some other approach—perhaps one of the two just mentioned—can do better.[26]

References

Adler, Matthew. 2012. *Well-being and Fair Distribution: Beyond Cost-benefit Analysis*. Oxford: Oxford University Press.

——. 2014. Extended preferences and interpersonal comparisons: A new account. *Economics and Philosophy*, 30(2), 123–162.

——. 2016a. Extended preferences. Pages 476–517 of: Adler, Matthew & Fleurbaey, Marc (eds), *Oxford Handbook of Well-Being and Public Policy*. Oxford: Oxford University Press.

——. 2016b. Aggregating moral preferences. *Economics and Philosophy*, 32(2), 283–321.

Alt, Franz. 1936. Über die messbarkeit des nutzens. *Journal of Economics*, 7(2), 161–169.

——. 1971. On the measurability of utility. Pages 424–431 of: Chipman, John, Hurwicz, Leonid, Richter, Marcel, & Sonnenschein, Hugo (eds), *Preferences, Utility and Demand: A Minnesota Symposium*. New York: Harcourt Brace Jovanovich. (This is a translation of Franz (1936).)

Arrow, Kenneth 1963. *Social choice and Individual Values*. 2nd edn. New York: John Wiley and Sons.

——. 1978. Extended sympathy and the possibility of social choice. *Philosophia*, 7(2), 223–237.

Broome, John. 1998. Extended preferences. Pages 271–287 of: Fehige, Christoph, & Wessels, Ulla (eds), *Preferences*. Berlin: W. de Gruyter.

——. 2008. Can there be a preference-based utilitarianism? Pages 221–238 of: Fleurbaey, Marc, Salles, Maurice, & Weymark, John (eds), *Justice, Political Liberalism and Utilitarianism: Themes from Harsanyi and Rawls*. Cambridge: Cambridge University Press.

Byrne, Alex & Hájek, Alan. 1997. David Hume, David Lewis, and decision theory. *Mind*, 106(423), 411–728.

Cotton-Barratt, Owen, MacAskill, William, & Ord, Toby. 2014. Normative uncertainty, intertheoretic comparisons, and variance normalisation. Unpublished manuscript.

Fine, Kit. 2012. The structure of joint intention. Unpublished manuscript.

Greaves, Hilary. 2016. A reconsideration of the Harsanyi-Sen-Weymark debate on utilitarianism Forthcoming in *Utilitas*.

—— & Lederman, Harvey. 2016. Aggregating extended preferences. Forthcoming in *Philosophical Studies*.

Griffin, James. 1986. *Well-being: Its Meaning, Measurement, and Moral Importance*. Oxford: Oxford University Press.

[26] Thanks to John Broome, Jake Nebel, Itai Sher and Teru Thomas for thoughtful comments on an earlier draft, and to the audiences of presentations at Boston University, the University of Southern California and the Australian National University for their comments and criticisms. The authors contributed equally to this paper.

Hammond, Peter. 1991. Interpersonal comparisons of utility: Why and how they are and should be made. Pages 200–254 of: Elster, Jon & Roemer, John E. (eds), *Interpersonal Comparisons of Well-being*. Cambridge: Cambridge University Press.

Harsanyi, John. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 61(5), 434.

——. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

——. 1982. Morality and the theory of rational behavior. Pages 39–62 of: Sen, Amartya & Williams, Bernard (eds), *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Hausman, Daniel. 1995. The impossibility of interpersonal utility comparisons. *Mind*, 104(415), 473–490.

Isbell, John Rolfe. 1959. Absolute games. Pages 257–296 of: Tucker, Albert William & Luce, Robert Duncan (eds), *Contributions to the Theory of Games*, vol. 4. Princeton: Princeton University Press.

Jeffrey, Richard. 1971. On interpersonal utility theory. *Journal of Philosophy*, 68(20), 647–656.

Kolm, Serge-Christophe. 2002. *Justice and Equity*. Cambridge: MIT Press.

Krantz, David, Luce, Duncan, Suppes, Patrick, & Tversky, Amos. 1971. *Foundations of Measurement (Additive and Polynomial Representations)* vol. 1. Mineola, New York: Dover Publications.

Lewis, David. 1979. Attitudes de dicto and de se. *Philosophical Review*, 88(4), 513–543.

——. 1988. Desire as belief. *Mind*, 97(387), 323–332.

——. 1996. Desire as belief II. *Mind*, 105(418), 303–313.

——. 2004. Letters to Priest and Beall. Pages 176–177 of: Priest, Graham, Beall, J. C., & Armour-Garb, Bradley (eds), *The Law of Non-contradiction*. Oxford: Oxford University Press.

McKay, Thomas, & Nelson, Michael. 2014. Propositional attitude reports. In: Zalta, Edward (ed), *The Stanford Encyclopedia of Philosophy*, spring 2014 edn.

Pattanaik, Prasanta 1968. Risk, impersonality, and the social welfare function. *Journal of Political Economy*, 76(6), 1152–1169.

Perry, John. 1979. The problem of the essential indexical. *Noûs*, 13(1), 3–21.

Price, Huw. 1989. Defending desire-as-belief. *Mind*, 98(389), 119–127.

Rachels, Stuart. 1998. Counterexamples to the transitivity of 'better than'. *Australasian Journal of Philosophy*, 76(1), 71–83.

Rawls, John. 1999. *A Theory of Justice*. Revised edn. Cambridge: Belknap Press of Harvard University Press.

Schick, Frederic. 1971. Beyond utilitarianism. *The Journal of Philosophy*, 68(20), 657–666.

Sen, Amartya. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

——. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision*, 7(4), 243–262.

——. 1977. Non-linear social welfare functions: A reply to Professor Harsanyi. Pages 297–302 of: Butts, Robert, & Hintikka, Jaakko (eds), *Foundational Problems in the Special Sciences*, vol. 2. Dordrecht, Holland: D. Reidel Publishing Company.

Temkin, Larry. 1987. Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, 16(2), 138–187.

——. 2014. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford: Oxford University Press.

Voorhoeve, Alex. 2014. Review of Matthew D. Adler: Well-being and fair distribution: Beyond cost-benefit analysis. *Social Choice and Welfare*, 42(1), 245–254.

Weymark, John. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. Pages 255–320 of: Elster, John & Roemer, John (eds), *Interpersonal Comparisons of Well-being*. Cambridge: Cambridge University Press.

——. 2005. Measurement theory and the foundations of utilitarianism. *Social Choice and Welfare*, 25(2–3), 527–555.