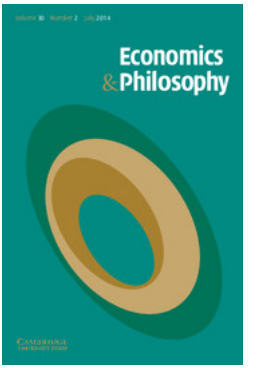

Matthew D. Adler

Economics and Philosophy / Volume 30 / Issue 02 / July 2014, pp 123 - 162
DOI: 10.1017/S0266267114000133, Published online: 02 June 2014

http://journals.cambridge.org/abstract_S0266267114000133

Matthew D. Adler (2014). EXTENDED PREFERENCES AND INTERPERSONAL COMPARISONS: A NEW ACCOUNT . Economics and Philosophy, 30, pp 123-162 doi:10.1017/S0266267114000133

Economics and Philosophy, **30** (2014) 123–162 © Cambridge University Press
doi:10.1017/S0266267114000133

EXTENDED PREFERENCES AND INTERPERSONAL COMPARISONS: A NEW ACCOUNT

MATTHEW D. ADLER
Duke University, USA
adler@law.duke.edu

This paper builds upon, but substantially revises, John Harsanyi's concept of 'extended preferences'. An individual 'history' is a possible life that some person (a subject) might lead. Harsanyi supposes that a given spectator, formulating her ethical preferences, can rank histories by empathetic projection: putting herself 'in the shoes' of various subjects. Harsanyi then suggests that interpersonal comparisons be derived from the utility function representing spectators' (supposedly common) ranking of history lotteries. Unfortunately, Harsanyi's proposal has various flaws, including some that have hitherto escaped scholarly attention. In particular, it ignores the limits of personal identity. If the subject has welfare-relevant attributes that the spectator cannot acquire without changing who she is, full empathetic identification of the latter with the former becomes impossible. This paper proposes instead to use *sympathy* as the attitude on a spectator's part that allows us to make sense of her extended preferences. Sympathy – an attitude of care and concern – is a psychological state quite different from empathy. We should also allow for hetereogeneity in spectators' extended preferences. Interpersonal comparisons emerge from a plurality of sympathetic spectators, not (as per Harsanyi) from a common empathetic ranking.

1. INTRODUCTION

Welfare economics has never resolved the puzzle of interpersonal welfare comparisons. While some economists continue to hold the view

Many thanks to Koen Decanq, Marc Fleurbaey, Francois Maniquet and an anonymous referee for *Economics and Philosophy* for very helpful comments. All errors are my own.

(dominant around the middle of the 20th century) that such comparisons are to be avoided, there are many subfields of modern economics where interpersonal comparability is accepted. In particular, this is true of subfields that use 'social welfare functions' to evaluate policies or institutions. The social-welfare-function framework works as follows. A policy or institution corresponds to some probability distribution across possible outcomes. Each outcome, in turn, corresponds to a vector of utility numbers – one for each of the individuals in the population, with $u_i(x)$ the utility of a given individual i in a given outcome x. Outcomes are ordered using some rule for ranking their corresponding utility vectors.

Critically, the social-welfare-function framework presupposes that utility numbers represent inter- as well as intrapersonal well-being comparisons. The *numerical* fact that $u_i(x) > u_j(y)$ is supposed to indicate the *well-being* fact that individual i in outcome x is better off than individual j in outcome y. But what is the basis for this supposed well-being fact?

Economists, at least, tend to reduce well-being to preferences. Each individual is seen to have preferences over outcomes. An outcome-preference is a psychological state on the part of the holder of the preference, which takes the form of a relation between him and outcomes (individual i prefers outcome x to outcome y), and plays the psychological role of motivating his choices (depending on the possible outcomes of the choices available to him). While each individual's outcome-preferences seem to offer a straightforward basis for well-being facts regarding that individual – if individual i prefers outcome x to y, then *he* is better off in x than y, or so economists usually suppose – these preferences provide no apparent basis for well-being comparisons across persons. How are we to move from a relation between a particular person and outcomes, to an interpersonal well-being comparison – which is a relation involving outcomes and *multiple* persons? This, in a nutshell, is the puzzle of interpersonal comparisons; and welfare economics has yet to arrive at a clear, consensus solution to it.

In this paper, I propose a solution to the puzzle.[1] The solution is based upon the idea of 'extended preferences' – an idea that John Harsanyi pioneered,[2] and that other scholars have employed as well.[3] An extended

[1] The concept of 'equivalent income' provides a different possible solution. Fleurbaey and Blanchet (2013: ch. 4); Fleurbaey (forthcoming). I contrast equivalent incomes and extended preferences in Adler (forthcoming).

[2] Harsanyi (1977: ch. 4). See also Harsanyi (1953, 1955, 1982). For a lucid presentation of Harsanyi's views, see Weymark (1991); Mongin and d'Aspremont (1998: 444–459); Mongin (2001).

[3] For a review of the broader literature on extended preferences, see Suzumura (1996). Recent examples include Gajdos and Kandil (2008); and Grant *et al.* (2010, 2012a, 2012b). This literature is also discussed by Mongin and d'Aspremont (1998: 444–459); Mongin

preference is a ranking, not of outcomes, but of *histories*: pairings of outcomes and individuals ('subjects').

However, Harsanyi's specific conception of 'extended preferences' is problematic. His is an *empathy-based conception*. On this view, a person (the 'spectator') develops extended preferences via empathetic projection. She compares histories by considering what it would be like for her to 'stand in the shoes' of each subject – to acquire the attributes that each subject has in his history. The empathy-based conception of extended preferences has two key flaws. (1) *The essential attribute problem*. There are attributes that subjects might possess which the spectator necessarily lacks: attributes that the spectator cannot acquire without changing who she is. Insofar as subjects possess these properties, the spectator cannot really 'stand in their shoes'. (2) *The 'wrong kind of preference' problem*. The spectator might develop an extended preference for reasons that have nothing to do with well-being, e.g. for moral reasons. Where some or all spectators have the 'wrong kind of preference' for one history over a second, it would be unwarranted to use these preferences as the basis for a well-being comparison between the histories.

Part 2 of the paper sets forth both the general concept of extended preferences, and Harsanyi's empathy-based conception thereof. Part 3 criticizes the empathy-based conception. The presentation and critique in these parts are focused on Harsanyi's scholarship – since his writings about extended preferences have been highly influential and, in his book (1977), are quite fully developed. But let me suggest (without exhaustively demonstrating) that other writers about extended preferences have also tended to adopt the problematic view that a spectator develops an extended preference over histories by empathetically projecting herself into the positions of the histories' subjects.[4]

(2001), and placed in a yet broader intellectual tradition of 'impartial observer' theories of morality.

[4] To be sure, much work on extended preferences is formal. Preferences over histories are defined as abstract objects, i.e. rankings, and the logical consequences of various axiomatic restrictions on these rankings and associated social welfare functions are explored. Such axiomatic analysis does not depend upon the interpretation of such rankings – the psychological content ascribed to them. However, to the extent that scholars in this area *do* express a position on the psychological content of an extended preference, they seem to see it as a preference to acquire the subject's attributes. See, e.g. Sen (1970: 131, 152); Arrow (1977: 224–225); Grant *et al*. (2010: 1957–1958; 2012b: 834–835).

Unfortunately, this empathy-based interpretation of an extended preference is obscured by terminology. The term 'extended sympathy' is sometimes used to describe the generic concept of extended preferences (which is neutral between empathy- and sympathy-based approaches) and, indeed, even to refer to the empathy-based conception. For example, in his survey of the extended-preferences literature, Kotaro Suzumura writes: 'Interpersonal welfare comparisons of the so-called *extended sympathy type* may be formulated in operational form as follows: it is better in my judgement to be put in your position in social

Still, extended preferences remain a fruitful basis for interpersonal comparison. In Part 4, I propose to analyse them in terms of *sympathetic* rather than empathetic spectators. This sympathy-based conception derives a spectator's ranking of histories from her ranking of outcomes under a condition of care and concern for subjects, and from her well-being judgements. In Parts 5 and 6, I discuss how the sympathy-based conception provides the tools for constructing interpersonal comparisons of well-being levels and differences.

The terms 'sympathy' and 'empathy' are often used interchangeably, or without close attention to their meaning. But the terms are *not* synonyms; they pick out two, quite distinct psychological states. And this difference, I suggest, has real significance for welfare economics, since by attending to it we can make progress on the problem of interpersonal comparisons.

Finally, I should note that the analysis in this paper is independent of debates about the appropriate form of the social welfare function. While Harsanyi developed the concept of extended preferences in the course of his defence of utilitarianism, interpersonal comparisons of levels and/or differences are also generally required for non-utilitarian social welfare functions (for example, leximin or 'prioritarian' functions). The sympathy-based conception I develop is meant to mesh with such functions, as well as with utilitarianism.[5]

2. INTERPERSONAL COMPARISONS, EXTENDED PREFERENCES AND HARSANYI'S APPROACH

Formally, the problem of interpersonal comparisons can be expressed as follows. Let **O** be a set of outcomes, i.e. states of affairs, $\{x, y, z \ldots\}$. Let **N** be a set of individuals, $\{1, 2, \ldots, N\}$. **N** has a finite number of members N, each of whom exists in all of the outcomes (problems of variable and infinite populations are ignored here). A well-being comparison is not a ranking of outcomes, simpliciter, but rather a ranking of outcomes relativized to individuals. We say that one outcome is better *for* Jim than a second – and, if interpersonal comparisons are possible, that *Sheila* is better off in some outcome than *Jim* in some outcome. This relativization can be captured via the concept of an individual 'history'. As I will define it, a history $(x; i)$ is a pairing of an individual, i, and an outcome, x. The set **H** is the set

state x than to be put in somebody else's position in social state y' (1996: 202). Suzumura uses the term 'sympathy' to mean what is properly called 'empathy'.

[5] Whatever the form of the social welfare function, important questions arise about how it should incorporate the information in $\succcurlyeq^{WB}$ and $\succcurlyeq^{DIFF}$ if these well-being structures are not represented by a single utility function. This is a topic I have addressed elsewhere, and will not discuss here. See Adler (2012: chs 2, 5).

of all histories. In other words, $\mathbf{H}$ is $\mathbf{O} \times \mathbf{N}$. A well-being ranking, denoted '$\succcurlyeq^{WB}$', is a quasiordering on $\mathbf{H}$.'$(x;i)\succcurlyeq^{WB}(y;j)$' should be interpreted as 'individual i in outcome x is at least as well off as individual j in outcome y'. Having strictly greater well-being ($\succ^{WB}$) and being equally well off ($\sim^{WB}$) are derived from $\succcurlyeq^{WB}$ in the standard way.

A quasiordering, recall, is a binary relation which is transitive and reflexive but not necessarily complete.[6] For a given pair of histories (x; i) and (y; j), it is possible that neither $(x;i)\succcurlyeq^{WB}(y;j)$, nor $(y;j)\succcurlyeq^{WB}(x;i)$. We can therefore say that $\succcurlyeq^{WB}$ makes some interpersonal comparisons iff: there exists at least one pair of histories (x; i) and (y; j), with i and j distinct individuals ($i \neq j$), such that $(x;i)\succcurlyeq^{WB}(y;j)$. Conversely, $\succcurlyeq^{WB}$ is only intrapersonally comparable iff, for every pair of histories such that $(x;i)\succcurlyeq^{WB}(y;j), i = j$.

While the formalization thus far expresses comparisons of well-being *levels*, we should also keep in view comparisons of well-being *differences*. Consider the set $\mathbf{H} \times \mathbf{H}$, comprised of all *pairs* of histories. Then $\succcurlyeq^{DIFF}$ is a quasiordering on $\mathbf{H} \times \mathbf{H}$.[7] '$((x;i),(y;j))\succcurlyeq^{DIFF}((z;l),(w;m))$' should be interpreted as: the difference between the well-being of individual i in x and the well-being of individual j in y is at least as large as the difference between the well-being of individual l in z and individual m in w. $\succcurlyeq^{DIFF}$ makes some interpersonal difference comparisons iff there is at least one case in which $((x;i),(y;j))\succcurlyeq^{DIFF}((z;l),(w;m))$ and it is not the case that $i = j = l = m$.

The puzzle of interpersonal comparisons – thus – is to use information about individual preferences so as to construct a $\succcurlyeq^{WB}$ that makes *some* interpersonal level comparisons and/or a $\succcurlyeq^{DIFF}$ that makes *some* interpersonal difference comparisons.

What is an extended preference? We should distinguish between the general *concept* of an extended preference, and particular *conceptions* or versions of this general concept. A preference, on the part of some individual (Raj), is a ranking that connects to Raj's choices. While an outcome-preference is a choice-connected ranking of outcomes, an 'extended preference' is a choice-connected ranking of individual *histories*. Generically, I will use the term 'spectator' to refer to the *holder* of an extended preference, and the term 'subject' to refer to the individuals *in* histories. Phil and Jim are the 'subjects', respectively, of the histories (x; Phil) and (y; Jim). Raj, in holding extended preferences regarding these histories, is a 'spectator'.

But what does it *mean* to rank histories? How are we to analyse the content of Raj's preference for (x; Phil) over (y; Jim)? Here, Harsanyi

[6] See, e.g. Donaldson and Weymark (1998).

[7] In order to conform to truisms about well-being differences, $\succcurlyeq^{DIFF}$ must also satisfy additional constraints. See below, Part 6.

offers a particular *conception* of extended preferences. Raj is supposed to imagine 'standing in Phil's shoes' in outcome x. In other words, Raj is supposed to imagine what it would be like to possess the physical, social and other attributes that Phil has in outcome x, and also to have Phil's tastes. Raj is next supposed to imagine 'standing in Jim's shoes' in outcome y: possessing now the physical, social, etc. attributes of Jim in outcome y, along with Jim's tastes. And, finally, Raj is supposed to rank (x; Phil) versus (y; Jim) by deciding whether he prefers to have the first attribute bundle or the second.

As Harsanyi explains:

> [T]he basic intellectual operation in … interpersonal comparisons is imaginative empathy. We imagine ourselves to be in the shoes of another person, and ask ourselves the question, 'If I were now really in *his* position, and had *his* taste, *his* education, *his* social background, *his* cultural values, and *his* psychological makeup, then what would now be *my* preferences between various alternatives … ? (An 'alternative' here stands for a given bundle of economic commodities plus a given position with respect to various non-economic variables, such as health, social status, job situation, family situation, etc.)[8]

[8] Harsanyi (1982: 50). Harsanyi makes similar statements – expressing the standing-in-the-shoes or empathy-based conception of extended preferences – at other junctures. For example:

> Our model [of interpersonal comparisons] is based on the assumption that … each individual i will try to assess the utilities $U_j(A)$ that any *other* individual j would derive from alternative social situations A and will try to compare these with the utilities $U_i(A)$ that he *himself* would derive from these (or from other) social situations. That is, he will try to make *interpersonal utility comparisons*. Moreover, we have assumed that i will attempt to assess these utilities $U_j(A)$ by some process of *imaginative empathy*, i.e. by imagining himself to be *put in the place* of individual j in social situation A.
>
> This must obviously involve his imagining himself to be placed in individual j's *objective position*, i.e. to be placed in the objective conditions (e.g. income, wealth, consumption level, state of health, social position) that j would face in social situation A. But it must also involve assessing these objective conditions in terms of j's own *subjective attitudes and personal preferences* … (Harsanyi 1977: 51–52)

A few paragraphs later, discussing i's comparison between the extended alternatives [A_i, P_i] and [B_j, P_j], with A_i and B_j the objective positions of the two individuals and P_i and P_j their subjective attitudes, Harsanyi explains: '[This comparison] will really amount to [i] trying to decide whether he himself would *prefer* to be in the objective position A_i with his *own* subjective attitudes P_i, or rather to be in the objective position B_j with j's subjective attitudes P_j ….' *Id.* at 52. And again, further down: '[Extended preferences] are preferences between partly *imaginary* alternatives [by allowing for a change of tastes], for example, between eating meat with one's actual taste and eating fish with a taste quite different from one's actual taste'. *Id.* at 53. Thus the holder of the extended preference imagines *his* eating meat or fish and *his* having certain tastes.

Harsanyi also repeatedly expresses his equiprobability model of moral judgements in terms of preferences with respect to being 'put in the place' of different individuals. See Harsanyi (1953: 435; 1955: 316; 1977: 50).

For short, I will term the particular conception of extended preferences which Harsanyi adopts the *empathy*-based conception, and I will express it as follows.

The Empathy-Based Conception of Extended Preferences

As above, **O** is the set of outcomes, **N** a finite set of individuals (each of whom exists in all outcomes), and $\mathbf{H} = \mathbf{O} \times \mathbf{N}$ the set of all histories $\{(x; i)\}$. Let **K** be the set of spectators. For simplicity, assume that $\mathbf{K} = \mathbf{N}$; each individual is both the subject of histories, and someone who develops extended preferences over histories.

The outcomes in **O** are arbitrarily detailed specifications of possible worlds, but do not specify individuals' preferences. Let $R = (R_1, R_2, \ldots, R_N)$ be a possible profile of outcome and choice preferences on the part of individuals 1, 2 $\ldots$, N. I will refer to R_i as the 'tastes' of individual i. (This is done purely for terminological convenience, and does not imply a position about the content or rational grounding of i's outcome and choice preferences.)[9]

For any given R, each spectator k has *extended* preferences over **H**. Denote these extended preferences as $\succcurlyeq^k(R)$, with '(R)' indicating that the spectator's extended preferences depend upon the profile of

[9] It is important not to conflate outcome and choice preferences, on the one hand, with extended preferences; and the term 'taste' is less clumsy than 'outcome and choice preferences'. Sometimes the word 'taste' is used to denote a particular kind of outcome or choice preference, e.g. one that is arbitrary and cannot be given any substantive justification; but that is not my intention here.

Why define outcomes to exclude tastes? The framework outlined in Part 5 for aggregating extended preferences is a 'multiprofile' framework, whereby different possible profiles of extended preferences are mapped onto different well-being quasiorderings of a single history set associated with a single outcome set. Because there are systematic connections between extended preferences and individual tastes, it must also be possible (if this multiprofile approach is adopted) to have different profiles of tastes associated with the very same outcome set. Thus outcomes cannot include tastes.

The general set-up put forth here is somewhat different from Harsanyi's. Among other things, it allows for the possibility that well-being-relevant attributes include individuals' *essential* attributes (see below Part 3.1), since these may be included in the description of the various outcomes; while Harsanyi stipulates at the outset that the attributes over which individuals have extended preferences are attributes each person could have (Harsanyi 1977: 53). However, the central feature of the set-up (that spectators have extended preferences over hybrid bundles, consisting of both non-taste attributes and tastes) is exactly what Harsanyi proposes.

Each R_i might be quite complex, since a particular individual might have different rankings of outcomes depending upon his attitude. In particular, his *moral* ranking of outcomes will differ from his *self-interested* ranking. See below Part 3.2. An individual's ranking of choices reflects not only his outcome ranking, but his risk aversion. For Harsanyi, at least, it is critical that R_i include such information.

tastes. (To avoid clutter, however, '(R)' will generally be dropped in my presentation and '$\succcurlyeq^k$' used to indicate $\succcurlyeq^k(R)$.)

Formally, $\succcurlyeq^k$ is a quasiordering of $\mathbf{H}$.[10] *Substantively*, k develops $\succcurlyeq^k$ via empathetic projection. Let $A_i(x)$ denote the (non-taste) attributes of individual i in outcome x. Let $(A_i(x), R_i)$ denote a hybrid attribute bundle, consisting of both the attributes of individual i in outcome x, and tastes R_i. Then $(x; i) \succcurlyeq^k (y; j)$ iff k weakly prefers to have as his own attributes the attributes in the hybrid bundle $(A_i(x), R_i)$, as compared to having as his own attributes the attributes in the hybrid bundle $(A_j(y), R_j)$.

'Empathy' – a word that Harsanyi himself uses – is the capacity to assume someone else's perspective. On the 'empathy' based conception, the spectator develops extended preferences by exercising this capacity. In order to compare two histories, the spectator empathetically projects herself into the position of the first subject (considering both that subject's non-taste attributes in the first history, and his tastes), and then empathetically projects herself into the position of the second subject (considering now that subject's non-taste attributes in the second history, and his tastes). This empathetic projection by the spectator is nothing other than the thought exercise of taking on the properties (including the feelings and tastes) of each subject. Thus the 'empathy-based' conception of extended preferences might also be termed an 'attribute-acquisition' conception.

Having articulated a conception of extended preferences, Harsanyi proceeds to arrive at interpersonal comparisons of well-being levels and differences by making two further assumptions. For short, call these the 'Convergence' premise and the 'Bernoulli' premise. The Convergence premise is that spectators have the very same extended preferences. If Raj prefers (x; Phil) to (y; Jim), then so does Sally. The 'Bernoulli' premise is that spectators' extended preferences over history lotteries comply with the requirements of von-Neumann Morgenstern (vNM) expected utility theory, hence can be expectationally represented by vNM utility functions; and that a measure of well-being differences is derivable (in a linear fashion) from these utility functions.

The Convergence and Bernoulli premises, together, allow Harsanyi to define $\succcurlyeq^{\text{WB}}$ and $\succcurlyeq^{\text{DIFF}}$ as follows. Let $v(.)$ be a vNM function expectationally representing the spectators' common extended preferences over history lotteries. Then $(x; i) \succcurlyeq^{\text{WB}} (y; j)$ iff $v(x; i) \geq v(y; j)$. $((x; i), (y; j)) \succcurlyeq^{\text{DIFF}} ((z; l), (w; m))$ iff $v(x; i) - v(y; j) \geq v(z; l) - v(w; m)$.

[10] While Harsanyi assumed a complete ordering, I will generalize and allow for $\succcurlyeq^k$ to be incomplete.

The critical literature on Harsanyi has discussed the Convergence and Bernoulli premises at length.[11] The first seems untrue. There is no particular reason to believe that individuals will have identical extended preferences. Harsanyi offers an argument for Convergence, but the argument is flawed.

The Bernoulli premise remains open to reasonable debate, but is much more contestable than Harsanyi realized. Consider the simplest case, where Convergence happens to hold true and spectators *do* have identical extended preferences over histories and lotteries. Thus there is a common extended vNM utility function $v(.)$. Why should it be true that $((x;i),(y;j)) \succcurlyeq^{\text{DIFF}} ((z;l),(w;m))$ iff $v(x;i) - v(y;j) \geq v(z;l) - v(w;m)$? Isn't it possible that spectators, in ranking lotteries, are risk-averse or risk-prone in well-being – in other words, that $\succcurlyeq^{\text{DIFF}}$ corresponds to differences in some convex or concave transformation of $v(.)$?

These criticisms of Harsanyi are familiar from the literature. In the next Part, I present a different and more novel criticism.[12] This criticism is logically independent of the familiar critiques of Convergence and Bernoulli. And, in a sense, the criticism is deeper. I argue that Harsanyi's account of the *content* of extended preferences – the empathy-based conception – is problematic. In effect, Harsanyi stumbles from the get-go. An extended preference is *some* kind of preference or judgement on the part of the spectator; but to characterize it, specifically, as a preference to acquire subjects' attributes and tastes faces serious difficulties. The critique of the empathy-based conception of extended preferences, which I develop in this paper, would also apply to a neo-Harsanyi view that relaxes Convergence and/or Bernoulli but retains the problematic analysis of extended preferences in terms of empathetic projection.

3. THE EMPATHY-BASED CONCEPTION OF EXTENDED PREFERENCES: A CRITIQUE

This Part describes two important flaws in the empathy-based conception of extended preferences – flaws that have received little attention in the literature. More precisely, these are flaws in any account of well-being

[11] On Convergence, see, e.g. Broome (1998); Mongin (2001). On Bernoulli, see Sen (1976, 1986: 1122–1123); Weymark (1991, 2005); Broome (1995); Mongin and d'Aspremont (1998); Risse (2002); Roemer (2008); Fleurbaey and Mongin (2012). The term 'Bernoulli' to mean the derivation of a measure of well-being differences from lottery preferences is taken from Broome.

[12] The empathy/sympathy distinction is brought to light, with reference to Harsanyi's work, by Mongin (2001), and more generally with reference to economic theory by Fontaine (1997). However, much of the critical analysis presented in Part 3 of this paper, as well as the positive proposal in Part 4, is – I believe – novel.

comparisons (such as Harsanyi's) that builds from this conception.[13] The first flaw is the 'essential attribute' problem; the second, the 'wrong kind of preference' problem.

3.1. The Essential-Attribute Problem

Recall that, on the empathy-based conception, the spectator k is asked to compare $(x; i)$ and $(y; j)$ by comparing the state of affairs in which he possesses attribute bundle $(A_i(x), R_i)$, to the state of affairs in which he possess $(A_j(y), R_j)$. Consider, now, that there are attributes which the spectator essentially possesses, and also attributes that he essentially lacks.[14] For short, call the latter EL_k properties. EL_k properties are properties that k cannot possibly possess – properties that k does not have in any possible world where he exists.

The essential-attribute problem is this: if $A_i(x)$, or $A_j(y)$, or both include EL_k attributes, the empathy-based conception asks k to compare *impossible* states of affairs.[15]

For example, imagine that the spectator is a woman, Sue Dean, currently living and born in 1980. Sue is told about a possible life that Cleopatra might have led, (x; Cleopatra), and a possible life that Shakespeare might have led, (y; Shakespeare). She is also told that Cleopatra had tastes $R_{\text{Cleopatra}}$, and that Shakespeare had tastes $R_{\text{Shakespeare}}$.

Sue is then asked to formulate her extended preferences between the two lives. In x, Cleopatra is described as having various (non-taste) attributes: she was born in the first century BC; she was female; she was beautiful, rich and powerful; she was the last ruler of the Ptolemaic Dynasty of Egypt; she had affairs with Julius Caesar and Marc Antony; she was deposed and imprisoned for life by the Romans after Octavian defeated Antony at the battle of Actium.[16] In y, Shakespeare's attributes include: he was born in the 16th century AD; he was male; he was ugly and,

[13] I take no position about the role of empathy-based extended preferences except as a basis for constructing $\succcurlyeq^{\text{WB}}$ and $\succcurlyeq^{\text{DIFF}}$.

[14] For a general discussion of essential properties, see Lowe (2002: ch. 6); Mackie (2006); Roca-Royes (2011). Characterizing the essential properties of *persons* is one aspect of the vast literature on personal identity. See sources cited in Adler (2012: 409).

[15] Since tastes are simply rankings of outcomes or choices, it is hard to see how the attribute of having R_i or R_j could be impossible for k. Someone's essential attributes might be something like her DNA, or the circumstances of her birth, but not her preferences. I thus focus on the possibility that some of the *non-taste* attributes in $A_i(x)$, or $A_j(y)$, or both include EL_k attributes. Including taste attributes as essential simply compounds the difficulties for the empathy approach. (Perhaps a certain variant of Psychological Essentialism, see below, *would* count some taste attributes as essential.)

[16] Remember that x is a possible outcome, not necessarily actual. Although Cleopatra in fact died at her own hand after Antony's defeat, a life in which the Romans imprison her is also possible.

although respected as a playwright in his own time, not famous until after his death; he struggled financially; he was married to Anne Hathaway, a loyal and loving wife, for his entire adult life; he wrote some of the greatest plays ever written, including *Hamlet*, *Macbeth* and *King Lear*.

On the empathy-based conception, the spectator, Sue Dean, is meant to rank these two histories by determining her preference as between the following two states of affairs: (1) Sue Dean is born in the first century BC; is female, beautiful, rich and powerful; is the last ruler of the Ptolemaic Dynasty; has affairs with Julius Caesar and Marc Antony; is deposed and imprisoned for life by the Romans after the battle of Actium; and has tastes $R_{\text{Cleopatra}}$; (2) Sue Dean is born in the 16th century AD; is an ugly male playwright, respected but not famous, who struggles financially but is happily married to Anne Hathaway; writes great plays including *Hamlet*, *Macbeth* and *King Lear*; and has tastes $R_{\text{Shakespeare}}$.

Neither state of affairs (1) nor (2) seems possible. To be sure, Sue Dean's gender, physical appearance, occupation, political power, social and financial status, and whether she marries or has intimate affiliations are *contingent* properties of her. So it *is* possible for Sue Dean to have been male or female, ugly or beautiful, a playwright or a monarch, imprisoned or free, rich or poor, married or not. But Sue Dean necessarily was born in 1980 or thereabouts, and thus necessarily was *not* born in the first century BC, or in the 16th century AD. Plausibly, the precise or at least rough timing of someone's birth *is* one of her essential properties.[17]

Moreover (on the premise that birth timing is an essential property), it is not possible for Sue Dean to have been a pharaoh of the Ptolemaic Dynasty (a particular kingdom that lasted from the 4th to the 1st centuries BC); nor to have written *Hamlet*, *Macbeth* or *King Lear* (plays that were written at the beginning of the 17th century AD); nor to have been the lover of Julius Caesar or Marc Antony, nor the spouse of Anne Hathaway (particular individuals all of whom died centuries or millennia before Sue Dean's birth date of 1980).

One might deny that birth timing is an essential property – but this hardly meets the challenge. *Which* properties are essential to human persons is a matter of philosophical dispute; but surely *some* are. Consider, then, *any* case in which i in x has some EL_k property. Then for k to formulate an extended preference as between (x; i) and some other history – on the empathy account – involves k's preferring an impossible state of affairs.

affairs in which k has all the properties of i in x *except* the EL_k properties of i in x, plus tastes R_i, to a state of affairs in which k has all the properties of j in y *except* the EL_k properties of j in y, plus tastes R_j. This means that Sue Dean compares (x; Cleopatra) to (y

attributes are (some of) her essential properties. Perhaps it is impossible for Sue Dean to have had a radically different series of perceptions, memories and experiences ('psychological history') than those of the actual Sue Dean. Call this Psychological Essentialism.[18] If Psychological Essentialism is true, the severance strategy clearly fails. Consider asking Sue Dean to compare (x^*; Cleopatra) and (y^*; Shakespeare). Outcomes x^* and y^* are the same as x and y above, respectively, but supplemented with rich detail about Cleopatra's and Shakespeare's psychological histories. x^+ and y^+ are outcomes in which Sue has the contingent non-psychological attributes of Cleopatra and Shakespeare in x^* and y^*, respectively, and the psychological attributes of theirs which Sue Dean can possibly possess (given Psychological Essentialism). If many of Sue Dean's psychological attributes in x^+ and y^+ are different from Cleopatra's and Shakespeare's in x^* and y^*, then Sue's preference as between x^+ and y^+ doesn't tell us much at all about how (x^*; Cleopatra) and (y^*; Shakespeare) compare in terms of well-being.

Finally, even if it *is* true (as an empirical rather than conceptual feature of well-being) that well-being depends solely on individuals' contingent properties, the severance strategy is problematic. Remember that this strategy asks the spectator k to compare (x; i) and (y; j) by comparing a state of affairs in which k has all of the properties of i in x except the EL_k properties, plus tastes R_i, to a state of affairs in which k has all of the properties of j in y except the EL_k properties, plus tastes R_j. Many of the subjects' properties in these outcomes might be properties which the spectator *could* have, but which the spectator finds it very difficult to imagine having. For example, someone's *gender*, unlike his or her DNA, is very plausibly a contingent attribute, since dependent on social construction and macroscopic physical features; but many individuals, still, find it difficult to imagine changing their gender. Thus, in the Shakespeare/Cleopatra case, Sue Dean may hardly be able to imagine the state of affairs in which she has Shakespeare's occupation, income level, etc., and in which she is *male* (Shakespeare's gender) – even though none of these are $EL_{\text{Sue Dean}}$ attributes. Why does this matter? Presumably the extended preferences relevant to well-being satisfy standard idealizing conditions; these are the preferences that emerge and remain stable after deliberation with good information. If Sue can barely imagine one or both states of affairs that the severance strategy asks her to consider in comparing (x; Cleopatra) and (y; Shakespeare) – the state in which she has all of Cleopatra's properties except the $EL_{\text{Sue Dean}}$ properties, and the state in which she has all of Shakespeare's properties except the $EL_{\text{sue Dean}}$

[18] Psychological Essentialism would be a plausible consequence of the view that each particular person is a particular psychological entity, rather than a particular human being ('animalism'). See Brown (2003).

properties – then she can hardly form a stable preference between these states.[19]

(2) *The* De Se *Strategy.* Philosophers of language and mind distinguish between '*de se*' and '*de re*' beliefs, which can come apart under conditions of imperfect self-knowledge. *De se* beliefs are expressed by sentences with 'I' in the subject position; *de re* beliefs, by sentences with the proper name of some person in the subject position.[20]

Imagine that John Perry wakes up in the middle of the night in a dark room, after an operation that has temporarily caused amnesia. He is bewildered, having forgotten everything about his life before the operation. He sees a TV in the corner of the room. At that moment, John Perry has the *de se* beliefs, 'I am seeing a TV' and 'I am in a dark room', but not the *de re* beliefs 'John Perry is seeing a TV' or 'John Perry is in a dark room'. Although John Perry is aware of himself saying or thinking the word 'I', he does not have the *de re* beliefs just mentioned because – by virtue of his amnesia – John Perry does not realize that the particular person who is the referent for the word 'I' is the very same person who is the referent for the proper name 'John Perry'.

Related to a distinction between *de se* and *de re* beliefs is a distinction between *de se* and *de re imagining*.[21] It is impossible for John Perry (the 20th century philosopher) to be the very same person as the historical figure Napoleon Bonaparte. And it is presumably difficult for John Perry, or anyone else, even to imagine the state of affairs that 'John Perry is Napoleon Bonaparte'. But John Perry may find it relatively easy to imagine the *de se* proposition, 'I am Napoleon Bonaparte'.

[19] Mongin (2001) is sensitive to the difficulties regarding personal identity that arise in asking the spectator to assume the subject's standpoint. See, e.g. 2001: 161–162. He proposes to avoid those difficulties via the following construal of extended preferences: spectator k has an extended preference for $(x; i)$ over $(y; j)$ just in case k prefers that outcome x obtain and that k be 'under the influence of the factors determining [i]'s preferences', rather than that outcome y obtain and that k be under the influence of the factors determining j's preferences (2001: 160).

On the supposition that an individual's preferences, and the factors determining them, are not essential attributes of hers (see 2001: 156–157), Mongin's proposal is a version of the severance strategy. An objection to this proposal is that the well-being relevance of a given contingent attribute should not hinge on whether the attribute is preference-determining. For example, assume that there is some mild health impairment which has no causal influence on individuals' preferences. (Individuals prefer not to have the impairment; but having it does not change how individuals rank health states, health-income bundles, etc.). Then, on Mongin's proposal, k should ignore whether i or j has the impairment in determining his extended preferences as between $(x; i)$ and $(y; j)$. That seems arbitrary, *if* extended preferences are meant to provide an analysis of well-being.

[20] See, e.g. Lewis (1979), Perry (1979), Recanati (2007), Feit (2008).

[21] See Williams (1973: ch. 3), Vendler (1976), Reynolds (1989); Walton (1990: 28–35), Gordon (1995), Velleman (1996), Nichols (2008), Ninan (2009).

Why this difference? The relative ease of *de se* versus *de re* imagining is a matter of ongoing discussion. Various philosophers have suggested that a *de se* imagining occurs when someone, perhaps with accurate beliefs about his actual attributes, engages in an act of 'pretense' or 'make believe' that shifts the 'I' concept so that it refers to someone else.

> Many actors speak of transforming themselves, of becoming the characters they play. This typically involves an imaginative shift in the reference of indexicals. There is a character in the dramatis personae who becomes in your imagination the referent of the pronoun 'I', and his time and place become the referents of 'now' and 'here'. One very perceptive actor, Ray McAnally … describes his thoughts while filming a scene in which he plays a future British prime minister: 'I had a very interesting moment in 10 Downing Street, surrounded by pictures of all the previous Prime Ministers *and me in the middle of it*. And I realized it was true, I *was* the Prime Minister …' Of course he is in fact pretending to be the prime minister, only he is doing it so well that he is oblivious to that fact. (By the way, he is not pretending that the following counteridentical is true: 'Ray McAnally is the prime minister'. Analogously, although I may pretend on July first that 'It is now New Year's Eve', I am not pretending that *July first* is New Year's Eve.)
>
> …
>
> The imaginative shift in the reference of indexicals reflects a much deeper, more important shift. Many of our tendencies to action or emotion appear to be specially keyed to an egocentric map. What triggers the action or emotion is the lion coming toward *me*, the meeting I am supposed to be at *now*, the insult directed to *me*, the award given to *my* child. …
>
> What the actor can do is to recenter his egocentric map. Think of one of those transparent overlays on a map, with concentric circles showing the distance from any point you center it on. The actor can shift his egocentric overlay until it is centered on a particular character, place, and time, rather than on, say, Ray McAnally and his place and time.[22]

With the *de re*/*de se* distinction in hand, one might try to specify the empathy-based account of extended preferences in *de se* fashion, so as to avoid the essential-attribute problem. In comparing $(x; i)$ and $(y; j)$, the spectator k is not meant to compare the state of affairs in which the particular person referred to by the proper name of k has all the attributes of i in x, to the state of affairs in which that particular person has all the attributes of j in y. Rather, k is meant to consider her preference as between scenarios specified with 'I', namely 'I have all the attributes of person i in outcome x' versus 'I have all the attributes of person j in outcome y'. For example, in the Shakespeare/Cleopatra case, Sue Dean is not meant to wrap her head around the impossibility, 'Sue Dean is born

[22] Gordon (1995: 733–734). For similar views, see Velleman (1996), Nichols (2008).

in the first century BC, Sue Dean is a Ptolemaic queen, lover of Antony and Julius Caesar', etc., but rather to think of the history (x; Cleopatra) by entertaining a *de se* scenario in which 'I am born in the first century BC, I am a Ptolematic queen, I am the lover of Antony and Caesar', etc. Similarly, Sue Dean is meant to think of the history (y; Shakespeare) by entertaining a *de se* scenario in which 'I am born in the 16th century, I am married to Anne Hathaway, I write *Hamlet, King Lear* and *Macbeth*', etc.

But if Sue Dean knows who she is (born in 1980, etc.), then she can only see the *de se* scenarios corresponding to (x; Cleopatra) and (y; Shakespeare) as *fictions* – pretence, make-believe, acts of imagination. It is hard to see why a spectator's preferences regarding fictions corresponding in some way to genuine possible lives have much to tell us about the well-being associated with those lives. In general, someone's ordinary preferences regarding some outcome might be quite different from what she prefers if the outcome is taken as fiction. For example, Sue Dean might prefer the outcome in which Sue Dean lives a contented but uneventful life as a married accountant, dying at the age of 80, to one in which Sue Dean becomes a drug addict who loses her job, destroys her marriage, becomes homeless, and is killed at age 50 in a violent street encounter at just the point where she has resolved to beat her addiction. But Sue might prefer the latter scenario, taken as fiction, because of its dramatic interest.[23]

Alternatively, the empathy-based conception might ask for the spectator to be placed under a kind of 'veil of ignorance' about her own attributes, as in the example of the amnesiac John Perry earlier. Spectator k is deprived of information about her name, place of birth, etc. Under that condition, she is asked to develop a preference as between (x; i) and (y; j) by considering, first, the outcome in which 'I have all the attributes of i in x' and second, the outcome in which 'I have all the attributes of j in y'. Because the spectator does not know the referent for 'I', she can entertain these outcomes as real possibilities, not fictional scenarios.

But this veil-of-ignorance construal of the empathy conception is in serious tension with the truism that well-being-relevant preferences must satisfy various idealizing conditions, including good information. Such information includes not merely information about the outcomes being considered, but also self-knowledge on the part of the preference-holder, for example about whether her preferences were the result of parental indoctrination, adaptation to adverse circumstances, etc.[24] By depriving Sue Dean of facts about her birth date (and other autobiographical data) when she considers 'I am born in the 1st century BC, I am a queen of the

[23] Some philosophers have responded to the various differences between ordinary desires, and desires about fictional scenarios, by arguing that the latter are not really desires at all. (For a critical overview, see Kind 2011). In any event, the differences are substantial.

[24] See Brandt (1998).

Ptolemies, . . . ', we allow her to think of these scenarios as real possibilities (not fictions), but also thereby suction away some of the crucial data that Sue would need for her preferences to be well-informed.[25]

The term 'veil-of-ignorance' is sometimes used to denote *any* construction that translates outcomes into the bundles of histories therein, and appeals to spectators' preferences (extended preferences) over these histories for some normative purpose, e.g. to determine what justice requires. My objection here is not to the veil-of-ignorance in this generic sense, but much more specifically to the proposal that extended preferences involve a condition of ignorance on the spectator about who she is.

(3) *The Psychological Strategy.* Perhaps in comparing $(x; i)$ and $(y; j)$, the spectator is meant to compare the *experience of being* the subjects in the two outcomes. In other words, the spectator is not supposed to consider the acquisition of non-mental attributes, such as Cleopatra's or Shakespeare's birth dates or their having lovers or spouses with particular identities, but instead to compare having all of the non-taste *psychological* attributes (feelings, perceptions, beliefs, emotions) of i in x, plus tastes R_i, to having all of the non-taste psychological attributes of j in y, plus tastes R_j.

However, as already noted, asking the spectator to consider a state of affairs in which she has psychological attributes different from her own *might* be asking her to consider an impossible state of affairs – if Psychological Essentialism is true. A quite different problem is this: a subject's psychological attributes are not the sole, intrinsic determinants of her well-being. Non-experiential attributes also matter, as vividly illustrated by Robert Nozick's 'experience machine'.[26] $(y;$ Shakespeare) is much better for well-being than $(y^{++};$ Fakespeare), with Fakespeare in y^{++} someone who has the very same experiences as Shakespeare in y (including the very same *beliefs* about the artistic quality of his plays), but with the histories differentiated by the fact that Shakespeare actually produces great art while Fakespeare's plays are junk.

3.2. The 'Wrong Kind of Preference' Problem

In *Reasons and Persons*, Derek Parfit notes that the occurrence of one outcome rather than a second may make no difference to someone's well-being, even though the person prefers the first outcome, and even though this preference satisfies normal idealizing conditions (good information, rationality, deliberation). Parfit illustrates the problem with the following example, the 'stranger' case:

[25] See also Voorhoeve (2014), pointing to difficulties that may arise for the veil-of-ignorance approach if some essential attributes are psychological attributes.

[26] Nozick (1974: 42–44)

> Suppose that I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I strongly want this stranger to be cured. We never meet again. Later, unknown to me, this stranger is cured. On the [actual preference theory of well-being], this event is good for me, and makes my life go better. This is not plausible. We should reject this theory.[27]

Note that Parfit, in this story, supposes himself to have a preference the content of which is (1) 'that the stranger be cured', and *not* a preference with the content (2) 'that the stranger be cured and I experience happiness at his being cured', or (3) 'that I believe the stranger to be cured'. Parfit, if motivated by a benevolent concern for the stranger, could easily develop a well-informed preference with the content of (1); such a preference would be fulfilled by the sheer fact of the stranger's being cured, *whether or not Parfit learns about that fact, or feels good about it*; but the sheer fact of the stranger's being cured does not, without more, improve Parfit's own well-being. This is the point of the example.[28]

Many other philosophers concur with Parfit. The 'stranger' case has been widely discussed in the contemporary philosophical literature on well-being, and the lesson it suggests has become virtually a truism in that literature. For example, Stephen Darwall writes:

> There are many things I rationally take an interest in, such as the survival of the planet and the happiness of my children long after I am dead, that will make no contribution to my welfare. A person may have rational *interests* that go well beyond what is for her good or *in her interest*. A person's good – what benefits her or advances her welfare – is different from what is good from her point of view or standpoint. The latter is the perspective of what she herself cares about, whereas her own good is what is desirable from the perspective of someone (perhaps she herself) who cares for her.[29]

[27] Parfit (1987: 494).

[28] An anonymous referee has suggested that the preference-based view of well-being can be rescued from Parfit's objection by analysing well-being as the combination of preference-realization plus the knowledge thereof. However, if my preference is morally motivated, it is hard to see why the preference's realization *plus* my justified true belief in this occurrence benefits me (without more). Perhaps this shows that the knowledge condition needs to be strengthened to enjoyment or happiness. However, there will be plausible cases in which the realization of my preference (for example, a preference for success at long-standing goals, for the flourishing of my children, for the fidelity of my spouse, or for the respect of my peers) can benefit me even though I never feel happiness about the fact that the preference has been realized – indeed, arguably, even if I never learn of that fact.

For a discussion of the possible restrictions on preferences to ensure their well-being relevance, see Adler (2012: 174–181). However Parfit's example is handled, it shows the need for *some* such restriction.

[29] Darwall (2002: 53).

Tim Scanlon writes:

> [Desire theories of well-being are] open to serious objection. The most general view of this kind – it might be called the unrestricted actual-desire theory – holds that a person's well-being is measured by the degree to which all the person's actual desires are satisfied. Since one can have a desire about almost anything, this makes an implausibly broad range of considerations count as determinants of a person's well-being. Someone might have a desire about the chemical composition of some star, about whether blue was Napoleon's favorite color, or about whether Julius Caesar was an honest man. But it would be odd to suggest that the well-being of a person who has such desires is affected by these facts themselves (as opposed to the pleasure he or she derives from having certain beliefs about them). The fact that some distant star is made up of the elements I would like it to be does not seem to make my life better (assuming that I am not an astronomer whose life work has been devoted to a theory that would be confirmed or refuted by this fact).[30]

Richard Arneson writes: '[N]ot all of an agent's desires plausibly bear on her well-being. I might listen to a televised plea for famine relief, and form the desire to aid distant starving strangers, without myself thinking (and without its being plausible for anyone else to think) that the fulfilment of this desire would in any way make my life go better'.[31]

I will term the problem described by Parfit *et al*. the 'wrong kind of preference' problem. As already explained, an outcome-preference is a ranking of outcomes, on the part of some person, that constitutes a choice-disposition on his part: what choices the preference-holder is disposed to make depends upon where their possible outcomes are located in this ranking. However, nothing in the sheer existence of this disposition guarantees that the choice-motivating features of outcomes are connected to the preference-holder's well-being. If someone has a preference for outcome x over outcome y (perhaps a fully informed, rational, and stable preference) but this is the 'wrong kind of preference', he need not be better off in x than y.

Moral preferences provide a particularly clear illustration of the problem. Jim might judge that x is morally better than y, and come to prefer x, even though his own interests favour y. But the problem is more general. Jim might be motivated to pursue x in virtue of considerations that are neither moral, nor connected to his interests (for example, aesthetic considerations, or perceived compliance with the will of God).

[30] Scanlon (1998: 113–114).

[31] Arneson (1999: 124). See also Overvold (1980, 1982, 1984), Gibbard (1986), Griffin (1986: chs. 1–2), Kagan (1992), Sumner (1996: ch. 5), Bernstein (1998); Brandt (1998: ch. 17); Hausman and McPherson (2009).

The 'wrong kind of preference' problem clearly undermines an analysis of intrapersonal well-being comparisons in terms of unrestricted outcome-preferences. But it also – less obviously – undermines the proposal to analyse intrapersonal or interpersonal comparisons in terms of extended preferences in the empathy-based sense.

Why? Let us ignore, for the moment, the essential-attribute problem and assume that spectator k can engage in the thought exercise of acquiring all of the attributes of various subjects. In ranking $(x; i)$ versus $(y; j)$, the spectator is asking whether she would prefer to have an attribute bundle $(A_i(x), R_i)$ consisting of individual i's attributes in x, plus tastes R_i, as against a bundle $(A_j(y), R_j)$ consisting of individual j's attributes in y, plus tastes R_j. Since these two bundles consist of the subjects' attributes (both non-taste attributes and tastes), how *can* the spectator's preference between the two diverge from the subjects' well-being?

Note, to begin, that a given subject's attributes include her *relational* as well as *non-relational* attributes. Thus other subjects' attributes are 'built into' the bundle of a given subject, as relational attributes of hers. But, clearly, spectators can have non-self-interested preferences for attribute bundles of this sort that don't track the well-being associated with the bundles.

To see this in a simple case, imagine that there are five people in the population: i, j, k, l, m. Outcome x is one in which individuals' incomes range in \$20 000 increments from \$20 000 to \$100 000. Individual i has income \$20 000, and individual j has income \$100 000; while the other three have, respectively, incomes of \$40 000, \$60 000 and \$80 000. Individual i has tastes R_i, etc.

Then $(A_i(x), R_i)$ is the bundle (having an income of \$20 000; having tastes R_i; being part of a population of five individuals where the other incomes are \$40 000, \$60 000, \$80 000, \$100 000 and where the other individuals have tastes R_k, R_l, R_m and R_j).[32] And $(A_j(x), R_j)$ is the bundle (having an income of \$100 000; having tastes R_j; being part of a population of five individuals where the other incomes are \$20 000, \$40 000, \$60 000, \$80 000 and where the other tastes are R_i, R_k, R_l and R_m).

Imagine, now, that k is an *impartial* spectator. In the exercise of ranking hybrid bundles, she assumes an attitude, not of self-interest, but rather of impartiality between her interests and everyone else's. If so k will be *indifferent* between the bundles $(A_i(x), R_i)$ and $(A_j(x), R_j)$. She doesn't care, from this impartial perspective, whether *she* is the one with \$20 000 and particular tastes in a given population distribution of income and tastes, or *she* is the one with \$100 000 and particular tastes in the very same

[32] Although the bundle $(A_i(x), R_i)$ does not itself describe the tastes of the other subjects, that information *is* part of the profile of tastes, and is available to spectator k in determining which bundle she prefers.

distribution of income and tastes. But, of course, $(x; i)$ and $(x; j)$ are *not* equally good for well-being. It is worse for well-being, ceteris paribus, to be the person with the lowest income in a given distribution of income, rather than the person with the highest (at least if R_i and R_j both include a taste for more income rather than less).

The 'wrong kind of preference' problem can arise even apart from the inclusion of the population distribution of attributes in each attribute bundle as a relational attribute of the subject who possesses that bundle. Consider the following sort of case. Subjects i and j, in outcomes x and y respectively, both engage in wrongdoing. But only i is punished (by prolonged incarceration), while j escapes punishment. Thus $A_i(x)$ includes the attributes of engaging in wrongdoing and being punished, while $A_j(y)$ includes the attributes of engaging in wrongdoing and escaping punishment. Let us also suppose that R_i and R_j are tastes of the ordinary sort, including a moral preference that wrongdoing be punished, and a self-interested preference not to be incarcerated.

Now, spectator k himself also has a moral preference that wrongdoing be punished. If motivated by this preference, he ranks $(A_i(x), R_i)$ over $(A_j(y), R_j)$. He has this ranking even though $(x; i)$ may well be worse for well-being than $(y; j)$, since i suffers the harm of prolonged incarceration. Spectator k would *morally* prefer to be (1) a wrongdoer who self-interestedly prefers not to be incarcerated, but is in fact incarcerated as punishment for the wrongdoing, as opposed to (2) a wrongdoer who self-interestedly prefers not to be incarcerated, and escapes punishment.

4. A SYMPATHY-BASED CONCEPTION OF EXTENDED PREFERENCES

I propose a sympathy-based conception of extended preferences.[33] The generic set-up remains the same one I used above, in discussing the empathy-based conception. $\mathbf{O}$ is the set of outcomes, $\mathbf{N}$ a finite set of individuals (each of whom exists in all outcomes), and $\mathbf{H} = \mathbf{O} \times \mathbf{N}$ the set of all histories $\{(x; i)\}$. $\mathbf{K}$ is the set of spectators (for simplicity, assume that $\mathbf{K} = \mathbf{N}$). The outcomes in $\mathbf{O}$ are arbitrarily detailed specifications of possible worlds, but do not specify individuals' preferences. Let $R = (R_1, R_2, \ldots, R_N)$ be a possible profile of 'tastes' (outcome and choice preferences) on the part of individuals 1, 2 … N. For any given R, each spectator k has *extended* preferences $\succcurlyeq^k(R)$ (or, for short, $\succcurlyeq^k$). Each $\succcurlyeq^k$ is a quasiordering of $\mathbf{H}$.

The difference between the sympathy-based and empathy-based conception of extended preferences lies not in this set-up (which is quite

[33] This Part builds upon, and refines, the analysis of Adler (2012: ch. 3). See also Adler (forthcoming), further developing the approach set forth here, with a fuller treatment of comparisons of well-being differences.

general), but in the substantive content of $\succcurlyeq^k$. In explicating the sympathy approach, I will distinguish between 'intrapersonal' and 'interpersonal' extended preferences. A spectator's extended preference as between two histories with the same subject is an 'intrapersonal' extended preference. A spectator's extended preference as between two histories with different subjects is an 'interpersonal' extended preference.

The Sympathy-Based Conception of Extended Preferences

(1) *Intrapersonal case*: $(x;i)\succcurlyeq^k(y;i)$ iff spectator k, under a condition of unreserved[34] sympathy for subject i, and given the profile of tastes in R, weakly prefers[35] outcome x to outcome y.

(2) *Interpersonal case*: $(x;i)\succcurlyeq^k(y;j)$, with i and j distinct, iff spectator k makes the judgement that i in x is at least as well off as j in y, given the profile of tastes in R.

What exactly is the connection between extended preferences, in the sympathy-based sense, and $\succcurlyeq^{WB}$? Part 5 will address this question for the general case, where Convergence may fail. At a minimum, however, if Convergence holds true (everyone has the same extended preferences, i.e. $\succcurlyeq^1 = \succcurlyeq^2 = \ldots = \succcurlyeq^N = \succcurlyeq$), then the sympathy-based conception says: $(x;i)\succcurlyeq^{WB}(y;j)$, the two outcomes and subjects the same or different, iff $(x;i)\succcurlyeq(y;j)$.

What is sympathy? While I *empathize* with you by projecting myself into your position, I *sympathize* with you by adopting an attitude of care and concern for you. Darwall has emphasized the distinction between empathy and sympathy, and the connection of sympathy (not empathy) to well-being.

> [Sympathy] is a feeling or emotion that (i) responds to some apparent obstacle to an individual's welfare, (ii) has that individual himself as object, and (iii) involves concern for him, and thus for his welfare, for his sake. Seeing the child on the verge of falling [down the well], one is concerned for his safety, not just for its (his safety's) sake, but for *his* sake. One is concerned for *him*. Sympathy for the child is a way of caring for (and about) him.
>
> Sympathy differs in this respect from several distinct psychological phenomena usually collected under the term 'empathy' that may involve no such concern. What these phenomena have in common is their involving feelings that are 'congruent with the other's emotional state or condition', as one psychologist puts it. Here it is the way things seem from the other's standpoint that is salient, in this case, the prospect of falling down the well.

[34] That is, k's sympathy is wholly directed on i, and no one else.

[35] Since $\succcurlyeq^k$ itself is weak (with 'strict' extended preference $\succ^k$ and extended indifference $\sim^k$ derived in the standard fashion from $\succcurlyeq^k$), it should be defined in terms of a weak outcome preference.

> Empathy consists in feeling what one imagines he feels, or perhaps should feel (fear, say), or in some imagined copy of these feelings, whether one comes thereby to be concerned for the child or not. Empathy can be followed by the indifference of pure observation or even the cruelty of sadism. It all depends on why one is interested in the other's perspective. Sympathy, on the other hand, is felt, not as from the child's perspective, but as from the perspective of 'one caring'.
>
> …
>
> [T]he concern we experience for people in sympathy is central, not just to seeing individuals and their well-being as having categorical importance, but also to the very concept of well-being or personal good. … It is because we can take up the standpoint of one caring toward ourselves and others and ask what it makes sense to want from that point of view that we have a need for the concept [of well-being].[36]

The sympathy-based conception of extended preferences readily avoids the essential-attribute problem and the wrong-kind-of-preference problem. Consider first the intrapersonal case. In ranking $(x; i)$ and $(y; i)$, the spectator does *not* engage in the thought experiment of acquiring i's attributes in x and in y. Instead, his extended preference is reduced to a preference of the ordinary sort – an *outcome* preference – with an attitudinal restriction, namely a ranking of those outcomes while in the grip of an attitude of unreserved sympathy for the subject, i.

The fact that some of the subject's attributes are EL_k attributes in no way frustrates this exercise. For example, Sue Dean, with an attitude of concern for Cleopatra, can ask herself whether she prefers the outcome x' in which Cleopatra dies after Actium from a self-imposed snakebite, or instead the outcome x in which Cleopatra is imprisoned for life by the Romans. Because x' and x are not outcomes in which Sue Dean herself is supposed to be born in the first century BC, to be the lover of Marc Antony, etc., the outcomes are possible. And the fact that Sue Dean and Cleopatra are distinct individuals, each with some essential attributes that the other lacks, obviously does not prevent Sue from being sympathetic to Cleopatra. Sympathy *often* takes the form of being targeted upon some person distinct from the sympathizer.

The 'wrong kind of preference problem', remember, was that the sheer possession of a preference on someone's part for x over y does not guarantee that x is better than y (for the holder of the preference, or for anyone else), since preferences can be motivated by a wide range of considerations (moral, aesthetic, etc.). But the sympathy-based conception avoids this problem, in the intrapersonal case, by virtue of the connection between sympathy and well-being emphasized by

[36] Darwall (2002: 51, 72). See also Fontaine (1997), Darwall (1998), Eisenberg (2000).

Darwall – more precisely, the connection between having an attitude of sympathy, and being motivated to pursue what you believe lies in the interests of the sympathy target. Where spectator *k* holds an attitude of

for (x; k) over (y; k). The sympathy-based conception analyses this as an outcome preference on k's part (a preference for x over y) under a condition of wholehearted self-sympathy. 'Self-sympathy' means an attitude of care and concern for the very same person who has the attitude. The person who is the target of the sympathizer's attitude is, now, the sympathizer himself. 'Self-sympathy' is, perhaps, an unfamiliar term. Here's a synonym: 'self-interest'. In short, the spectator's extended preferences for the case where he is subject are nothing other than his self-interested outcome preferences.

The observation can be inverted. Wholehearted sympathy is a *generalization* of self-interest – an attitude of interest *in* some particular person, be it the holder of the attitude (self-interest), or someone else. The spectator's intrapersonal extended preferences are just rankings of outcomes with this generalized self-interest directed at the appropriate person, the subject.

Consider now the interpersonal case. It is tempting to propose: k has an extended preference for (x; i) over (y; j), i and j distinct, iff k prefers x over y under a condition of unreserved sympathy with 'the subject'. However, this proposal fails. There is no one person to be the target of the spectator's sympathy in the interpersonal case. And holding simultaneous attitudes of unreserved sympathy with *two* different subjects is psychologically impossible. The spectator Robert cannot, at the same time, care *only* about Maurice and *only* about Jean.[40] Thus, it will not work to analyse Robert's extended preference for (x; Maurice) over (y; Jean) as a preference for x over y while simultaneously being unreservedly sympathetic to both Maurice and Jean. Asking what Robert would prefer with this impossible combination of attitudes would be like asking what he prefers when simultaneously calm and panicky.

To be sure, Robert's attitudes of sympathy can change. He can, at one moment, feel unreservedly sympathetic to Maurice and, later, unreservedly sympathetic to Jean. But a preference for one possible outcome over a second is a synchronic relation between a preferrer and the two possibilities, each simultaneously presented to his mind via some mental representation. For Robert to stand in that synchronic relation to outcomes x and y, and at that one moment in time to be wholeheartedly sympathetic to two different people, is impossible.

Thus, in the interpersonal case, the account of extended preferences presented here asks the spectator to make a judgement about well-being. However, this analysis is a close cousin to the intrapersonal analysis. Although sympathy does not necessitate explicit well-being judgements – I can hold an attitude of sympathy for you without explicitly thinking

[40] The same problem would arise if the requirement of unreserved sympathy were weakened to predominant sympathy. See above note 39.

about your well-being (or so it is plausible to believe) – sympathy is responsive to such judgements, as already discussed. If I am sympathetic to you, and come to believe/judge that some course of action benefits you, then I am motivated to pursue that course of action. The *affective* component of the intrapersonal analysis (wholehearted concern for the sympathy target) cannot be transposed to the interpersonal case, but the *valuational* component (the judgements about the target's well-being that motivate the sympathizer) can be. And that is what my account does. Indeed, the very same implicit or explicit views about the nature of well-being that motivate the sympathizer, in the intrapersonal case, will determine his well-being ranking when he makes explicit well-being judgements in the interpersonal case. For this reason, I describe the entire account of extended preferences, both (1) and (2), as the 'sympathy-based conception'.[41]

Note that the interpersonal analysis, like its intrapersonal cousin, avoids the 'essential attribute' and 'wrong kind of preference' objections. A spectator can take account of all of the subjects' attributes, including their essential attributes, in arriving at his well-being judgements. A judgement made by one spectator to the effect that i in x is at least as well off as j in y is evidence that $(x;i) \succcurlyeq^{\mathrm{WB}} (y;j)$, at least if the spectator is suitably idealized. And such a judgement by all such spectators is substantial evidence that $(x;i) \succcurlyeq^{\mathrm{WB}} (y;j)$ – indeed, it arguably guarantees that $(x;i) \succcurlyeq^{\mathrm{WB}} (y;j)$.

What exactly is the relation between $\succcurlyeq^k$ and the profile R? Remember that extended preferences (on both the empathy- and sympathy-based account) are potentially dependent on individual tastes. But what does such dependency involve, on the sympathy-based view?

First, there is a logical link between $\succcurlyeq^k$ and R_k. Given the definition of intrapersonal extended preferences, it follows that spectator k has a weak extended preference for $(x;k)$ over $(y;k)$ iff R_k is such that k has a self-interested weak preference for outcome x over outcome y. Second, there are empirical (not logical) links between $\succcurlyeq^k$ and each R_i, $i \neq k$. In developing his intrapersonal extended preferences for histories belonging to subjects other than himself, and in making across-subject comparisons, the spectator k is permitted to take account of everyone else's tastes.

[41] Why not adopt the simpler strategy of defining both intra- and interpersonal extended preferences in terms of well-being judgements – omitting any reference to sympathy? The spectator's extended preferences are genuinely *preferences* only in virtue of their motivational connection to his choices. The account I have offered preserves that connection: directly, in the intrapersonal case (the sympathetic spectator's ranking of outcomes, like other genuine outcome preferences, helps motivate his choices) and indirectly, in the interpersonal case (in virtue of those judgements flowing from the very same views about well-being that figure in the intrapersonal case). A single, purely valuational account would sever this nexus to choice.

Whether k does so depends upon his own tastes and values. For example, k might make the judgement that $(x; i)$ is better for well-being than $(y; j)$ given R_i and R_j, but not given R_i^* and R_j^*.

In the next Part, I will discuss how to derive $\succcurlyeq^{WB}$ from an array of extended preferences $(\succcurlyeq^1, \ldots, \succcurlyeq^k, \ldots, \succcurlyeq^N)$, where Convergence fails. But we should first consider three potential criticisms of the sympathy-based conception of extended preferences, as well as the question of representing $\succcurlyeq^k$ via utility functions.

One criticism is that the sympathy-based conception, used as an analysis of well-being, is *circular*. In the interpersonal case, the circularity is patent. In that case, the view says: Given Convergence, $(x; i) \succcurlyeq^{WB} (y; j)$ iff everyone judges that i in x is at least as well off as j in y. So well-being is 'analysed' in terms of well-being judgements – an obvious circularity. In the intrapersonal case, the circularity is more subtle, but still real. Even if sympathy does not require explicit thinking about the target's well-being, the attitude of sympathy has important connections to well-being, described in the previous paragraphs. But (so the critique goes) these connections are not *happenstance*. It is not as if sympathy is some independently specified attitude which, as it happens, motivates the holder to act in line with his judgements about the target's well-being, if he makes such judgements. Rather, this is a defining feature of the attitude: sympathy has conceptual, not just empirical, connections to well-being judgements and beliefs.[42] Insofar as the concept of well-being is used to identify the attitude of sympathy, which in turn is used – by the view tendered here – to analyse intrapersonal extended preferences and, thereby, intrapersonal well-being comparisons, the view is circular.

However, the circularities just described are not *vicious* circularities. Consider the interpersonal case (if the circularity is not vicious here, then a fortiori it is not in the intrapersonal case). To analyse a value relation between two items in terms of that very value relation is viciously circular. But the sympathy-based conception does not do that. Rather, it analyses a value relation in terms of individuals' beliefs and judgements about that relation. What we have, here, is not troubling circularity, but a kind of self-reference. The value relation $\succcurlyeq^{WB}$ is instantiated just in case individuals have certain thoughts, thoughts about that very relation. This sort of self-reference is familiar from the literature on secondary properties.[43] An object has the property of redness if it has surface

[42] Darwall seems to suggest otherwise: '[C]are or concern exists as a natural psychological kind for us to refer to … If concern or care for someone for his sake is a natural psychological kind, then we can make use of it in a theory of welfare without having to define it' (Darwall 2002: 12). If Darwall is correct, here, then the circularity challenge to the sympathy-based conception of extended preferences is yet weaker.

[43] See Smith *et al*. (1989), Darwall *et al*. (1992), Casati and Tappolet (1998), Miller (2003: chs 7, 9–10).

reflectance characteristics which make it look red to normal observers. Note, here, how persons' perceptions or judgements of redness are an ineliminable aspect of the property of redness.

The characterization of interpersonal comparisons in terms of well-being judgements and beliefs is, indeed, quite useful. One way of gaining evidence about whether $(x;i) \succcurlyeq^{WB} (y;j)$ is by asking individuals (with good information, thinking clearly, etc.) whether they *believe* that i in x is better off than j in y. (By analogy, the process of constructing a successful device to detect red objects would involve asking observers which objects *look* red.) Turn this coin over: the very usefulness of this characterization underscores that it is not viciously circular. By contrast, the sham 'analysis' of $\succcurlyeq^{WB}$ in terms of well-being (rather than thoughts about well-being) offers no basis for determining whether i or j is indeed better off.

To be sure, the sympathy-based conception is *not* value-free. It links well-being to thoughts about well-being. Thus the analysis, although not viciously circular, is not as 'nice' as a characterization of well-being wholly in terms of non-value facts and non-evaluative thoughts. But philosophical efforts to produce a value-free analysis of well-being have been a failure. In particular, no value-free solution to the 'wrong kind of preference' problem has yet succeeded

A second critique of the sympathy-based conception is that it fails to satisfy what Harsanyi calls the 'Principle of Acceptance'. This principle says that a spectator's extended preferences track the subject's outcome preferences in the intrapersonal case.

> **Principle of Acceptance**: $(x;i) \succcurlyeq^{k} (y;i)$ iff R_i is such that i weakly prefers x to y.

The Principle of Acceptance was a central element of Harsanyi's account of extended preferences.[44]

However, the Principle of Acceptance is not a plausible component of the sympathy-based conception of extended preferences. The Principle, as just formulated, runs afoul of the 'wrong kind of preference' problem. Imagine that i prefers x to y on non-self-interested grounds. In such a case – given the connection between extended preferences and $\succcurlyeq^{WB}$ – it is problematic to require that k extendedly prefer $(x; i)$ to $(y; i)$.

A refinement to the Principle of Acceptance makes it more plausible.

> **Modified Principle of Acceptance:** $(x;i) \succcurlyeq^{k} (y;i)$ iff R_i is such that i self-interestedly weakly prefers x to y.

However, this principle is far from compelling. I may care wholeheartedly about you, but refuse to take your views about your well-being as decisive. Sympathetic preferences may, in some cases, be paternalistic. This can occur, for example, where (1) the sympathizer gives more weight

[44] Harsanyi (1977: 52)

to some aspect of well-being than the target; and (2) the sympathizer believes that the realization of this aspect of well-being is sufficiently insensitive to the target's preferences. For example, the sympathizer may give greater emphasis to health; the target, to enjoyment. Moreover, the sympathizer believes (correctly) that a strong preference for health is not a precondition for its realization; people who care little for their health can still end up healthy. In one outcome, the target has a healthy but less enjoyable lifestyle; in a second, he is more indulgent but harms his health. The target self-interestedly prefers the second outcome; the sympathizer, caring wholeheartedly about the target, and knowing of his preference, believes that he undervalues health, and prefers the first outcome.

The Modified Principle of Acceptance precludes the spectator from *ever* having an intrapersonal history ranking that deviates from the subject's. This is arguably too strong. By contrast, the sympathy-based conception (without that principle) *permits* the spectator to take account of the subject's preferences, without *requiring* the spectator to take those preferences as decisive. For example, the spectator believes that reading great literature or seeing great art makes a larger contribution to the target's well-being than watching sit-coms on TV, but only if the target himself has a taste for these more refined pursuits. The spectator thus extendedly prefers a history in which the target spends his free time experiencing art or literature, rather than sit-coms, iff the target self-interestedly prefers to spend his time this way. The sympathy-based conception of extended preferences (without the Modified Principle of Acceptance) permits this.

Still, some readers may wish to embrace the Modified Principle of Acceptance. They may argue: 'It is true that a sympathizer k whose attitude of sympathy is targeted at some other person i may act paternalistically towards i. However, the best account of well-being is strongly non-paternalistic, in the following sense. If i is sufficiently idealized, and is ranking outcomes under a condition of *self-sympathy* (self-interest), then one outcome is at least as good for i as a second iff i weakly prefers the first – regardless of what others may prefer, and even if these others are also idealized, and also sympathetic towards i'.

Whether the most attractive account of well-being *is* strongly non-paternalistic, in the sense just described, is controversial. The sympathy-based model of extended preferences will yield a strongly non-paternalistic account of well-being *if* it is revised to incorporate the Modified Principle of Acceptance.

The Sympathy-Based Conception (incorporating the Modified Principle of Acceptance)

(1) *Intrapersonal case*: $(x;i) \succcurlyeq^{k} (y;i)$ iff R_i is such that i has a self-interested weak preference for outcome x over outcome y.

(2) *Interpersonal case*: $(x;i) \succcurlyeq^k (y;j)$, with i and j distinct, iff spectator k makes the judgement that i in x is at least as well off as j in y, given the profile of tastes in R.

A third criticism of the sympathy-based conception (in both the original version and revised to incorporate the Modified Principle of Acceptance) is that nothing in my definitions of intra- and interpersonal extended preferences ensures that $\succcurlyeq^k$ is a quasiordering of **H**. For example, the spectator might have well-behaved intrapersonal rankings, and then make interpersonal judgements that yield an intransitivity.

The answer, here, is that k should treat his initial intra- and interpersonal rankings as provisional, and be willing to revise them to avoid intransitivity. Think of this as a rationality requirement on extended preferences. How k should engage in this revision process, so as to maximize consistency with his initial rankings, is a complicated topic that I will not attempt to address. It is clear, however, that if each of k's N intrapersonal rankings (his rankings of outcomes under a condition of sympathy with each subject) is itself a quasiordering, then these and his interpersonal rankings can always be revised in some fashion so that $\succcurlyeq^k$ is a quasiordering which never contradicts (either conforms to or 'extends') the initial intrapersonal rankings.[45]

[45] Assume that each initial intrapersonal ranking is a quasiordering. $\succcurlyeq^k$ conforms to such ranking where: $(x; i) \succcurlyeq^k (y; i)$ iff k initially ranks $(x; i)$ at least as good as (*sh at9f.911 TD.1 Tc(in7*

A final topic concerns the representation of extended preferences via utility numbers. Assume, first, that $\succcurlyeq^k$ is a complete quasiordering of the set **H** of histories. Standard results in utility theory show that, if **H** is finite or countable, $\succcurlyeq^k$ can be represented by a single utility function $v^k(.)$ such that $(x;i)\succcurlyeq^k(y;j)$ iff $v^k(x;i) \geq v^k(y;j)$. Such a $v^k(.)$ may (but need not) exist if $\succcurlyeq^k$ is complete and **H** is uncountable;[46] and of course no single utility function can represent an incomplete quasiordering of **H**. However, every $\succcurlyeq^k$ – whether complete or incomplete, and regardless of the cardinality of **H** – can be represented by some *set* of utility functions $\mathbf{V}^k$, such that $(x; i) \succcurlyeq^k (y; j)$ iff $v^k(x; i) \geq v^k(y; j)$ for all $v^k(.)$ in $\mathbf{V}^k$.[47]

5. CONSTRUCTING WELL-BEING COMPARISONS

Convergence can fail – even if the Modified Principle of Acceptance is adopted, and a fortiori if it is rejected. For example, two spectators can make different interpersonal judgements, because of a disagreement about the sources of well-being. We thus face the interesting question of constructing $\succcurlyeq^{WB}$ where spectators need not have the same extended preferences.[48]

Formally, we are looking for a rule that takes a profile P of extended preferences over set **H**, $P = (\succcurlyeq^1, \succcurlyeq^2, \succcurlyeq^3, \ldots, \succcurlyeq^N)$, and maps this profile onto a well-being quasiordering of **H**, $\succcurlyeq^{WB}$.[49] I will use the symbol $\succcurlyeq^k_P$ to mean the extended preferences of spectator k under profile P, and $\succcurlyeq^{WB}(P)$ to mean the well-being quasiordering associated with P – dropping the 'P' where the particular profile at issue is clear from context. A statement of identity between extended preferences, namely, $\succcurlyeq^k_P = \succcurlyeq^l_Q$, means that for all $(x; i)$ and $(y; j)$ in **H**, $(x;i)\succcurlyeq^k_P(y;j)$ iff $(x;i)\succcurlyeq^l_Q(y;j)$. Similarly, $\succcurlyeq^{WB}(P) = \succcurlyeq^{WB}(Q)$ means that, for all $(x; i)$ and $(y; j)$ in **H**, $(x;i)\succcurlyeq^{WB}(P)(y;j)$ iff $(x;i)\succcurlyeq^{WB}(Q)(y;j)$. Extended preferences can be

these revised rankings will yield a quasiordering $\succcurlyeq^k$ that extends k's initial intrapersonal rankings.

[46] See Kreps (1988).

[47] See Evren and Ok (2011); see also Donaldson and Weymark (1998).

[48] This problem seems to have been little discussed. Roberts (1997) is an important exception. However, he focuses on individual extended utility functions on **H** as the inputs and a collective ('objective') extended utility function as the output – thus presupposing that $\succcurlyeq^{WB}$ is complete. The literature on 'extensive social choice' (see Suzumura 1996; Ooghe and Lauwers 2005) is also relevant, but focuses on the social ranking of outcomes given the plurality of extended utility functions, rather than determining a collective judgement of well-being.

[49] Although P itself depends on the profile R of individual tastes, I assume that $\succcurlyeq^{WB}$ does not depend directly on those tastes. Let $P(R)$ be the profile of extended preferences associated with R. Strictly, $\succcurlyeq^{WB}$ is a function of both R and $P(R)$, but I assume that $\succcurlyeq^{WB}(R, P(R))$ is such that if $P(R) = P(R^*)$, then $\succcurlyeq^{WB}(R, P(R)) = \succcurlyeq^{WB}(R^*, P(R^*))$.

heterogeneous (Convergence can fail), i.e. it is possible that $\succcurlyeq_P^k \neq \succcurlyeq_P^l$, with k and l different individuals.

This problem, obviously, has an Arrovian 'feel'. But it differs in critical respects from the classic Arrow problem. First, the 'inputs' and 'outputs' need only be quasiorderings, not complete orderings. Second, what we are after is not a social ranking of outcomes, but a ranking of histories in light of well-being.

One plausible rule is the **Pooling Rule**: $(x;i)\succcurlyeq^{\mathrm{WB}}(P)(y;j)$ iff $(x;i)\succcurlyeq_P^k(y;j)$ for all spectators k. This Rule can also be expressed in terms of utility functions. The quasiordering $\succcurlyeq^{\mathrm{WB}}$, *however* constructed, can always be represented by some set $\mathbf{V}$ of utility functions such that $(x;i)\succcurlyeq^{\mathrm{WB}}(y;j)$ iff $v(x;i)\geq v(y;j)$ for all $v(.)$ in $\mathbf{V}$. In the case of the Pooling Rule, $\mathbf{V}$ is simply the union of $\mathbf{V}^1, \mathbf{V}^2, \ldots, \mathbf{V}^N$, where $\mathbf{V}^k$ represents $\succcurlyeq^k$.

The Pooling Rule satisfies various plausible axioms: Unanimous Indifference and Weak Superiority, Independence, Spectator and Subject Anonymity, and either Strong or Weak Non-paternalism.

'Unanimity' axioms say that *universal* spectator judgements are decisive with respect to well-being. Well-being facts are *accessible* to the community of spectators (at least idealized ones). It is not possible for *everyone* to be wrong about well-being. Formally, these are analogous to the Pareto principles in social choice.

> **Unanimous Indifference**: $(x;i)\sim_P^k(y;j)$ for all k implies $(x;i)\sim^{\mathrm{WB}}(P)(y;j)$.
> **Unanimous Weak Superiority**: $(x;i)\succ_P^k(y;j)$ for all k implies $(x;i)\succ^{\mathrm{WB}}(P)(y;j)$.

The Independence Axiom says that the well-being ranking of any pair of histories is wholly determined by spectators' preferences between those histories.

> **Independence**: Let profiles P and Q and histories $(x;\ i)$ and $(y;\ j)$ be such that: for all $k, (x;i)\succcurlyeq_P^k(y;j)$ iff $(x;i)\succcurlyeq_Q^k(y;j)$. Then $(x;i)\succcurlyeq^{\mathrm{WB}}(P)(y;j)$ iff $(x;i)\succcurlyeq^{\mathrm{WB}}(Q)(y;j)$.

'Independence' is, of course, the analogue of Arrow's 'Independence of Irrelevant Alternatives'. The former seems considerably more plausible than the latter. In determining whether some outcome x is socially preferred, dispreferred, or indifferent to some other outcome y (the Arrow problem), we may well want to take into consideration more information than the affected individuals' ordinal rankings of the outcomes – in particular, information about the pattern of interpersonally comparable well-being in both outcomes. By contrast, the problem at hand is to arrive

at a basis *for* ascriptions of interpersonally comparable well-being, as a function of *spectators*' extended preferences. What relevant information about spectators' extended preferences is excluded by Independence?

Spectator anonymity expresses the idea that everyone's well-being views have equal weight. Subject anonymity expresses the idea that the well-being ranking of a given subject's histories does not depend upon the identity of the subject (were the spectators to have the same extended preferences with respect to someone else's histories, the well-being ranking would be the same).

Let π be any permutation mapping on the set of N subjects and spectators.

Spectator Anonymity: If $\succcurlyeq_P^k = \succcurlyeq_Q^{\pi(k)}$ for all k, then $\succcurlyeq^{\mathrm{WB}}(P) = \succcurlyeq^{\mathrm{WB}}(Q)$.
Subject Anonymity: For all k, let $\succcurlyeq_P^k$ and $\succcurlyeq_Q^k$ be such that, for all histories $(x;\ i)$ and $(y;\ j)$, $(x;i)\succcurlyeq_P^k(y;\ j)$ iff $(x;\pi(i))\succcurlyeq_Q^k(y;\pi(j))$. Then $(x;i)\succcurlyeq^{\mathrm{WB}}(P)(y;\ j)$ iff $(x;\pi(i))\succcurlyeq^{\mathrm{WB}}(Q)(y;\pi(j))$.

Non-paternalism axioms give a special role to each subject in ranking his own histories. Strong non-paternalism says that the well-being ranking of a given subject's histories is identical to the subject's extended preferences over those histories. If spectators are required to construct extended preferences in accordance with the Modified Principle of Acceptance, the Pooling Rule satisfies strong non-paternalism. If extended preferences can violate that principle, the Pooling Rule can violate strong non-paternalism, but it still satisfies weak non-paternalism – namely, that the well-being ranking of a subject's histories will never directly contradict the subject's own ranking.

Strong Non-paternalism: $(x;i)\succcurlyeq^{\mathrm{WB}}(P)(y;i)$ iff $(x;i)\succcurlyeq_P^i(y;i)$.
Weak Non-paternalism: If $(x;i)\succ_P^i(y;i)$, then not $(y;i)\succcurlyeq^{\mathrm{WB}}(P)(x;i)$.

Notwithstanding its axiomatic virtues, the Pooling Rule might be criticized for producing 'too much' incompleteness in $\succcurlyeq^{\mathrm{WB}}$. Note, specifically, that two histories will be non-comparable if some spectator ranks them as non-comparable, or if two spectators have conflicting strict extended preferences over the histories.[50]

[50] Two histories $(x;\ i)$ and $(y;\ j)$ are non-comparable if not $(x;\ i)\succcurlyeq^{\mathrm{WB}}(P)$ $(y;\ j)$ and not $(y;\ j)\succcurlyeq^{\mathrm{WB}}(P)$ $(x;\ i)$. Under the Pooling Rule, either of the following suffices for non-comparability: (1) there exists k such that not $(x;\ i)\succcurlyeq_P^k$ $(y;\ j)$ and not $(y;\ j)\succcurlyeq_P^k$ $(x;\ i)$; (2) there exist $k,\ l$, such that $(x;\ i)\succ_P^k$ $(y;\ j)$ and $(y;\ j)\succ_P^l$ $(x;\ i)$. Note also that if the Modified Principle of Acceptance is posited, non-comparability arises between two histories with the same subject iff she ranks them as non-comparable:

The force of this criticism is open to dispute. It might be argued that if facts about well-being just *are* facts about the convergent idealized extended preferences of a community of spectators, divergence in their views about two histories *should* yield well-being non-comparability between the two.[51] However, it is certainly important to investigate plausible rules for constructing $\succcurlyeq^{WB}$ other than the Pooling Rule. This is a topic for future research.

6. WELL-BEING DIFFERENCES

Extended preferences (as described thus far) provide *ordinal* information about spectators' views regarding the histories in **H**. Whether $(x; i) \succcurlyeq^{k} (y; j)$ depends upon k's ordering of $(x; i)$ and $(y; j)$, more specifically whether he judges $(x; i)$ to be at least as good as $(y; j)$ if the two subjects are distinct, and whether he ranks x at least as good as y under a condition of sympathy with the subject if $i = j$. Similarly, $\succcurlyeq^{WB}$ is ordinal: $(x; i) \succcurlyeq^{WB} (y; j)$ means that the well-being level of i in x is at least as large as the well-being level of j in y.

The difference quasiordering $\succcurlyeq^{DIFF}$ is a natural way to represent *cardinal* well-being facts regarding the histories in **H**. $\succcurlyeq^{DIFF}$ is a quasiordering on $\mathbf{H} \times \mathbf{H}$: $((x; i), (y; j)) \succcurlyeq^{DIFF} ((z; l), (w; m))$ is to be interpreted as: the difference in well-being between $(x; i)$ and $(y; j)$ is at least as large as the difference in well-being between $(z; l)$ and $(w; m)$. Moreover, so as to reflect truisms about well-being differences, and their connection to well-being levels, $\succcurlyeq^{DIFF}$ must satisfy certain additional 'Difference Constraints'.[52]

not $(x; i) \succcurlyeq^{WB}(P) (y; i)$ and not $(y; i) \succcurlyeq^{WB}(P) (x; i)$ iff not $(x; i) \succcurlyeq^{i}_{P} (y; i)$ and not $(y; i) \succcurlyeq^{i}_{P} (x; i)$.

[51] Cf. Smith (1994: 173), proposing that normative facts are facts about convergent idealized preferences.

[52] If $\mathbf{S} = \{a, b, c \ldots\}$ is an arbitrary set, with $\succcurlyeq^{D}$ understood as a ranking of the differences between the elements of **S**, $\succcurlyeq^{D}$ cannot merely be a quasiordering (reflexive, transitive, binary relation) on **S** x **S**. In order to capture our intuitive understanding of how differences behave, $\succcurlyeq^{D}$ will need to satisfy additional constraints. What these might be is suggested by the scholarly literature on difference quasiorderings. See, e.g. Köbberling (2006); Krantz *et al.* (2007: 150–157). Difference Constraints should surely include the following: **Reversal**: $(a, b) \succcurlyeq^{D} (c, d)$ iff $(d, c) \succcurlyeq^{D} (b, a)$. **Separability**: If $(a, b) \succcurlyeq^{D} (c, b)$ then $(a, b^*) \succcurlyeq^{D} (c, b^*)$ for all b^*; and if $(a, b) \succcurlyeq^{D} (a, c)$, then $(a^*, b) \succcurlyeq^{D} (a^*, c)$ for all a^*. **Neutrality**: $(a, a) \sim^{D} (b, b)$ for all a, b. **Concatenation**: If $(a, b) \succcurlyeq^{D} (a', b')$ and $(b, c) \succcurlyeq^{D} (b', c')$ then $(a, c) \succcurlyeq^{D} (a', c')$. **Linkage**: The ranking $\succcurlyeq$ of **S** to which $\succcurlyeq^{D}$ is meant to correspond must satisfy the following requirement: $a \succcurlyeq b$ iff $(a, b) \succcurlyeq^{D} (b, b)$.

These would apply to $\succcurlyeq^{WB}$ and $\succcurlyeq^{DIFF}$ as follows. First, $\succcurlyeq^{DIFF}$ will not only be a quasiordering of **H** x **H**, but will satisfy Reversal, Separability, Neutrality and Concatenation (and any other such Difference Constraints not logically implied by these, if there are any). For example, Reversal applied to $\succcurlyeq^{DIFF}$ means: $((x; i), (y; j)) \succcurlyeq^{DIFF} ((z; l), (w; m))$ iff $((w; m), (z; l)) \succcurlyeq^{DIFF} ((y; j), (x; i))$. Second, $\succcurlyeq^{DIFF}$ will satisfy Linkage vis-à-vis $\succcurlyeq^{WB}$, that is: $(x; i) \succcurlyeq^{WB} (y; j)$ iff $((x; i), (y; j)) \succcurlyeq^{DIFF} ((y; j), (y; j))$.

If $\succcurlyeq^{WB}$ is a *complete* quasiordering on **H**, and $\succcurlyeq^{DIFF}$ a *complete* quasiordering on **H** x **H** that satisfies the Difference Constraints, then (depending on whether further technical conditions are satisfied) they may be jointly representable by a single utility function $v(.)$, such that: $(x;i)\succcurlyeq^{WB}(y;j)$ iff $v(x;i)\geq v(y;j)$ and $((x;i),(y;j))\succcurlyeq^{DIFF}((z;l),(w;m))$ iff $v(x;i)-v(y;j)\geq v(z;l)-v(w;m)$. By extension, if $\succcurlyeq^{WB}$ is a possibly incomplete quasiordering on **H**, and $\succcurlyeq^{DIFF}$ a possibly incomplete quasiordering on **H** x **H** that satisfies the Difference Constraints, they may be jointly representable by a *set* of utility functions **V** such that: $(x;i)\succcurlyeq^{WB}(y;j)$ iff $v(x;i)\geq v(y;j)$ for all $v(.)$ in **V**; and $((x;i),(y;j))\succcurlyeq^{DIFF}((z;l),(w;m))$ iff $v(x;i)-v(y;j)\geq v(z;l)-v(w;m)$ for all $v(.)$ in **V**.[53]

How are we to construct $\succcurlyeq^{DIFF}$, as a function of spectators' preferences and well-being judgements? This is a large and complicated topic, which I lack space to discuss in detail here. However, two basic possibilities should be noted. The first strategy is to invite spectators to make their own judgements about well-being differences.[54] For any given four-tuple of histories, spectator k asks herself: do I believe the well-being difference between the first two to be at least as large as the well-being difference between the second two? If spectator k's difference judgements are sufficiently coherent, they will take the form of a personal difference quasiordering $\succcurlyeq^{k\text{-Diff}}$: a quasiordering on $\mathbf{H}\times\mathbf{H}$ which satisfies the Difference Constraints vis-à-vis $\succcurlyeq^{k}$. We can then formulate some rule (e.g. a Pooling Rule) for mapping a profile of such judgements $(\succcurlyeq^{1\text{-Diff}},\succcurlyeq^{2\text{-Diff}},\ldots,\succcurlyeq^{N\text{-Diff}})$ onto $\succcurlyeq^{DIFF}$.

Note that this strategy does not create a direct connection between spectator k's personal difference quasiordering, $\succcurlyeq^{k\text{-Diff}}$, and her preferences when sympathetic to various subjects. Still, there remains a substantial, indirect connection – since $\succcurlyeq^{k\text{-Diff}}$ must satisfy Difference Constraints relative to $\succcurlyeq^{k}$, and $\succcurlyeq^{k}$ *does* have a direct nexus to the spectator's sympathetic preferences. Recall that $(x;i)\succcurlyeq^{k}(y;i)$ iff spectator k weakly prefers x to y when unreservedly sympathetic to subject i. These

[53] On the representation of a complete difference quasiordering by a utility function, see Köbberling (2006); Krantz *et al.* (2007: 150–157). By 'technical conditions', I mean axioms such as 'Archimedean' or 'solvability' which figure in these representation proofs, but do not seem to be part of the very concept of a ranking of differences – by contrast with the requirements I have labelled 'Difference Constraints'.

Adler (2012: ch. 3) constructs a set **V** that pools the extended utility functions of a group of spectators, and then defines $\succcurlyeq^{WB}$ and $\succcurlyeq^{DIFF}$ from **V** using the set-valued rule stated in the text. (No attempt is made there to describe the technical conditions *generally* guaranteeing that any $\succcurlyeq^{WB}$ and $\succcurlyeq^{DIFF}$ are representable by a set **V**.) It is worth noting that if a quasiordering $\succcurlyeq^{WB}$ on **H** and a difference quasiordering $\succcurlyeq^{DIFF}$ on $\mathbf{H}\times\mathbf{H}$ are indeed representable by a set **V** using the set-valued rule, $\succcurlyeq^{DIFF}$ will satisfy all of the Difference Constraints mentioned in the footnote immediately above.

[54] See Abdellaoui *et al.* (2007).

sympathetic preferences will, in turn, help structure $\succcurlyeq^{k\text{-Diff}}$. For example, if k under a condition of unreserved sympathy with i weakly prefers outcome x to y to z, i.e. $(x;i)\succcurlyeq^k(y;i)\succcurlyeq^k(z;i)$, then $\succcurlyeq^{k\text{-Diff}}$ must assign a difference between $(x;i)$ and $(z;i)$ at least as large as the difference between $(y;i)$ and $(z;i)$.

A second strategy (call it the Bernoulli strategy) is to involve sympathy more directly in $\succcurlyeq^{k\text{-Diff}}$.[55] Let $u_i^k(.)$ be a vNM utility function that expectationally represents spectator k's preferences over outcome *lotteries* under a condition of unreserved sympathy with subject i. Then the Bernoulli strategy uses $u_i^k(.)$ to define $\succcurlyeq^{k\text{-Diff}}$ in the intrapersonal case, with respect to subject i's histories: $((x;i),(y;i))\succcurlyeq^{k\text{-Diff}}((z;i),(w;i))$ iff $u_i^k(x)-u_i^k(y)\geq u_i^k(z)-u_i^k(w)$.[56] As with the first strategy, $\succcurlyeq^{\text{DIFF}}$ is constructed from $(\succcurlyeq^{1\text{-Diff}},\succcurlyeq^{2\text{-Diff}},\ldots,\succcurlyeq^{N\text{-Diff}})$ using a Pooling Rule or some other rule.

While Harsanyi's derivation of difference comparisons from vNM utility functions was empathy-based, the Bernoulli strategy here is *sympathy*-based. Harsanyi asked spectators to rank lotteries over histories, with the spectator meant to think about the probability assigned by a lottery to a history $(x;i)$ as the probability that the spectator will 'stand in the shoes' (acquire the attributes) of subject i in outcome x. By contrast, the Bernoulli strategy under discussion now asks each spectator to rank, in turn, N subsets of lotteries over histories – the subset consisting of lotteries over histories that have individual 1 as subject and thus take the form $(x;1)$, the subset consisting of lotteries over histories of the form $(x;2)$, ..., the subset consisting of lotteries over histories of the form $(x;N)$ – with each such ranking understood as the spectator's ranking of lotteries over *outcomes* when unreservedly sympathetic with the given subject. The spectator is never asked to imagine acquiring someone else's identity.

Still, the Bernoulli strategy will be controversial.[57] It is a truism about sympathy that my judgements about your welfare levels in various outcomes are mirrored by my preferences over those outcomes when unreservedly sympathetic to you. It may be true, but is hardly a *truism*, that my judgements about *differences* between your welfare levels in

[55] See Adler (2012: ch. 3).

[56] What about the interpersonal case, i.e., $((x;i),(y;j))\succcurlyeq^{k\text{-Diff}}((z;l),(w;m))$, where it is not the case that $i=j=l=m$? The Bernoulli strategy cannot define $\succcurlyeq^{k\text{-Diff}}$ in the interpersonal case with reference to k's ranking of outcomes under a condition of unreserved sympathy with 'the subject', since there are multiple subjects. (This is, of course, exactly the same problem that arose in using sympathy to define $\succcurlyeq^k$ in the interpersonal case, as discussed in Part 4.) Rather, in the interpersonal case, the Bernoulli strategy must be either (a) to ask k to make judgements of well-being differences; or (b) to derive $\succcurlyeq^{k\text{-Diff}}$ from $u_i^k(.)$ and $\succcurlyeq^k$. Adler (2012: ch. 3) in effect does the latter.

[57] As already noted, the existing literature on Harsanyi contains criticisms of his use of vNM utility functions to derive well-being differences. See above note 11. These criticisms are independent of his empathy-based conception of extended preferences, and would therefore also apply to the sympathy-based Bernoulli strategy now under discussion.

various outcomes are mirrored by differences in the values of a vNM utility function expectationally representing my preferences for lotteries over those outcomes when unreservedly sympathetic to you. Imagine that k, under a condition of sympathy with subject i, strictly prefers outcome z to y to x; and k makes the judgement that the difference between i's well-being in z and y is exactly equal to the difference in his well-being between y and x. Why are we justified in insisting (as the Bernoulli strategy does) that spectator k, under a condition of sympathy with i, must now be indifferent between y and a lottery giving a 50% probability of z and a 50% probability of x? Why would it be incoherent or problematic for k to rank y above or below this lottery?

A full analysis of the pros and cons of the two strategies for defining $\succcurlyeq^{k\text{-Diff}}$ must be left to another day.[58] What should be emphasized, here, is that each represents an intellectual *extension* of the sympathy-based conception of extended preferences. The first strategy asks spectator k to make well-being judgements, regarding differences, just as the construction of $\succcurlyeq^{k}$, in the interpersonal case, asks spectator k to make judgements of well-being levels. The second, Bernoulli, strategy brings sympathy into the picture more directly. Moreover, as already noted, the first strategy is indirectly constrained by the use of sympathy in defining $\succcurlyeq^{k}$. Neither strategy employs the problematic device of empathetic projection.

REFERENCES

Abdellaoui, M., C. Barrios and P. P. Wakker. 2007. Reconciling introspective utility with revealed preference: experimental arguments based on prospect theory. *Journal of Econometrics* 138: 356–378.

Adler, M. D. 2012. *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford: Oxford University Press.

Adler, M. D. Forthcoming. Extended preferences. In *Oxford Handbook of Well-Being and Public Policy*, ed. M. D. Adler and M. Fleurbaey. Oxford: Oxford University Press.

Arneson, R. J. 1999. Human flourishing versus desire satisfaction. In *Human Flourishing*, ed. E. F. Paul, F. D. Miller and J. Paul, 113–142. Cambridge: Cambridge University Press.

Arrow, K. J. 1977. Extended sympathy and the possibility of social choice. *American Economic Review: Papers and Proceedings* 67: 219–225.

Bernstein, M. 1998. Well-being. *American Philosophical Quarterly* 35: 39–55.

Brandt, R. B. 1998. *A Theory of the Good and the Right*. Amherst, MA: Prometheus Books.

Broome, J. 1995. *Weighing Goods: Equality, Uncertainty and Time*. Oxford: Blackwell.

Broome, J. 1998. Extended preferences. In *Preferences*, ed. C. Fehige and U. Wessels, 271–287. Berlin: Walter de Gruyter.

Brown, M. 2003. The elimination of personal identity. *Southwest Philosophy Review* 19: 239–247.

Casati, R. and C. Tappolet, eds. 1998. *European Review of Philosophy, vol. 3: Response-Dependence*. Stanford: CSLI Publications.

Darwall, S. L. 1998. Empathy, sympathy, care. *Philosophical Studies* 89: 261–282.

[58] See Adler (forthcoming).

Darwall, S. L. 2002. *Welfare and Rational Care*. Princeton, NJ: Princeton University Press.
Darwall, S. L., A. Gibbard and P. Railton. 1992. Toward fin de siècle ethics: some trends. *Philosophical Review* 101: 115–189.
Donaldson, D. and J. A. Weymark. 1998. A quasiordering is the intersection of orderings. *Journal of Economic Theory* 78: 382–387.
Eisenberg, N. 2000. Empathy and sympathy. In *Handbook of Emotions*, ed. M. Lewis and J. M. Haviland-Jones, 677–691. 2nd edition. New York, NY: Guilford Press.
Evren, Ö. and E. A. Ok. 2011. On the multi-utility representation of preference relations. *Journal of Mathematical Economics* 47: 554–563.
Feit, N. 2008. *Belief about the Self: A Defense of the Property Theory of Content*. Oxford: Oxford University Press.
Fleurbaey, M. Forthcoming. Equivalent income. In *Oxford Handbook of Well-Being and Public Policy*, ed. M. D. Adler and M. Fleurbaey. Oxford: Oxford University Press.
Fleurbaey, M. and D. Blanchet. 2013. *Beyond GDP: Measuring Welfare and Assessing Sustainability*. Oxford: Oxford University Press.
Fleurbaey, M. and P. Mongin. 2012. The utilitarian relevance of the aggregation theorem. Working paper, October 2012.
Fontaine, P. 1997. Identification and economic behavior: sympathy and empathy in historical perspective. *Economics and Philosophy* 13: 261–280.
Gajdos, T. and F. Kandil. 2008. The ignorant observer. *Social Choice and Welfare* 31: 193–232.
Gibbard, A. 1986. Interpersonal comparisons: preference, good, and the intrinsic reward of a life. In *Foundations of Social Choice Theory*, ed. J. Elster and A. Hylland, 165–193. Cambridge: Cambridge University Press.
Gordon, R. M. 1995. Sympathy, simulation, and the impartial spectator. *Ethics* 105: 727–742.
Grant, S., A. Kajii, B. Polak and Z. Safra. 2010. Generalized utilitarianism and Harsanyi's impartial observer theorem. *Econometrica* 78: 1939–1971.
Grant, S., A. Kajii, B. Polak and Z. Safra. 2012a. Equally-distributed equivalent utility, ex post egalitarianism and utilitarianism. *Journal of Economic Theory* 147: 1545–1571.
Grant, S., A. Kajii, B. Polak and Z. Safra. 2012b. A generalized representation theorem for Harsanyi's ('impartial') observer. *Social Choice and Welfare* 39: 833–846.
Griffin, J. 1986. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press.
Harsanyi, J. C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61: 434–435.
Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.
Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.
Harsanyi, J. C. 1982. Morality and the theory of rational behaviour. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams, 39–62. Cambridge: Cambridge University Press.
Hausman, D. M. and M. S. McPherson. 2009. Preference satisfaction and welfare economics. *Economics and Philosophy* 25: 1–25.
Kagan, S. 1992. The limits of well-being. *Social Philosophy and Policy* 9: 169–189.
Kind, A. 2011. The puzzle of imaginative desire. *Australasian Journal of Philosophy* 89: 421–439.
Köbberling, V. 2006. Strength of preference and cardinal utility. *Economic Theory* 27: 375–391.
Krantz, D. H., R. D. Luce, P. Suppes and A. Tversky. 2007. *Foundations of Measurement*. Vol. 1, *Additive and Polynomial Representations*. Mineola, NY: Dover.
Kreps, D. M. 1988. *Notes on the Theory of Choice*. Boulder, CO: Westview Press.
Lewis, D. 1979. Attitudes *de dicto* and *de se*. *Philosophical Review* 88: 513–543.
Lowe, E. J. 2002. *A Survey of Metaphysics*. Oxford: Oxford University Press.
Mackie, P. 2006. *How Things Might Have Been: Individuals, Kinds, and Essential Properties*. Oxford: Clarendon Press.

Miller, A. 2003. *An Introduction to Contemporary Metaethics*. Cambridge: Polity Press.
Mongin, P. 2001. The impartial observer theorem of social ethics. *Economics and Philosophy* 17: 147–179.
Mongin, P. and C. d'Aspremont. 1998. Utility theory and ethics. In *Handbook of Utility Theory*, Vol. 1 (*Principles*), ed. S. Barberà, P. J. Hammond, and C. Seidl, 371–481. Dordrecht: Kluwer Academic.
Nichols, S. 2008. Imagination and the *I*. *Mind and Language* 23: 518–535.
Ninan, D. 2009. Persistence and the first-person perspective. *Philosophical Review* 118: 425–464.
Nozick, R. 1974. *Anarchy, State, and Utopia*. New York, NY: Basic Books.
Ooghe, E. and L. Lauwers. 2005. Non-dictatorial extensive social choice. *Economic Theory* 25: 721–743.
Overvold, M. C. 1980. Self-interest and the concept of self-sacrifice. *Canadian Journal of Philosophy* 10: 105–118.
Overvold, M. C. 1982. Self-interest and getting what you want. In *The Limits of Utilitarianism*, ed. H. B. Miller and W. H. Williams, 186–194. Minneapolis, MN: University of Minnesota Press.
Overvold, M. C. 1984. Morality, self-interest, and reasons for being moral. *Philosophy and Phenomenological Research* 44: 493–507.
Parfit, D. 1987. *Reasons and Persons*, reprint with corrections. Oxford: Clarendon Press.
Perry, J. 1979. The problem of the essential indexical. *Noûs* 13: 3–21.
Recanati, F. 2007. *Perspectival Thought: A Plea for (Moderate) Relativism*. Oxford: Oxford University Press.
Reynolds, S. L. 1989. Imagining oneself to be another. *Noûs* 23: 615–633.
Risse, M. 2002. Harsanyi's 'utilitarian theorem' and utilitarianism. *Noûs* 36: 550–577.
Roberts, K. 1997. Objective interpersonal comparisons of utility. *Social Choice and Welfare* 14: 79–96.
Roca-Royes, S. 2011. Essential properties and individual essences. *Philosophy Compass* 6: 65–77.
Roemer, J. E. 2008. Harsanyi's impartial observer is *not* a utilitarian. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles and J. A. Weymark, 129–135. Cambridge: Cambridge University Press.
Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press.
Sen, A. 1970. *Collective Choice and Social Welfare*. San Francisco, CA: Holden-Day.
Sen, A. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision* 7: 243–262.
Sen, A. 1986. Social choice theory. In *Handbook of Mathematical Economics*, Vol. 3, ed. K. J. Arrow and M. D. Intriligator, 1073–1181. Amsterdam: North-Holland.
Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.
Smith, M., D. Lewis and M. Johnston. 1989. Dispositional theories of value. *Proceedings of the Aristotelian Society* (Supplementary vols.) 63: 89–174.
Sumner, L. W. 1996. *Welfare, Happiness, and Ethics*. Oxford: Clarendon Press.
Suzumura, K. 1996. Interpersonal comparisons of the extended sympathy type and the possibility of social choice. In *Social Choice Re-Examined: Proceedings of the IEA Conference held at Schloss Hernstein, Berndorf, Vienna, Austria*, vol. 2, ed. K. J. Arrow, A. Sen and K. Suzumura, 202–229. Houndmills: Macmillan Press.
Velleman, J. D. 1996. Self to self. *Philosophical Review* 105: 39–76.
Vendler, Z. 1976. A note to the paralogisms. In *Contemporary Aspects of Philosophy*, ed. G. Ryle, 111–121. Stocksfield: Oriel Press.
Voorhoeve, A. 2014. Book review (M.D. Adler, *Well-Being and Fair Distribution*). *Social Choice and Welfare* 42: 245–254.

Walton, K. L. 1990. *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Cambridge, MA: Harvard University Press.

Weymark, J. A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. E. Roemer, 255–320. Cambridge: Cambridge University Press.

Weymark, J. A. 2005. Measurement theory and the foundations of utilitarianism. *Social Choice and Welfare* 25: 527–555.

Williams, B. 1973. *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge: Cambridge University Press.