8-1-2011

Harsanyi 2.0

Matthew D. Adler
University of Pennsylvania Carey Law School

Harsanyi 2.0

Matthew D. Adler[1], draft of August 2011

I. Introduction

Welfare economics has never resolved the problem of interpersonal well-being comparisons. One important branch of the field eschews such comparisons, relying instead on the criterion of Kaldor-Hicks efficiency for evaluating governmental policies. But, intuitively, well-being *is* interpersonally comparable—at least to some extent—and, in any event, the Kaldor-Hicks criterion is hard to justify.

A different branch embraces interpersonal comparability. Many welfare economists, in various subfields such as optimal tax theory, growth theory, public finance, environmental economics, and social choice theory, employ "social welfare functions" (SWF) as a key tool. Let x, y, z, … denote different possible outcomes, i.e., states of affairs.[2] The SWF methodology assigns a utility number to each individual in each outcome: a number such as $V_i(x)$, the utility of individual i in outcome x. These utility numbers are taken to mirror inter- as well as intrapersonal comparisons of well-being. Supposedly, $V_i(x) > V_j(y)$ just in case individual i in outcome x is better off than individual j in outcome y. The SWF methodology then ranks outcomes as a function of their corresponding utility numbers.

However, economists in this tradition have never reached consensus about the grounding for interpersonally comparable utilities. The key difficulty is that there is no straightforward way to move from the *ordinary* utility function representing the ordinary preferences of each individual, to the *interpersonal* utility function $V(.)$. By an "ordinary preference," I mean a preference (a choice-connected pro-attitude) that takes states of affairs ("outcomes") as its arguments. What individual k prefers in this ordinary sense is that outcome x occur rather than outcome y. If k's ordinary preferences are well-behaved, there exists an ("ordinary") utility function, indexed to k – denote it as $u^k(.)$—that assigns numbers to outcomes mirroring these preferences. Individual k prefers x to y just in case $u^k(x) > u^k(y)$.

It is tempting, but naïve, to construct an interpersonal utility function by equating the *interpersonal* utility of a given individual in a given outcome with the number assigned by that individual's ordinary utility function to the outcome. In other words, $V_k(x) = u^k(x)$. This approach is naïve because ordinary utility functions are not *unique*. Imagine, for example, that there are four outcomes x through w, and that i prefers x to y to z to w, while j prefers y to w to z to x. Then i's preference ranking can be represented by an ordinary utility function that assigns

[1] Leon Meltzer Professor, University of Pennsylvania Law School. madler@law.upenn.edu.

[2] I use the term "outcome"—and economists in the SWF tradition sometimes use the term "social state" —but the item thus denoted is nothing other than what philosophers term a "state of affairs" or "proposition." An outcome need not be maximally specified; if it *is* maximally specified, an outcome is a possible world.

the numbers 100, 97, 93, 91 to the four outcomes, and *j*'s by an ordinary utility function that assigns them the numbers 20, 30, 22, 26. Using these numbers to define the interpersonal utility function $V(.)$, it looks like *i* is better off in all the outcomes than *j*. However, *i*'s preference ranking is equally well represented by a different utility function, namely one that assigns them 40, 39, 38, 37; similarly, *j*'s ranking is equally well represented by the numbers 70, 80, 71, 74. Using *these* numbers to define $V(.)$, it now looks like *j* is better off in all the outcomes than *i*.

John Harsanyi offers a clever proposal for circumventing the difficulty just described. Instead of using ordinary preferences as the basis for interpersonal comparisons, Harsanyi proposes to use "extended preferences." Rather than being a ranking of *outcomes*, an extended preference is a ranking of "extended alternatives." An "extended alternative" is a pairing of an individual's "personal position" and his ordinary preferences. As Harsanyi explains:

> [A]ny social situation can be regarded as a *vector* listing the economic, social, biological, and other variables that will affect the well-being of the individuals making up the society. Different social situations will be called *A*, *B*, ...
>
> … Let A_i denote *i*'s *personal position* in social situation *A* (i.e., the objective conditions that would face individual *i* in social situation *A*). Likewise, let B_j denote *j*'s personal position in social situation *B*. … Finally, let P_i and P_j denote *i*'s and *j*'s *subjective attitudes* (including their personal preferences), respectively.
>
> …[L]et *A* be a social situation where all individuals' diets consist mainly of fish, and let *B* be a social situation where all individuals' diets consist mainly of meat. Suppose that individual *i* has a *mild* preference for fish, while individual *j* has a very *strong* preference for meat (with a violent distaste for fish). Then individual *i*, his taste P_i being what it is, will obviously prefer fish to meat, which means that he will prefer $[A_i, P_i]$ to $[B_i, P_i]$. But he will presumably also recognize that it is better (less inconvenient) to eat meat with a *mild* distaste for meat than it is to eat fish with a *strong* distaste for fish. Therefore he will prefer $[B_i, P_i]$ to $[A_j, P_j]$. In terms of the language of interpersonal utility comparison, he will recognize that *j* would derive more disutility (i.e., would derive less utility) from eating fish than (*i*) himself would derive from eating meat.
>
> Hypothetical alternatives of the form $[A_i, P_i]$ or $[B_j, P_j]$, and so on, will be called *extended alternatives*. A given individual's (say *i*'s) preferences among such extended alternatives will be called his *extended preferences*.[3]

Simplifying Harsanyi's terminology and notation, I will denote what he calls an "extended alternative" as an "individual history" or, even more simply, a "history." A history for individual *i* is a joint specification of both the "objective conditions" that might affect her well-being, and her preferences. While Harsanyi uses a symbol such as "$[A_i, P_i]$" to denote an individual history, I will use a symbol such as "$(x; i)$." Unless otherwise noted, I will use the term "outcome" to mean a hybrid state of affairs consisting of *both* "a vector listing the economic, social, biological, and other variables that will affect the well-being of the individuals making up the society" *and* a vector listing the preferences of each individual—with specific

[3] Harsanyi (1986, pp. 49, 52-53). See also Harsanyi (1953; 1955; 1982). For a lucid presentation of Harsanyi's views, see Weymark (1991).

outcomes identified by lower-case letters such as "*x*," "*y*," or "*z*." The "individual history" (*x*; *i*) is, roughly, a description of what happens to individual *i* in outcome *x*: her own attributes, the attributes of others (insofar as these bear upon her well-being), and her tastes.

A given individual's "extended preferences," then, consist in her ranking of histories. For example, individual *k* might have an extended preference for individual history (*x*; *i*) over individual history (*y; j*): individual *k* might prefer a life in which her objective conditions are the same as *i*'s in *x*, and her preferences are the same as *i*'s in *x*, to a life in which her objective conditions are the same as *j*'s in *y*, and her preferences are the same as *j*'s in *y*. Individuals' extended preferences, if well behaved, will be representable by *extended* utility functions. Let us denote *k*'s extended utility function as $v^k(.)$. Unlike *k*'s ordinary utility function, her *extended* utility function assigns numbers to *histories* (not outcomes) and mirrors her ranking of histories (extended preferences) rather than her ranking of outcomes (ordinary preferences). In other words, $v^k(.)$ is such that *k* has an extended preference for (*x*; *i*) over (*y*; *j*) just in case $v^k(x; i) > v^k(y; j)$.

Harsanyi's clever proposal is that we use these extended utility functions as the basis for an interpersonal utility function. Moreover, we can do so in a manner that yields interpersonal comparisons of well-being differences as well as interpersonal comparisons of well-being levels.

How does this proposal work? First, Harsanyi posits that individuals have extended preferences over individual-history *lotteries*, not just histories; and that these preferences satisfy the axioms of von Neumann/Morgenstern ("vNM") utility theory, allowing them to be expectationally represented by a utility function. For short, call this the vNM premise. Let *L* be one lottery over histories, and *L** a second. The vNM premise means that, for any individual *k*, there exists a utility function $v^k(.)$ such that *k* prefers *L* to *L** just in case the expected utility assigned to the first lottery using $v^k(.)$ is greater than the expected utility thus assigned to the second.[4] Second, Harsanyi assumes that individuals will have the *same* extended preferences. For short, call this the "homogeneity" premise. It means that, for any two individuals *k* and *l*, and any two histories, *k* has an extended preference for the first over the second just in case *l* does.

---

[4] The general formulation of vNM theory is as follows. Let **R** be a set of "prizes" which, for simplicity, is finite. Let **L** be the set of all possible lotteries over those prizes. A given lottery *L* assigns a probability $\pi_L(r)$ to prize *r*, with these summing to unity: $\sum_{r\in\mathbf{R}} \pi_L(r) = 1$. Assume that individual *k* has a complete and transitive ranking of the lotteries. To say that this ranking can be "expectationally represented" by a utility function $w^k(.)$ means: *L* is ranked at least as good as *L** iff $\sum_{r\in\mathbf{R}} w^k(r)\pi_L(r) \geq \sum_{r\in\mathbf{R}} w^k(r)\pi_{L^*}(r)$. There are various groups of axioms that suffice for the existence of $w^k(.)$. One standard formulation is that individual *k*'s ranking of the lotteries must satisfy an "independence" and an "Archimedean" axiom. See Kreps (1988, ch. 5); Gilboa (2009, ch. 8).

In order to adapt this general framework to the case of extended preferences, simply make the prize set the set **H** of all histories. A lottery, now, assigns a probability $\pi_L(x;i)$ to history (*x*; *i*). To say that individual *k*'s ranking of the lotteries is "expectationally represented" by an extended utility function $v^k(.)$ means: *L* is ranked at least as good as *L** iff $\sum_{(x;i)\in\mathbf{H}} v^k(x;i)\pi_L(x;i) \geq \sum_{(x;i)\in\mathbf{H}} v^k(x;i)\pi_{L^*}(x;i)$. If *k*'s ranking of **L** is complete and transitive and satisfies an additional group of vNM axioms, $v^k(.)$ will exist.

Similarly, for any two individuals, and any two individual-history lotteries, *k* prefers the first lottery to the second just in case *l* does.

The vNM and homogeneity premises, together, mean that there will be an extended utility function $v(.)$ which expectationally represents the extended preferences of *every* person in the population. For *any* individual *k*, it will be the case that *k* prefers history $(x; i)$ to history $(y; j)$ just in case $v(x; i) > v(y; j)$; and it will be the case that *k* prefers lottery *L* to lottery *L** just in case the expected utility of the first, according to $v(.)$, is greater than the expected utility of the second, according to $v(.)$. The interpersonal utility function $V(.)$, in turn, can be immediately derived from $v(.)$ – or so Harsanyi suggests. Simply define $V(.)$ as follows: $V_i(x) = v(x; i)$. In short, the *interpersonal* utility assigned to a given individual in a given outcome is simply the extended utility of her history in that outcome.

For those inclined to see a close connection between preferences and well-being, this proposal is appealing because it retains such a nexus, yet circumvents the problem described earlier in deriving interpersonal utilities from ordinary utility functions. The problem, there, involved the non-uniqueness of ordinary utility functions. The extended-utility function, too, is non-unique. However, *its* non-uniqueness does not jeopardize the proposal to use extended preferences as the basis for interpersonal comparisons.

Why not? A well-known feature of von Neumann/Morgenstern utility theory is that utility functions which expectationally represent lottery preferences are unique *up to a positive affine transformation*. Assume that the vNM and homogeneity premises are true, and that function $v(.)$ is an extended utility function representing everyone's extended preferences over individual histories and lotteries. Then if $v^+(.)$ does the same, $v^+(.)$ must be a positive affine transformation of $v(.)$. This means that there must be a positive "scaling" factor *a*, and a "translation" factor *b*, which transforms the $v(.)$ utilities into the $v^+(.)$ utilities.[5] For example, imagine that there are four histories $(x; i)$, $(y; i)$, $(x; j)$, $(y; j)$, assigned numbers by $v(.)$ as follows: $v(x; i) = 10$, $v(y; i) = 15$, $v(x; j) = 17$, $v(y; j) = 23$. Then there must be *a* (>0) and *b* such that $v^+(x; i) = 10a + b$, $v^+(y; i) = 15a + b$, $v^+(x; j) = 17a + b$, and $v^+(y; j) = 23a + b$.

[5] Return to the general formulation of vNM theory in the previous footnote, with **R** a prize set and **L** a set of all lotteries over the prize set. Where $f(.)$ and $g(.)$ are real-valued functions on **R**, to say that $g(.)$ is a positive affine transformation of $f(.)$ means: there exists a number $a > 0$ and a number *b* such that, for every *r* in **R**, $g(r) = af(r) + b$. Assume that *k* has a ranking of **L** which satisfies the vNM axioms and can be expectationally represented by utility function $w(.)$. Then vNM theory shows the following: $w^+(.)$ *also* expectationally represents *k*'s ranking of **L** iff $w^+(.)$ is a positive affine transformation of $w(.)$. See Kreps (1988, ch. 5); Gilboa (2009, ch. 8).

Turn now to the specific case where **H** is the set of all histories; an individual *k* has extended preferences regarding these histories and lotteries over them; and $v(.)$ expectationally represents individual *k*'s ranking of the lotteries. Then $v^+(.)$ also expectationally represents *k*'s ranking of the lotteries iff $v^+(.)$ is a positive affine transformation of $v(.)$. Since **H**, now, is the set of all *histories*, what this means is: there exists a number $a > 0$ and number *b* such that, for any history $(x; i)$, $v^+(x; i) = av(x; i) + b$.

Because $v(.)$ is unique up to a positive affine transformation, the proposal to derive interpersonal comparisons from $v(.)$ yields *determinate* comparisons of well-being levels. Assume that $v(.)$ assigns a higher number to history $(x; i)$ than $(y; j)$. If $v^+(.)$ is a positive affine transformation of $v(.)$, then $v^+(.)$ must also assign a higher number to the first history. Moreover—an extra bonus—the proposal to derive interpersonal comparisons from $v(.)$ yields determinate comparisons of well-being differences. Assume that the difference between $v(x; i)$ and $v(y; j)$ is greater than the difference between $v(z; l)$ and $v(w; m)$. Then, if $v^+(.)$ is a positive affine transformation of $v(.)$, simple algebra shows that the difference between $v^+(x; i)$ and $v^+(y; j)$ must be greater than the difference between $v^+(z; l)$ and $v^+(w; m)$.[6]

Despite these virtues, Harsanyi's proposal has serious flaws. In Part II of this Article, I describe the flaws, namely these: (1) *Essential attributes*. Harsanyi assumes that the "objective conditions" constitutive of someone's "personal situation" (of her "individual history," in my terminology) are contingent attributes, such as income, health or social status. However, there seems to be no conceptual reason why well-being is *just* a function of contingent attributes. And, once a history is allowed to include essential (non-contingent) attributes, it becomes hard to understand what an extended preference *means*. (2) *Wrong attitude*. An individual may have the wrong attitude in formulating her extended preferences—an attitude that fails to explain the well-being ranking of histories. For example, she may rank histories on moral or aesthetic grounds, rather than out of any concern for the subjects. (3) *The "Principle of Acceptance" and Changing Preferences*. Harsanyi, in his so-called "principle of acceptance (discussed below), requires that anyone's extended preferences regarding the histories of one particular subject rank those histories so as to track the subject's ordinary preferences regarding the component outcomes. However, this principle, as Harsanyi presents it, does not allow for the possibility that subjects' ordinary preferences may not be fixed. (4) *Heterogeneity of Extended Preferences*. Harsanyi fails to establish the homogeneity premise; there is absolutely no reason to think that, in general, individuals will have the same extended preferences.

In Part III, I discuss how the flaws might be repaired, yielding an account of interpersonal comparisons that is quite different in its details from Harsanyi's, but retains his central idea that interpersonal utilities might be based upon extended preferences and vNM extended utility functions. In Part IV, I show how interpersonal utilities as thus constructed can be integrated with an SWF—be it a utilitarian SWF or some non-utilitarian SWF (for example, a prioritarian SWF). Although Harsanyi does in fact develop his views about interpersonal comparisons within the context of a defense of utilitarianism, there is no necessary connection between the *axiological* claim that utilities representing interpersonally comparable well-being are to be

[6] To say that $v^+(.)$ is a positive affine transformation of $v(.)$ relative to the set **H** of all histories, in the sense expressed in the previous footnote, is logically stronger than saying that there are subject-specific scaling and translation factors relating the two functions. The latter proposition means that: for each individual i, there exists $a_i > 0$, b_i, such that, for any $(x; i)$, $v^+(x; i) = a_i v(x; i) + b_i$. Note that, if $v^+(.)$ is related to $v(.)$ in this weaker sense, the two functions need not imply the same interpersonal level or difference comparisons.

derived from extended preferences, and the *moral* claim that the ranking of outcomes as morally better or worse is to be achieved by summing unweighted utilities (utilitarianism).

What is novel in this Article? The account presented in Parts III and IV, to be sure. But the critical discussion of Harsanyi's views in Part II also is, to a large extent. Much of the extant literature on Harsanyi concerns his "aggregation" theorem rather than his account of extended preferences and interpersonal comparisons. And scholarship addressing the latter topic[7] has failed fully to canvass the problems I describe. This is true, both of the critical literature reacting to Harsanyi, and of the work of other theorists (there are a few) who—like Harsanyi—have used the device of extended preferences as a basis for interpersonal comparisons.[8]

II. **A Critique of Harsanyi**

A. Essential Attributes

Recall Harsanyi's definitions. A social situation is a description of "the economic, social, biological, and other variables that will affect the well-being of the individuals making up the society." Given some social situation A, a "personal position," A_i, is "the objective conditions that would face individual i in social situation A." An "extended alternative" – what I term an individual history—is a combination of a personal position and individual tastes. "Extended preferences" are preferences among extended alternatives.

Let us refer to the person *in* a "history" as the "subject" of that history, and the person whose extended preferences are at issue a "spectator." Individual k, qua spectator, may prefer $(x; i)$ to $(y; j)$; individual i is the subject of the first history, j the subject of the second.

We can now turn the puzzle at hand. What exactly *is* a "personal position" or "individual history"? The ontological status of a "social situation" or an outcome is straightforward; these are states of affairs. The ontological status of a "personal position" or individual history is murkier. $(x; i)$ is, in some sense, a relativization of the state-of-affairs, x, *to* the subject i. The spectator's extended preference for $(x; i)$ over $(y; j)$ is meant to explain a well-being comparison *of* the two subjects. Moreover, $(x; i)$ and $(y; j)$ must be possible arguments for preferences.

These desiderata are satisfied by conceptualizing a history as a description of individual *properties*, namely those of the subject in the outcome; and by conceptualizing an extended preference *for* a history as a preference for a certain kind of state of affairs, namely the state of affairs *that* the spectator possess those properties. On this conceptualization, for k to have an extended preference for $(x; i)$ over $(y; j)$ means: k prefers *that k* have the properties of person i in

[7] See Binmore (1994, ch. 4; 2008); Broome (1998, 2008); Gajdos and Kandil (2008); Grant et al. (2010a, 2010b); Griffin (1991); Kaneko (1984); Karni (1998); Karni and Weymark (1998); MacKay (1986); Mongin (2001); Mongin and d'Aspremont (1998); Moreno-Ternero and Roemer (2008); Ng (1999); Pattanaik (1986); Roemer (2008); Safra and Weissengrin (2003); Sen (1986, pp. 1122-1123); Suzumura (1996); Weymark (1991, 2005)

[8] See, e.g., Arrow (1977); Kolm (1996, pp. 160-168).

outcome x, rather than *that* k have the properties of person j in outcome y. Call this the "property-possession" conceptualization of an extended preference.

Harsanyi is far from precise about the ontology of "personal position," "extended alternative," and extended preference. However, the textual evidence suggests that he intends something like the "property-possession" account. For example, he suggests that an extended preference regarding some history concerns the possibility of being "put in the place"[9] or standing "in the shoes" of the subject of the history.

> [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 noneconomic variables, such as health, social status, job situation, family situation, etc.)[10]

For a spectator to be "put in the place" or "stand in the shoes" of the subject is just for the spectator to have the same attributes as the subject. In discussing the case where a spectator is formulating an extended preference between a history in which the subject eats meat and has the spectator's tastes, and a history in which the subject eats fish and has different preferences, Harsanyi characterizes the extended preferences as follows: "[They are] preferences between eating meat with one's actual taste and eating fish with a taste quite different from one's actual taste."[11] In other words, the thought experiment that the spectator undertakes in comparing these histories is to compare *his* eating meat and having certain tastes, with *his* eating fish and having other tastes.

One worry about the property-possession conceptualization of extended preferences is that an individual's well-being may depend upon others' attributes, not just her own. For example, it is often thought that my children's happiness or well-being is one determinant of *my* well-being, or at least can be. However, the property-possession account may be able to address this worry by allowing for the subject's relational as well as intrinsic properties to be included in the description of histories, and to be part of the argument for extended preferences. Spectator k's preference for $(x; i)$ over $(y; j)$ is a preference that k possess all of the intrinsic *and* relational attributes that subject i has in outcome x, rather than that k possess all of the intrinsic and relational attributes that subject j has in outcome y.

A cross-cutting problem, and the one I wish to stress, concerns essential attributes—more precisely, properties of the subject that the spectator essentially lacks (for short, EL_k properties, with the "k" subscript referring to the spectator, k). EL_k properties are properties (intrinsic or

[9] Harsanyi (1986, p. 50).
[10] Harsanyi (1982, p. 50).
[11] Harsanyi (1986, p. 53).

relational) that *k* cannot possibly possess—properties that *k* does not have in any possible world where she exists.

For example, imagine that the spectator is a woman, Sue, currently living and born in 1980. Sue is told about a possible life that Cleopatra might have led, (*x*; Cleopatra), a possible life that Shakespeare might have led, (*y*; Shakespeare), and asked to formulate her extended preferences between the two. But—on the property-possession account—this thought exercise involves Sue comparing two *impossible* states of affairs. Cleopatra was a woman born in the first century B.C.; Shakespeare, a man born in the 16th century A.D. It is possible, perhaps, for Sue to have been male—her gender (by contrast with her genetic makeup) seems not to be an essential feature of her. Yet it is impossible for Sue to have been born when Cleopatra or Shakespeare was. The precise or at least rough timing of someone's birth *is* one of her essential properties, seemingly.

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 property-possession account—involves *k*'s preferring an impossible state of affairs.

The proponent of the property-possession account has two strategies for meeting this challenge. One is a "severance" strategy, namely to characterize outcomes so that individuals have *only* attributes that any spectator can possess; or, alternatively (a less cumbersome version of the severance approach) to instruct a given spectator *k* to compare a given pair of histories (*x*; *i*) and (*y*; *j*) by comparing a state of affairs in which *k* has all the properties of *i* in *x except the* EL_k properties of *i* in *x*, 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*.

The severance strategy is problematic. Nothing in the concept of well-being would seem to preclude certain individual attributes from playing a role in determining the welfare of some subject, merely because those attributes are essential (in any sense, either essentially possessed by the subject, or essentially lacked by some spectator). Well-being, conceptually, involves subject-relative goodness: if Jim is better off than Steve in some outcome, the outcome is better for Jim than for Steve. This goodness-for relation will be grounded in Jim's and Steve's attributes – and, if so, why not (in part) their essential attributes, even if these might be essentially lacked by some spectators? Why should we rule out this possibility *ab initio*—particularly since the nature of essential attributes is itself in dispute?

Perhaps the irrelevance of essential attributes to well-being is a *metaphysical* rather than *conceptual* fact—as in the famous example of the metaphysical but not conceptual necessity that water is H_20. However, metaphysical necessities emerge (paradigmatically) by empirical investigation. We learn by scientific investigation that the molecular makeup of the wet and

thirst-quenching stuff we call "water" is H_20. The severance strategy does not render essential attributes irrelevant to well-being in this empirical manner. Rather, the nexus between such items and well-being is—implausibly—eliminated by conceptual fiat.

Consider again the Cleopatra/Shakespeare case, on the premise that birth timing is an essential property. It seems very plausible, indeed, that the comparative well-being of Cleopatra and Shakespeare depends on their consumption, social status, hedonic states, health, accomplishments, sex lives, etc., and *not* the fact that Cleopatra was born in the first century B.C. while Shakespeare in the 16th A.D. But why *is* this well-being fact true? It is because humans, contemplating possible lives, *care* about consumption, etc., and not about birth timing per se. So it is an empirical fact about human valuations that explains the welfare-irrelevance of birth timing. The severance strategy, by contrast, makes subjects' birth timing (if an EL_k property) irrelevant to spectator *k*'s ranking of histories even if birth timing is something that *k* cares about.[12]

The second (non-severance) strategy for handling essential properties—without abandoning the property-possession account of extended preferences—is to include all of the subject's properties in the states of affairs that the spectator *k* is meant to contemplate, and to explain (somehow) why such contemplation is possible even if some of these properties are EL_k properties. This second strategy has different variants. (1) *Spectator Ignorance*. The spectator might be deprived of information about her own attributes (including essential attributes), or instructed to ignore such information. While it is impossible for a spectator *k* to *have* an EL_k property, it is possible for the spectator to *believe that she has* an EL_k property. Sue might not know her actual birth timing, and instead believe she was born when Cleopatra was. Or, she might be uncertain about her actual birth timing, and thus be able to contemplate as an epistemic possibility both that she was born in the first century B.C., and that she was born in the 16th A.D. However, asking spectators to rank histories under a condition of ignorance about their own attributes is in serious tension with the desideratum that the preferences constitutive of well-being be fully informed. (2) *"Conceivability*". Some impossible states of affairs are still "conceivable." For example, either Goldbach's conjecture is true and necessarily so, or it is necessarily false;[13] but both propositions seem conceivable.[14] Moreover, the conceivability of impossible states of affairs *might* transcend the epistemic considerations discussed under the heading of spectator ignorance; perhaps even a perfectly informed spectator might conceive both the truth and the falsity of Goldbach's conjecture. But can Sue, knowing for sure what her essential properties are (including, ex hypothesi, that she was born in 1980), conceive the impossible state of affairs of her being born around the time of Cleopatra or Shakespeare? (3)

[12] A second problem with the severance strategy is that the spectator, knowing his essential properties, may find it very difficult to imagine his having all of the subject's properties except the EL_k properties – because it may be *nomically* impossible (given the actual laws of physics, biology, psychology, or sociology) for the spectator to possess those properties together with his own essential properties.

[13] Mathematical propositions, if true, are necessarily true, and similarly if false.

[14] See Chalmers (2002).

The Semantics of "I." Pursing a line of analysis suggested by Zeno Vendler, we might distinguish between Sue's imagining that "Sue has all the properties of Cleopatra," and her imagining that "I have all the properties of Cleopatra." When Sue imagines the first sentence, she imagines something impossible; when Sue imagines the second (Vendler proposes) she does not.[15] However, Vendler does not succeed in clarifying what Sue *is* imagining in contemplating the second sentence, if not the first.

B. Spectator Attitude

In *Reasons and Persons*, Parfit notes that someone might prefer a state of affairs which is too spatially or temporally remote from her to make a difference to her well-being—illustrating the point with the following example.

> 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.[16]

Many other philosophers concur with Parfit —indeed, his observation that the satisfaction of certain preferences (however well-informed and rational) is not well-being enhancing has become virtually a truism in contemporary philosophical scholarship about well-being. For example, 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.[17]

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 Ceasar 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

[15] See Vendler (1976, p. 116): "Thus whereas the man known as Zeno Vendler cannot be imagined, and cannot be, the man known as Claudius [the Roman Emperor], I, who am known as Zeno Vendler, still can imagine being, and could be, the man known as Claudius."

[16] Parfit (1987, p. 494).

[17] Darwall (2002, p. 53).

> that I am not an astronomer whose life work has been devoted to a theory that would be confirmed or refuted by this fact).[18]

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 fulfillment of this desire would in any way make my life go better."[19]

Not only can someone's preference-satisfaction and well-being diverge in the case where she has moral, altruistic, aesthetic, or otherwise non-self-interested preferences for a spatially or temporally remote feature of outcomes. Divergence can also occur where her ranking of outcomes differentiated only by "proximate" features is driven (wholly or partly) by non-self-interested considerations. Felix, repentant over terrible past wrongdoing, believes in a version of retributivism that requires him to serve a very long term of incarceration, and therefore prefers on moral grounds that he serve this term, even though he would be better off with a lighter sentence.

Although Harsanyi eliminates certain kinds of extended preferences, namely poorly informed and sadistic ones, he does not recognize the problem adverted to by Parfit et al.—a problem that welfare economics, more generally, has failed to confront. Nothing in Harsanyi's proposal precludes a given spectator k from comparing $(x; i)$ to $(y; j)$ with the wrong sort of attitude, and thus making the comparison by virtue of features of the component outcomes, x and y, that have little or nothing to do with the subjects' welfare. For example, k might prefer $(x; i)$ to $(y; j)$ because of the fact that x is an outcome in which cancer is cured in the 22nd century, and y an outcome in which cancer is never cured, even though each subject dies long before the 22nd century. By "the right sort of attitude," I mean a concern or perspective on the spectator's part—whatever exactly it might be—that could serve to justify the putative link between his extended preferences over histories, and the well-being ranking of those histories. The "wrong sort of attitude" is an attitude not of this sort.

Indeed, the possibility of a spectator who is rational, well-informed, non-sadistic, and otherwise normal, but has the wrong sort of attitude, can be seen to pose deep issues for Harsanyi. Let us return to a very basic point. The well-being ranking of lives is not, merely, a person-neutral goodness ordering[20] of the outcomes in which these lives occur. Whether Felix's life in x is better or worse than Jane's in y is not, merely, a matter of the comparative goodness (in a moral sense, aesthetic sense, or any other person-neutral sense) of x and y themselves. If the well-being ranking of lives *were* merely a matter of the goodness of the underlying outcomes, then in each outcome all individuals would be equally well off—which is absurd.

---

[18] Scanlon (1998, pp. 113-114).

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

[20] By "person neutral" I mean an ordering of outcomes which is not biased toward the interests, concerns, or perspective of one person rather than another.

Harsanyi's construct of extended alternatives is meant as a device for relativizing outcomes *to* subjects, and thus for differentiating between the goodness ranking of outcomes and the well-being ranking of lives. In particular, on the property-possession account, this relativization is accomplished by asking k to rank $(x; i)$ as against $(y; j)$ via a comparison of the state-of-affairs in which he possesses all the properties of i in x, to the state of affairs in which he possesses all the properties of j in y. *But the success of this relativization hinges on k's having an appropriate attitude*. Imagine that k is completely impartial between his interests and everyone else's. What he cares about is not what *his* attributes are, but only the population distribution of attributes. Then k's ranking of $(x; i)$ versus $(y; j)$ will ignore where each subject is located in the population distribution of attributes—instead comparing the two histories *just* by comparing the x distribution to the y distribution.[21] In particular, on the property-possession account, a fully impartial spectator will be indifferent between histories $(x;1)$, $(x; 2)$, … $(x; N)$ for any outcome x. This is absurd (if extended preferences are meant to explain well-being), and *not* what Harsanyi intends.

If the property-possession account were otherwise attractive, a ready solution to the attitudinal problem would be to require that the spectator be *self-interested*: to elicit her ranking of $(x; i)$ and $(y; j)$ by asking her to adopt an attitude of self-concern and, in that frame of mind, to compare the state of affairs in which she has all the attributes of i in x, to the state of affairs in which she has all the attributes of j in y. However, the property-possession account runs aground on the shoal of essential attributes; and the replacement account that I argue for in Part III will need to employ a different (and more complicated) device to ensure that the spectator has the right sort of attitude.

C. The Principle of Acceptance

An important feature of Harsanyi's proposal has not yet been discussed: the Principle of Acceptance.[22] In ranking histories belonging to the same subject, each spectator is required to respect the subject's ordinary preferences. That is, each spectator's extended preference between $(x; i)$ and $(y; i)$ is determined by the subject i's ordinary preference between x and y.

The Principle of Acceptance is infected by the "wrong sort of attitude" problem: the subject might prefer x to y on moral, aesthetic, altruistic, or other non-self-interested grounds. It also presupposes temporally and modally fixed subject preferences. But of course that presupposition may prove false. Subject i might, at one time, prefer x to y, but later prefer y to x.

---

[21] I assume that the impartial spectator reasons as follows. Ranking $(x; i)$ over $(y; j)$, in accordance with the property-possession conceptualization, he compares a state of affairs in which he possesses all of i's relational and intrinsic properties in x, to one in which possesses all of j's relational and intrinsic properties in y. But (not having a particular bias in favor of his own interests), he makes *this* comparison by considering what *everyone*'s properties are in the two outcomes, i.e., by impartially comparing the state of affairs x to y.

[22] See Harsanyi (1986) ; Weymark (1991).

Or, what subject *i* would prefer might depend upon which outcome occurs. It might be the case that, were *x* to occur, *i* would prefer *x* to *y*; but that, were *y* to occur, she would prefer *y* to *x*.

D. Homogeneous Extended Preferences

Harsanyi argues that spectators will have the same extended preferences.

> Let P_j again denote individual *j*'s *subjective attitudes* (including his preferences), and let R_j denote a vector consisting of all *objective causal variables* needed to explain these subjective attitudes denoted by P_j. Our discussion suggests that the extended utility function V_i of each individual *i* should really be written as $V_i = V_i[A_j, R_j]$ rather than as $V_i = V_i [A_j, P_j]$. Written in this form, the utility function $V_i = V_i[A_j, R_j]$ indicates the utility that individual *i* would assign to the objective position A_j if the causal variables determining his preferences were R_j. Because the mathematical form of this function is defined by the basic psychological laws governing people's choice behavior, this function V_i must be the same for all individuals *i*, so that, for example, $V_h[A_j, R_j] = V_i[A_j, R_j]$ for each pair of individuals *h* and *i*.[23]

As John Broome has observed,[24] this argument is flawed. Let us grant Harsanyi his premise that an individual's ordinary preferences are fully determined by certain objective attributes of her—for short, Determinism about Ordinary Preferences. And, following Harsanyi, let R_j denote the features of an individual *j* that fully explain her preferences. Finally, for a given extended alternative (A_i, P_i), let us refer to the pairing (A_i, R_i) as the counterpart "fully objective" history.

Determinism about Ordinary Preferences, plus the further premise that the spectator *k* has full information, entails that *k*'s ranking of two extended alternatives is the same as his ranking of their counterpart fully objective histories. Moreover, in this case, $v^k(.)$ can be expressed as having fully objective histories for its arguments: $v^k(A_i, P_i) = v^k(A_i, R_i)$.

But why need one spectator's ranking of two histories be the same as some *other* spectator's ranking of the two? Harsanyi's thought seems to be that a given spectator's extended preferences are *themselves* determined by objective causal variables—for short, Determinism about Extended Preferences. Using R_k now to mean the objective features of *k* that account for *all* the aspects of his preference structure—both his ordinary and extended preferences—this means that spectator *k*'s extended preferences can be summarized in an extended utility function[25] fixed by those features. In other words, if $v^k(.)$ is *k*'s extended utility function, $v^k(.) = O(R_k)$. $O(.)$, here, is a mapping *from* the features of a given person, *k*, which determine his tastes (R_k) *to* an extended utility function.

Determinism about Ordinary Preferences and about Extended Preferences, together, mean that the extended utility number a given spectator assigns to a given history is equal to the number he assigns to its fully objective counterpart; and that these numbers, themselves, are

[23] Harsanyi (1986, p. 58).
[24] See Broome (1998).
[25] Unique up to a positive affine transformation.

determined by the spectator's preference-determining attributes. Formally, $v^k(.) = O(R^k)$, and $v^k(A_i, P_i) = v^k(A_i, R_i)$. But nothing here shows that $v^k(.) = v^h(.)$ for any two spectators k and h – the conclusion that Harsanyi wrongly leaps to at the end of the paragraph above. The determination of a spectator's extended preferences *by* his attributes hardly means that spectators have the same extended preferences. After all, spectators can have different preference-determining attributes. Which flavor ice cream I prefer may be fully explainable by my genetics; but if my genes are different from yours, I might crave chocolate while you vanilla. Nothing here changes if the arguments for preferences are more abstruse—extended alternatives, pairings of objective and subjective attributes, or their fully objective counterparts, understood (on the property-possession account) as complexes of properties that the spectator might possess, or in some other manner.

III. **Harsanyi 2.0: The Revised Account**

The account proposed here (for short, the "Revised Account") retains the most fundamental aspects of Harsanyi's framework: the concept of extended preferences; the representation of extended preferences by extended utility functions; the use of such functions as the basis for interpersonally comparable utilities. The Revised Account also retains a further (arguably less fundamental) feature of his approach, namely the vNM premise.

However, the account abandons the property-possession conception of extended preferences, and instead employs an attitudinal conception: a given spectator's extended preferences are equated with his ordinary preferences under a suitable attitude. At one stroke, this shift resolves the first two difficulties with Harsanyi's account described in the previous part, i.e., the problem of essential attributes and the possibility of spectators having the "wrong attitude." The third difficulty, namely the possibility of variable subject preferences, can be handled by conditionalizing or dropping the principle of acceptance; the fourth, the possibility of heterogeneous spectator preferences, by pooling such preferences.

I discuss these moves in turn. I then defend the vNM premise. This feature of Harsanyi's view has been controversial. Why do I retain it, even while I carve away other elements?

A. <u>The Attitudinal Conception of Extended Preferences</u>

The history $(x; i)$ is, in some sense, a relativization of the outcome x *to* the subject i. The property-possession conception is one (unsuccessful) attempt to achieve this relativization—but there is a different way to do so, namely by analyzing extended preferences as ordinary preferences with an appropriate attitudinal constraint. Call this analysis the "attitudinal conception" or "attitudinal approach." It is this conception of extended preferences that comprises the core of the Revised Account.

Consider, first, a given spectator k's within-subject ranking of histories: his preferences over histories all of which have the same subject. Spectator k is ranking the subset of histories of the form $(x; i)$, $(y; i)$, $(z; i)$ …. The Revised Account analyzes such preferences in terms of k's

ordinary preferences under a condition of wholehearted sympathy with the subject. In other words, it says the following:

> The Revised Account: A Spectator's Within-Subject Ranking of Histories and Lotteries
>
> Spectator *k* has an extended preference for (*x*; *j*) over (*y*; *j*) just in case spectator *k*, under a condition of wholehearted sympathy with subject *j*, has an ordinary preference for *x* over *y*.
>
> Similarly, imagine that *L* is one lottery over subject *j*'s histories, and *L** another such lottery. *M* is the outcome lottery corresponding to *L* (in other words, if *L* assigns a given probability to history (*x*; *j*), *M* assigns that probability to outcome *x*). *M** is the outcome lottery corresponding to *L**. Then *k* has an extended preference for *L* over *L** just in case *k*, under a condition of wholehearted sympathy with subject *j*, has an ordinary preference for *M* over *M**.

This account (which draws from a long philosophical tradition of analyzing well-being or morality in terms of the attitude of sympathy) presupposes that sympathy is a natural human attitude that is directed at other persons, and can be partial or wholehearted. At a given time, I may be sympathetic to multiple persons, or just to one.[26]

Note how the attitudinal conception relativizes histories *to* subjects. Spectator *k*, in ranking (*x*; *j*) and (*y*; *j*), is not just ranking the underlying outcomes, *x* and *y*, in any old manner. Rather, *k* is ranking those outcomes *with her care and concern targeted at j*. Similarly, *k*, in ranking (*x*; *i*) and (*y*; *i*), is ranking those outcomes with her care and concern targeted at *i*.

Note also how this conception solves the essential-attribute problem. The spectator *k* is not asked to imagine that she possesses the subject's attributes; and so there is no difficulty in understanding how an attribute which is essentially lacked by the spectator (an EL_k attribute) might nonetheless figure in her ranking of histories. To see in a simple way how this might happen, imagine that each person's DNA is an essential feature of him. Imagine, further, that spectator *k*'s DNA is different from subject *j*'s, in turn different from subject *i*'s. It is conceptually possible (although, to be sure, empirically implausible) that: *k*'s ranking of outcomes under a condition of sympathy with *j* is different from her ranking of outcomes under a condition of sympathy with *i*, and moreover that this difference is (wholly or partly) due to the fact that the two subjects have different DNA.

Finally, note how the attitudinal conception solves the "wrong attitude" problem. A spectator's ranking of histories cannot, in general, be equivalent to her *moral* ranking of the component outcomes—for example—or to her aesthetic ranking. Rather, the spectator's ranking must involve some concern or perspective on her part that serves to justify the putative link

[26] See generally Darwall (2002).

between extended preferences and the well-being ranking of histories. But *sympathy* is exactly the right attitude to justify this linkage. As Stephen Darwall has explored at length, sympathy is responsive to judgments of well-being.[27] If I am wholeheartedly sympathetic to you, and I judge that you are better off in x than y, then I prefer x to y. In particular, then—unless k makes the unlikely judgment that subject i's well-being is affected by events which are temporally or spatially remote from i—k will rank (x; i) and (y; i) as equal if the component outcomes, x and y, are differentiated only by such events.

It is also worth observing that the attitudinal conception bears out an important truism about well-being: that there is a close connection (of some kind) between well-being and *self-interest*. Consider a given spectator's ranking of his own life-histories. What the attitudinal conception says, in this case, is the following: k has an extended preference for (x; k) over (y; k) just in case k, under a condition of wholehearted sympathy *with himself*, prefers x to y. But wholehearted self-sympathy is nothing other than self-interest.

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 within-subject extended preferences are just rankings of outcomes with this generalized self-interest directed at the appropriate person, the subject.

I have said nothing, yet, about across-subject extended preferences. What does it mean for spectator k to have an extended preference for (x; i) over (y; j), where i and j are different persons? Elucidating the content of such preferences turns out to be essential in constructing an extended utility function that makes determinate interpersonal comparisons of well-being levels and differences. To see the difficulty here, note that if $v^k(.)$ represents k's within-subject extended preferences, there can easily be a $v^k{*}(.)$ which is not a positive affine transformation of $v^k(.)$ and yet which represents k's within-subject extended preferences equally well. Moreover, $v^k{*}(.)$ and $v^k(.)$ may well imply different interpersonal comparisons.

	Utility function $v^k(.)$			Utility function $v^{k*}(.)$		
	x	y	z	x	y	z
Subject 1	100	300	500	1	3	5
Subject 2	6	7	9	60	70	90

Utility function $v^{k*}(.)$ takes each history for subject 1—histories (x; 1), (y; 1), and (z; 1)—and assigns it a utility which is $1/100^{th}$ the utility assigned to that history by function $v^k(.)$. It takes each history for subject 2—histories (x; 2), (y; 2), and (z; 2)—and assigns it a utility which is 10 times the utility assigned to that history by function $v^k(.)$. It is easy to see that $v^{k*}(.)$ represents spectator k's within-subject preferences just as well as $v^k(.)$—

[27] See ibid.

because it multiplies each history by a subject-specific positive constant—but that it is not a positive affine transformation of $v^k(.)$.

An analysis of across-subject extended preferences in terms of sympathy is unsuccessful. We cannot say that *k* has an extended preference for (*x*; *i*) over (*y*; *j*) just in case *k*, under a condition of wholehearted sympathy with the subject, prefers *x* to *y*—because there is no one particular person to be the target of *k*'s care and concern. What mental operation is Sue supposed to go through when we ask her whether she prefers *x* under a condition of sympathy with Jean, to *y* under a condition of sympathy with Leslie? Asking Sue to make this comparison seems incoherent, in much the same way that it would be to ask her whether she prefers *x* when scared, to *y* when calm.

I therefore offer a different approach. Let us say that spectator *k* has an extended preference for (*x*; *i*) over (*y*; *j*), with *i* and *j* different subjects, just in case *k* judges that *i* in *x* is better off than *j* in *y*. The following can be shown: If *k* makes a small number of such across-subject judgments, and $v^k(.)$ represents such judgments as well as *k*'s within-subject extended preferences, $v^k(.)$ *will* normally be unique up to a positive affine transformation and *will* yield determinate interpersonal comparisons of levels and differences.[28]

The Revised Account might be criticized as furnishing a circular analysis of the concept of well-being. One kind of circularity is patent: the account purports to analyze one aspect of that concept, interpersonal comparisons, by positing that "spectators" have across-subject extended preferences; and yet such preferences are, in turn, understood by using the very concept being analyzed (namely, by supposing that spectators arrive at these preferences by making judgments of well-being). Another kind of circularity is more subtle. Sympathy—I have suggested—is itself connected to judgments of well-being. But this connection is not merely contingent. It is not as if sympathy is some independently specified attitude which, as is happens, motivates the holder to act in line with his judgments of well-being. Rather, the attitude of sympathy is just that attitude which plays the following role: if *P* is wholeheartedly sympathetic to *S* and judges *x* to be better for *S*'s well-being than *y*, *P* is motivated to pursue *x* rather than *y*. Insofar as the concept of well-being is used to identify the attitude of sympathy, which in turn is used—by the attitudinal approach—to make sense of within-subject extended preferences, the approach is circular.

But the circularities, here, are not vicious. They arise because well-being is a self-referential property. A property is self-referential if its instantiation depends upon how individuals use concepts expressing that very property. This phenomenon is familiar from the literature on secondary properties.[29] An object has the property of redness if it has surface reflectance characteristics which make it look red to normal observers. Note, here, how persons' perceptions or judgments of redness are an ineliminable aspect of the property of redness. Similarly (it seems) individuals' judgments of well-being are an ineliminable aspect of the

[28] See Adler (2011, ch. 3).
[29] See Darwall, Gibbard, and Railton (1992).

property of well-being. In particular, one person has the relational property of being better off than a second only if the two have various attributes (whatever exactly they may be) such that individuals generally judge the first to be better off than the second.

Necessarily, an accurate "account" of a self-referential property *cannot* be a reductive account—one that identifies the property without using its concept. However, an accurate account of a self-referential property can still be very useful—and the circularities inevitably arising in such an account are, therefore, not vicious. For example, the account may help us pick out the instances when the property is instantiated or (more generally) identify the physical, chemical, biological, or psychological supervenience base for such instantiation. And that is exactly what the attitudinal conception of extended preferences is meant to do. It is meant to help us construct a numerical device (utility functions) that will signal whether one person is better off than a second, or one well-being difference greater than a second, depending upon the health, consumption, social life, status, and other characteristics of the persons involved. That this device is constructed by asking individuals to make certain judgments of well-being is no bar to its construction, and the dependence of this construction on such judgments is no objection to it. (By analogy, the blueprint for building a successful instrument to detect redness might well involve asking observers which objects look red.)

B. Conditionalizing or Dropping the Principle of Acceptance

The principle of acceptance offers an alternative route to defining within-subject extended preferences. Rather than looking to spectators' ranking of outcomes under a condition of wholehearted sympathy with the subject, the principle—as modified to require an appropriate attitude on the *subject*'s part—says: subject *i*'s self-interested preferences between *x* and *y* determine what any spectator's extended preferences between (*x*; *i*) and (*y*; *i*) are. But the principle is problematic, even in this modified form, because *i*'s preferences might vary over time or across outcomes.

The Revised Account might accommodate variable subject preferences by (a) conditionalizing or (b) dropping the principle of acceptance. The conditionalized principle says: *If* subject *i* at all times and in all outcomes self-interestedly prefers *x* to *y* (for short, the subject "steadfastly" self-interestedly prefers *x* to *y*), *then* any spectator must have an extended preference for (*x*; *i*) over (*y*; *i*). Otherwise, spectator *k*'s extended preferences between these two histories depend upon his outcome preferences under a condition of wholehearted sympathy with *i*.

Dropping the principle of acceptance makes the spectator's outcome preferences, under this condition, determinative in all cases. The spectator is allowed to prefer (*x*; *i*) to (*y*; *i*) even if the subject has a steadfast self-interested preference for *y* over *x*.

A live question in the theory of well-being—one of the aspects of the debate about whether well-being consists in "objective goods"—concerns the extent to which individuals are

sovereign with respect to their own well-being. If someone (under conditions of good information and rationality) favors one outcome over another, does that make the first outcome better for him? The Revised Account with a conditionalized principle of acceptance is "sovereignty-respecting" in quite a strong way—as we shall see more clearly in a moment, once the account's mechanism of pooling utility functions to arrive at well-being comparisons has been presented. However, the Revised Account without the principle is also sovereignty-respecting, albeit more weakly. First, spectators are still *permitted* to follow subjects' steadfast preferences. It might be an empirical fact about human psychology that—when a subject has a steadfast self-interested preference for x over y—spectators will tend to prefer x to y under a condition of wholehearted sympathy with the subject. (In other words, as an empirical matter, spectators might tend to be deferential to subjects). Second, even if spectators do not tend to be deferential, the pooling mechanism makes the approach sovereignty respecting in at least a weak sense.

C. Pooling Extended Utility Functions

For simplicity, assume a fixed population of interest, each member of which exists in all the outcomes under consideration. What have we established to this point? Each individual k in this population has an extended utility function $v^k(.)$ which represents his within-subject and across-subject extended preferences (as those are understood by the Revised Account); and this extended utility function is normally unique up to a positive affine transformation.

Harsanyi's argument for the homogeneity of extended preferences is a failure—and this argument also fails in the context of the Revised Account. One spectator might make different across-person judgments than another. Even with a conditionalized principle of acceptance, two spectators might disagree in their within-subject preferences when the subjects are not steadfast—and, a fortiori, they might disagree if the principle of acceptance is dropped. Thus, if k and l are different, there is no reason to suppose that $v^k(.)$ will be the same function as $v^l(.)$, or a positive affine transformation thereof.

In turn, the heterogeneity of extended preferences would seem to pose a grave obstacle to the construction of the interpersonal utility function $V(.)$. However, that obstacle can be avoided by reconceptualizing $V(.)$ – not as a single function, but as a *set* of functions. Let us indicate this reframing with bold-face: $\mathbf{V}$ is the set $\{v^1(.), v^2(.). \ldots v^N(.)\}$, where $v^1(.)$ is an extended utility function representing individual 1's preferences (as per the Revised Account), $v^2(.)$ an extended utility function representing individual 2's, …, $v^N(.)$ an extended utility function representing individual N's. $v^k(.)$, for each k, can be any one of the family of functions (unique up to a positive affine transformation) that represent his extended preferences.[30]

[30] This description of the construction of $\mathbf{V}$ is, in one important respect, too simplified—because it ignores the possibility of *intra*spectator variation in extended preferences. A given spectator's extended preferences can vary across times or outcomes. (Indeed, this possibility is not logically independent from the kind of temporal and modal

Inter- as well as intrapersonal comparisons of well-being levels and differences can now be defined with reference to **V**.

> Level comparisons: Individual i in outcome x is at least as well off as individual j in outcome y iff, for all $v(.)$ in **V**, $v(x; i) \geq v(y; j)$.
>
> Difference comparisons: The difference between the well-being of individual i in outcome x and the well-being of individual j in outcome y is at least as great as the difference between the well-being of individual l in outcome z and the well-being of individual m in outcome w iff, for all $v(.)$ in **V**, $v(x; i) - v(y; j) \geq v(z; l) - v(w; m)$.

This account has technical virtues. It yields a well-behaved (transitive, reflexive) ranking of levels and differences.[31]

Substantively, the Revised Account succeeds in delivering interpersonal comparisons. In the case where individuals have identical extended preferences, **V** is such that each member is a positive affine transformation of every other, and inter- as well as intrapersonal comparisons are fully determinate. (For every pair of histories, either the first is better than the second, or the second is better than the first, or the two are equally good. For every four histories, either the well-being difference between the first two is greater than the well-being difference between the second two, or vice versa, or the well-being differences are the same.) Where extended preferences diverge, **V** no longer has this structure, and there may be "pockets" of indeterminacy in the ranking of levels and differences—the size of which depends upon the extent of divergence in these preferences.

Moreover, the Revised Account fulfills two desiderata for a theory of well-being. On the one hand, interpersonal comparisons must have intersubjective validity: No single person's preferences or judgments suffice to establish interpersonal well-being facts. Indeed, according to the Revised Account, well-being comparisons (both intra-[32] and interpersonal) rest upon

---

variation which *has* been discussed, namely that *subjects* might have variable self-interested preferences. Note that if individual k's self-interested preferences as between outcomes x and y are not fixed, it follows that k's extended preferences as between $(x; k)$ and $(y; k)$ will not be fixed, either—given how the Revised Account defines extended preferences.)

Where *intra*spectator variation in extended preferences occurs—where k's extended preferences at time t and in outcome x are different from his preferences at time t^* and in outcome y—we need to include in **V** both an extended utility function representing the first preferences, and an extended utility function representing the second. See Adler (2011, ch. 3).

[31] To be precise, consider the binary relation *Lev* on the set **H** of life-histories, defined as follows: $(x; i)$ *Lev* $(y; j)$ iff, for all $v(.)$ in **V**, $v(x;i) \geq v(y; j)$. *Lev* is a quasiordering. Now consider the binary relation *Diff* on the set **H** x **H**, the set of all pairs of life-histories, defined as follows: $((x; i), (y; j))$ *Diff* $((z; l), (w; m))$ iff, for all $v(.)$ in **V**, $v(x; i) - v(y; j) \geq v(z; l) - v(w; m)$. *Diff* is a quasiordering on **H** x **H**.

[32] It might be wondered why the Revised Account needs to analyze *intra*personal comparisons with reference to **V**. Why not, instead, say that within-subject level and difference comparisons are just a matter of that person's preferences?

If the Revised Account incorporates a conditionalized principle of acceptance, the two approaches amount to the same thing. Alternatively, the very same "objective-good" intuitions about well-being that might motivate

universally shared judgments and preferences. One history is at least as good as a second only if *everyone's* extended utility function assigns the first a utility number at least as large. In the case where the two histories belong to the same subject, this means that everyone (under a condition of wholehearted sympathy with the subject) prefers the first outcome or is indifferent; in the case where the subjects are different this means that everyone judges the first history to be at least as good.[33]

The second desideratum is that an account should be "sovereignty respecting," at least to some extent. This is a constraint upon how the account makes *intra*personal comparisons. It can now be seen that the Revised Account– even though it makes intra- as well as interpersonal comparison by pooling extended utility functions – *is*. Consider the case where individual *i* has a steadfast self-interested preference for outcome *x* over *y*. With the conditionalized principle of acceptance on hand, this fact suffices to establish that *the individual is better off in x*: every utility function in **V**, conforming to that principle, will assign (*x*; *i*) a higher utility than (*y*; *i*). Even without the principle of acceptance (and even if spectators are not deferential to subjects as an empirical matter), this fact suffices to establish that the individual is *not* better off in *y*. The individual's own extended utility function will assign a lower number to (*y*; *i*) and thus it will not be the case that every utility function in **V** assigns (*y*; *i*) a value at least as large as the value it assigns (*x*; *i*).

Finally, it can be asked: is the Revised Account true to the rationale for preference-based views of well-being? What *that* rationale is can itself be contested—but, plausibly, it is the *motivational* features of well-being that warrant an account thereof linked to preferences. To say that someone is necessarily motivated by her own well-being is too strong; after all, the individual's actual goals might be other-regarding. But, more carefully, we *can* say this: An account of well-being should be such that an individual *is* motivated by her welfare, when she is self-interested – and that others who care about her are also thus motivated.

The account of well-being I have proposed fulfills this motivational requirement. How? Note that if (*x*; *i*) is better than (*y*; *i*), at least some individuals have extended utility functions that assign a higher number to (*x*; *i*), and none have extended utility functions that assign a higher number to (*y*; *i*). Because these utility functions are derived from extended *preferences* (analyzed in turn in terms of ordinary preferences under a condition of sympathy with the subject), it follows that: if (*x*; *i*) is better than (*y*; *i*), at least some of those who care about *i* are motivated to pursue *x* rather than *y*, and none who do are motivated to pursue *y* rather than *x*.

D. Why the vNM Premise?

---

rejecting even a conditionalized principle of acceptance would also cut against the proposal to analyze within-subject comparisons by looking just at the subject's (possibly idiosyncratic) preferences.

[33] Similarly, on the Revised Account, difference comparisons rest upon universally shared judgments and preferences.

On Harsanyi's account, a spectator's extended utility function is a vNM function: it expectationally represents her preferences regarding lotteries over histories. This feature is retained by the Revised Account, as follows. For any spectator k, $v^k(.)$ is such that: for any lottery L over one subject's histories, and any other lottery L^* over that same subject's histories, k prefers L to L^* just in case the expected value of L according to $v^k(.)$ is greater than the expected value of L^* according to $v^k(.)$.

But why should the utilities that represent well-being differences or levels be grounded in lottery preferences?[34] Comparisons of well-being levels and differences have nothing, essentially, to do with uncertainty. At the limit, every outcome might be a *possible world*—a maximally specified state of affairs—and yet, on Harsanyi's account, and my own, lottery preferences would still be invoked to explain well-being comparisons with respect to these outcomes.

vNM utilities have a key technical advantage: uniqueness up to a positive affine transformation. My construction, like Harsanyi's simpler approach, indeed yields a $v^k(.)$ that is unique up to a positive affine transformation; and this in turn means that (absent heterogeneity in extended preferences) well-being level and difference comparisons are fully determinate. But there are other ways—without invoking lottery preferences—to arrive at a $v^k(.)$ unique up to a positive affine transformation. Consider the following approach. Spectator k's within-subject extended preferences with respect to histories are analyzed as before: k prefers $(x; i)$ to $(y; i)$ just in case k, under a condition of wholehearted sympathy with i, prefers x to y. Such preferences, plus information about k's across-subject judgments—to the effect that $(x; i)$ is better than $(y; j)$, the two subjects different-- yield a $v^k(.)$ unique up to an ordinal transformation.[35] We then appeal to k's *judgments about well-being differences*[36], eliminating ordinal transformations of $v^k(.)$ that do not conform to these judgments, and arriving eventually at a $v^k(.)$ unique up to a positive affine transformation.

I see no argument that decisively refutes this alternative approach. Still, the following considerations can be marshaled against it: (1) *Are the relevant judgments possible*? The approach presupposes that individuals can make *direct* judgments of well-being differences—"direct" in the sense of being underived from lottery preferences. It is contestable whether this is true. (2) *Complexity*. The account is more complex than the Revised Account, which constructs $v^k(.)$ using the twin elements of sympathetic preferences and well-being judgments, not the third element of direct difference judgments. (3) *Motivational Link*. The Revised Account retains a

[34] On this issue, see Broome (1995); Ng (1999); Risse (2002); Roemer (2008); Sen (1976; 1986, pp. 1122-1123); Weymark (1991, 2005).

[35] Assume that $v(.)$ is an extended utility function on the set **H** of life-histories. To say that another function $w(.)$ is an "ordinal transformation" of $v(.)$ means: there exists some strictly increasing function $g(.)$ such that, for every $(x; i)$ in **H**, $w(x; i) = g(v(x; i))$. To say that a $v^k(.)$ which represents k's extended preferences (of some sort) is "unique up to an ordinal transformation) means: some other $v^{k*}(.)$ also represents those extended preference iff $v^{k*}(.)$ is an ordinal transformation of $v^k(.)$.

[36] See Abdellaoui, Barrios, and Wakker (2007).

strong link to motivation. According to that account (but not the alternative proposal now on the table) *every* intrapersonal fact about a subject's well-being, both level facts and difference facts, entails a corresponding motivational fact (about the motivation of individuals who care about the subject). (4) *Inferences*. On the Revised Account, all information regarding *k*'s within-subject ranking, relevant to the construction of $v^k(.)$, can be inferred from his choices. This is not true of the alternative account.[37] (5) *No intuitions*. Although the critical literature on Harsanyi stresses the *theoretical* point that $v^k(.)$ need not be derived from lottery preferences, no concrete example of how such a derivation is counterintuitive has been provided.

IV. **Social Welfare Functions**

The Revised Account represents inter- as well as intrapersonal comparisons of well-being via a set **V**, pooling extended utility functions. I have said nothing, yet, about how this utility information is to be integrated into a social welfare function (SWF). The SWF tradition, remember, ranks outcomes based upon individual utility numbers, which are taken to be interpersonally comparable; yet the theoretical basis for these numbers remains obscure. An account of interpersonal well-being measurement, if successful, will fill this gap. Does the Revised Account do so?

The traditional definition of an SWF is as follows: it is some rule *R* for ranking the utility vectors that correspond to outcomes.[38] The interpersonally comparable utility function $V(.)$ maps a given outcome *x* onto a list, or "vector," of *N* numbers representing the well-being of each of the *N* persons in the population. $V(x) = (V_1(x), \ldots, V_N(x))$. Any two outcomes are then compared by using *R* to rank their respective vectors: *x* at least as good as *y* iff $V(x)$ ranked by *R* at least as good as $V(y)$. *R* might be a *utilitarian* SWF, which compares vectors by summing unweighted utilities. However, *R* need not be utilitarian. Indeed, various standard non-utilitarian views in normative ethics have their counterpart SWFs. *R* might be a *prioritarian* SWF, which sums concavely transformed utilities; a *leximin* SWF, which gives absolute priority to worse-off individuals; or a *sufficientist* SWF, which uses a threshold to mark the point of absolute priority, and is utilitarian in making tradeoffs between above-threshold individuals and prioritarian in making tradeoffs between below-threshold individuals.[39]

The argument for the Revised Account of interpersonal comparisons, if persuasive, also shows that this traditional definition of an SWF needs revision. The definition presupposes that a *single* utility function $V(.)$ fully captures the well-being information that a utilitarian,

---

[37] On the Revised Account, *k*'s extended preferences regarding the histories of subject *i* (perhaps himself) are defined in terms of *k*'s ranking of outcomes and outcome-lotteries under a condition of sympathy with *i*. This ranking can—in principle—be inferred from *k*'s choices under a condition of sympathy with *i*. By contrast, nothing about *k*'s choices reveals his ranking of differences between histories (even histories belonging to one subject, even himself).

[38] See generally Blackorby, Bossert, and Donaldson (2005, chs. 3-4); Mongin and d'Aspremont (1998); Bossert and Weymark (2004).

[39] See ibid. On the possibility of a sufficientist SWF, see Brown (2005).

prioritarian, sufficientist, or leximiner would need to rank outcomes. But the argument for the Revised Account denies that. Instead (given the possible heterogeneity of extended utility functions), we will generally need a non-singleton set **V** in order to numerically represent facts about well-being levels and differences. **V** can be a singleton only in the limiting case where extended preferences are identical.

However, the requisite amendment to the traditional definition is straightforward. An SWF is the following recipe for ranking outcomes, which integrates a rule *R* for ordering utility vectors, and a set **V** of utility functions. **V** consists of scalar-valued utility functions, which map histories onto real numbers, and this set represents the well-being levels of histories and differences between them as per the representational rules discussed above.[40] Each member $v(.)$ of **V** is also naturally associated with a *vector-valued* function defined on outcomes,[41] namely: $v(x) = (v(x; 1), v(x; 2), \ldots, v(x; N))$. And an SWF combines *R* and **V** to rank outcomes, as follow:

> Outcome *x* at least as good as outcome *y* iff, for all $v(.)$ in **V**, $v(x)$ ranked by *R* at least as good as $v(y)$.

The "better than" and "equally good as" relations between outcomes can be derived from this "at least as good as" relation in the standard way, namely: *x* better than *y* iff *x* at least as good as *y* and not *y* at least as good as *x*; *x* and *y* equally good iff *x* at least as good as *y* and *y* at least as good as *x*.

SWF scholarship typically imposes certain minimal constraints on the rule *R*.[42] It is straightforward to show that—if *R* is required to satisfy these constraints--the recipe in the previous paragraph produces a "well-behaved" ranking of outcomes, one satisfying the accepted transitivity, symmetry/asymmetry, and reflexivity/irreflexivity properties of the "better than" and "equally good as "relations. It can also be shown that, with *R* thus constrained, this recipe satisfies what John Broome calls the "principle of personal good."[43] If each person is equally well off in *x* and *y*, then *x* and *y* are equally good. If everyone is at least as well off in *x* as compared to *y*, and at least one is better off, then *x* is better than *y*.[44]

---

[40] See Section III.C.

[41] I see little risk of confusion in using "$v(.)$" to denote this function as well.

[42] Those constraints are as follows: (1) *Complete ordering*. *R*, used to rank any given set of utility vectors, produces a complete ordering of that set. (2) *Anonymity*. If $v(x)$ is a permutation of $v(y)$, *R* ranks $v(x)$ and $v(y)$ as equally good. (3) *Strong Pareto*. If every entry in $v(x)$ is at least as large as its corresponding entry in $v(y)$, with at least one strictly larger, *R* ranks $v(x)$ as better than $v(y)$.

[43] Broome (1995).

[44] (1) Assume that each person is equally well off in *x* and *y*. Because **V** represents well-being via the representational rules discussed in Section III.C., it follows that, for each *i* in the population, $v(x; i) = v(y; i)$ for all $v(.)$ in **V**. Thus the vector $v(x)$ is the same as the vector $v(y)$ for every $v(.)$ in **V**, and thus $v(x)$ and $v(y)$ are ranked equally good by *R* for every $v(.)$ in **V**. (2) Assume that everyone is at least as well off in *x* as *y*, and at least one is better off. Then, for each *i*, $v(x; i) \geq v(y; i)$ for all $v(.)$ in **V**, and moreover there is some *j* such that $v^*(x; j) > v^*(y; j)$

To be sure, the recipe need *not* produce a complete ranking of outcomes, except in the limiting case where **V** is a singleton. Otherwise, there may well be a pair of outcomes, such that it is *not* the case that one is better than the other, nor is it the case that the two are equally good. But this is not troubling per se.

This amended account of an SWF—like the traditional definition—is agnostic about the form of the rule *R*, beyond requiring that it satisfy the minimal constraints. Although Harsanyi is, famously, utilitarian—he presents his account of interpersonal comparisons within the context of a defense of utilitarianism—these two aspects of his view are logically separable. He arrives at utilitarianism by combining that account with the *further* claim that an individual's *moral* preferences over outcomes are identical to her extended preferences over the corresponding equiprobability lotteries.

> Now if individual *i* wants to make a moral value judgment about the merits of alternative social situations *A*, *B*, …, he must make a serious attempt not to assess these social situations simply in terms of his own personal preferences and personal interests but rather in terms of some impartial and impersonal criteria. ….
>
> Individual *i*'s choice among alternative social situations would certainly satisfy this requirement of impartiality and impersonality, if he simply *did not know in advance* what his own social position would be in each social situation—so that he would not know whether he himself would be a rich man or a poor man, a motorist or a pedestrian, a teacher or a student, a member of one social group or a member of another social group, and so forth. More specifically this requirement would be satisfied if he thought that he would have an *equal probability* of being *put in the place* of any one among the *n* individual members of society ….

Let us capture this further aspect of Harsanyi's view—relaxed to allow for heterogeneity in extended preferences—in the Equiprobability Premise. Use the symbol E_z to mean the equiprobability lottery corresponding to outcome *z*: the lottery over life-histories that gives a 1/*N* chance of life history (*z*;1), a 1/*N* probability of life-history (*z*; 2), …, a 1/*N* probability of life-history (*z*; *N*), with *N* individuals in the population.

> <u>The Equiprobability Premise</u>
>
> Outcome *x* is morally at least as good as outcome y iff every spectator either has an extended preference for E_x over E_y, or is indifferent between them.[45]

---

for some $v^*(.)$ in **V**. Because *R* satisfies strong Pareto, it follows that $v(x)$ is ranked by *R* at least as good as $v(y)$ for all $v(.)$, and that $v^*(x)$ is ranked by *R* as better than $v^*(y)$.

[45] The Equiprobability Premise purports to state the conditions under which one outcome is *at least as good as* (better than or equally good as) a second, and thus those conditions are framed in terms of a universal spectator attitude of preference for the first *or* indifference.

Note also that, on the Revised Account, spectators are not asked to rank lotteries except in the context of making within-subject judgments. So one might ask: what does it mean for a spectator to have an extended preference for E_x over E_y? After all, E_x is a lottery over histories with different subjects, and so too is E_y. Here, we can provide a *derivative* definition. Where $v^k(.)$—unique up to a positive affine transformation-- represents *k*'s

It can be shown that the Equiprobability Premise, together with the Revised Account of interpersonal comparisons and the amended definition of an SWF, requires the SWF to be utilitarian. The following argument is valid.[46]

> (1) *Representability of Well-Being via a set* **V**. There is a set **V** of utility functions that assign numbers to histories, and it thereby represents well-being comparisons of levels and differences via the set-valued rules mentioned earlier.
>
> (2) *The Amended Definition of an SWF*. Outcome x is morally at least as good as outcome y iff, for all $v(.)$ in **V**, $v(x)$ is at least as good as $v(y)$ according to some rule R for ranking utility vectors.
>
> (3) *The Revised Account*. **V** is constructed as per the Revised Account.
>
> (4) *The Equiprobability Premise*

Conclusion:

> *Utilitarianism*. Outcome x is at least as good as outcome y iff, for all $v(.)$ in **V**, $v(x)$ is at least as good as $v(y)$ according to the utilitarian rule.

However, the conclusion obviously does *not* follow from the first three premises alone. The proponent of prioritarianism, sufficientism, leximin, or some other non-utilitarian view, can endorse the first three premises, but deny the conclusion, by denying the Equiprobability Premise.

A different argument for utilitarianism should also be noted. This concerns a troubling arbitrariness in the construction of **V**. Remember that **V** includes utility function $v^1(.)$ representing the extended preferences of individual 1, utility function $v^2(.)$ representing the extended preferences of individual 2, and so forth.

Each $v^k(.)$ is not unique. If $v^{k*}(.)$ is a positive affine transformation of $v^k(.)$, then $v^{k*}(.)$ represents k's extended preferences just as well as $v^k(.)$. Thus **V** itself is not unique. For a given **V**, consider **V*—**formed by swapping one or more of the utility functions in **V** with positive affine transformations thereof. Then: (1) the functions in **V*** represent each person's extended preferences just as well as the functions in **V**; and (2) **V*** produces the very same ranking of well-being levels and difference as **V**. For short, let us say that **V*** is an "admissible" transformation of **V**.

If the SWF is utilitarian, this non-uniqueness in **V** is not troubling. It can be shown that, if some **V** is exchanged for an admissible transformation thereof, and coupled with a utilitarian rule

---

within-subject preferences and across-subject judgments, k prefers *any* lottery L to *any* lottery L^* if the expected value of L, according to $v^k(.)$, is greater than the expected value of L^*, according to $v^k(.)$.

[46] See Adler (2011, ch. 5).

R for ranking vectors, the ranking of outcomes does not change. This is also true of the leximin SWF. However, various plausible *non*-utilitarian SWFs are not invariant to admissible transformations of **V**. For example, a prioritarian SWF that sums the square root of utilities is not, as illustrated by the following table:

	Utility function v(.)		Utility function v^*(.)	
	Outcome *x*	Outcome *y*	Outcome *x*	Outcome *y*
Individual 1	100	135	1100	1135
Individual 2	200	160	1200	1160
Utilitarian SWF	300	295	2300	2295
Sum of square root of utilities	24.14	24.27	67.81	67.75

Given utility function v(.), v^*(.) is defined as follows: $v^*(z; i) = v(z; i) + 1000$. The first row shows that utility function v(.) assigns individual 1 a utility of 100 in outcome *x* and 135 in outcome *y*; while v^*(.) assigns individual 1 a utility of 1100 in outcome *x* and 1135 in outcome *y*. The second row shows the numbers assigned by these two functions to individual 2. The row for each SWF shows the numbers it assigns to the two outcomes, according to each utility function. Note that the utilitarian SWF ranks *x* over *y* according to both v(.) and v^*(.). By contrast, the sum-of-square-root SWF ranks *y* over *x* according to v(.), but *x* over *y* according to v^*(.).

Indeed, it can be shown that a wide range of SWFs are sensitive to admissible transformations of **V**.[47]

Is this a devastating critique of the Revised Account, for those who endorse some such SWF? Does it prove that this account fails to furnish the kind of interpersonal utilities required by the prioritarian SWF or sufficientist SWF, for example? I believe the Revised Account can be defended from this attack, along the following lines. Consider some SWF which is sensitive to admissible transformations of **V**. A particular numerical level of individual utility U^* may possess a special kind of qualitative significance, for purposes of this SWF; how the SWF ranks utility vectors may change, in a qualitative manner, depending on whether the vectors contain entries above or below this number. (For example, the utility level 0 has special qualitative significance for purposes of the prioritarian SWF.[48]) Imagine, now, that some history (x^*; i^*) has just the kind of moral role corresponding to the qualitative significance of the number U^*.

[47] This follows from Theorem 13.4 in Bossert and Weymark (2004), reporting a result proved by Deschamps and Gevers (1978).

[48] More precisely, this is true of the standard "Atkinson" form of the prioritarian SWF, $(1-\gamma)^{-1}\sum_{i=1}^{N} u_i^{1-\gamma}$, which is undefined with negative utilities.

Then a set **V** which represents individuals' extended preferences, albeit "admissible" in that sense, is not admissible all-things-considered unless, for every $v(.)$ in **V**, $v(x^*; i^*) = U^*$.[49]

Unfortunately, I lack space to expand upon these brief remarks here. At a very minimum, the Revised Account *is* adequate for the needs of SWFs which are *insensitive* to "admissible" transformations of **V**—again, the replacement of one or more of the extended utility functions therein by a positive affine transformation. Such insensitivity is characteristic, not only of utilitarianism, but also of the leximin approach, as well as certain other standard SWFs[50]; and this in turn underscores that the Revised Account is severable from Harsanyi's commitment to utilitarianism.

V. **Conclusion**

Utility functions representing ordinary preferences are an insufficient basis for interpersonal comparisons. Assume that I rank a set of outcomes in some manner and you in some manner, and that my ranking is representable by a utility function, as is yours. Even the preferentialist about well-being must concede that there is nothing, yet, which establishes how my well-being level in some outcome compares to your well-being level in some outcome, or how the difference between my well-being in two outcomes compares to the difference between yours in two. Harsanyi's fertile insight was that interpersonal comparisons can be constructed by changing the *arguments* for preferences. Endow individuals with *extended* preferences, which take individual histories (not outcomes) and individual-history lotteries as their arguments; represent these extended preferences via *extended* utility functions.

This insight, however, was not adequately developed by Harsanyi. In this Article, I have both identified the gaps and flaws in Harsanyi's account of interpersonal comparisons, and suggested how they might be remedied. In particular, for an individual (a "spectator") to rank histories is *not* for her to imagine that she stands in the subjects' shoes—that she acquires their attributes—but, rather, to rank outcomes with an appropriate attitude (wholehearted sympathy) or making an appropriate judgment.[51] Nothing in the common causal basis for ordinary or extended preferences requires that extended preferences be identical. Extended preferences can be represented by a set **V** of extended utility functions (which becomes a singleton *only* in the limiting case where everyone's extended preferences happen to be identical); and although this set can provide the input for a utilitarian SWF, the SWF can also be non-utilitarian.

Sources

[49] See Adler (2011, ch. 3), "zeroing out" extended utility functions by requiring that they assign 0 to nonexistence; Blackorby, Bossert, and Donaldson (2005, pp. 37-38).

[50] In particular, it is true of the "rank-weighted" SWF, which is closely related to the Gini coefficient.

[51] In the case where the subjects of two histories are identical, the spectator's extended preferences regarding the histories are equivalent to his ranking of the component outcomes with an attitude of sympathy directed towards the subject. In the case where the subjects are non-identical, the spectator ranks the histories by making a judgment regarding the comparative well-being of the two subjects in the component outcomes. See Part III.A.

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. 2011. *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford: Oxford University Press (forthcoming).

Arneson, R.J. 1999. "Human Flourishing versus Desire Satisfaction." In E.F. Paul, F.D. Miller, and J. Paul, eds., *Human Flourishing*, pp. 113-142. Cambridge: Cambridge University Press.

Arrow, K.J. 1977. "Extended Sympathy and the Possibility of Social Choice." *The American Economic Review* 67: 219-225.

Bernstein, M. 1998. "Well-Being." *American Philosophical Quarterly* 35: 39-55.

Binmore, K.G. 1994. *Game Theory and the Social Contract*. Volume 1, *Playing Fair*. Cambridge, MA: MIT Press.

———. 2008. "Naturalizing Harsanyi and Rawls." In M. Fleurbaey, M. Salles, and J.A. Weymark, eds., *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pp. 303-333. Cambridge: Cambridge University Press.

Blackorby, C., W. Bossert, and D. Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press.

Bossert, W., and J.A. Weymark. 2004. "Utility in Social Choice." In S. Barberà, P.J. Hammond, and C. Seidl, eds., *Handbook of Utility Theory*, vol. 2 (*Extensions*), pp. 1099-1177. Boston: Kluwer Academic.

Brandt, R.B. 1998. *A Theory of the Good and the Right*. Amherst: Prometheus Books. First published in 1979 by Oxford University Press.

Broome, J. 1995. *Weighing Goods: Equality, Uncertainty and Time*. Paperback edition. Oxford: Basil Blackwell. First published in 1991.

———. 1998. "Extended Preferences." In C. Fehige and U. Wessels, eds., *Preferences*, pp. 271-287. Berlin: Walter de Gruyter.

———. 2008. "Can There Be a Preference-Based Utilitarianism?" In M. Fleurbaey, M. Salles, and J.A. Weymark, eds., *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pp. 221-238. Cambridge: Cambridge University Press.

Brown, C. 2005. "Priority or Sufficiency . . . or Both?" *Economics and Philosophy* 21: 199-220.

Chalmers, D.J. 2002. “Does Conceivability Entail Possibility?” In T.S. Gendler and J. Hawthorne, eds., *Conceivability and Possibility*, pp. 145-200. Oxford: Oxford University Press.

Darwall, S.L. 2002. *Welfare and Rational Care*. Princeton: Princeton University Press.

Darwall, S.L., A. Gibbard, and P. Railton. 1992. "Toward Fin de Siècle Ethics: Some Trends." *The Philosophical Review* 101: 115-189.

Deschamps, R., and L. Gevers. 1978. “Leximin and Utilitarian Rules: A Joint Characterization.” *Journal of Economic Theory* 17: 143-163.

Gibbard, A. 1986. "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of a Life." In J. Elster and A. Hylland, eds., *Foundations of Social Choice Theory*, pp. 165-193. Cambridge: Cambridge University Press.

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.

———. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63: 309-321.

———. 1982. "Morality and the Theory of Rational Behaviour." In A. Sen and B. Williams, eds., *Utilitarianism and Beyond*, pp. 39-62. Cambridge: Cambridge University Press.

———. 1986. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Paperback edition. Cambridge: Cambridge University Press. First published in 1977.

Hausman, D.M., and M.S. McPherson. 2009. "Preference Satisfaction and Welfare Economics." *Economics and Philosophy* 25: 1-25.

Gajdos, T., and F. Kandil. 2008. "The Ignorant Observer." *Social Choice and Welfare* 31: 193-232.

Gilboa, I. 2009. *Theory of Decision under Uncertainty*. Cambridge: Cambridge University Press.

Grant, S., A. Kajii, B. Polak, and Z. Safra. 2010a. "Ex Post Egalitarianism and Harsanyi’s Impartial Observer Theorem.” Working paper, June 2010.
http://www.asb.unsw.edu.au/schools/economics/Documents/S.%20Grant%20-%20Ex%20Post%20Egalitarianism%20and%20Harsanyi%27s%20Impartial%20Observer%20Theorem.pdf.

———. 2010b. "Generalized Utilitarianism and Harsanyi’s Impartial Observer Theorem." *Econometrica* 78: 1939-1971.

Griffin, J. 1991. "Against the Taste Model." In J. Elster and J.E. Roemer, eds., *Interpersonal Comparisons of Well-Being*, pp. 45-69. Cambridge: Cambridge University Press.

Kagan, S. 1992. "The Limits of Well-Being." *Social Philosophy & Policy* 9: 169-189.

Kaneko, M. 1984. "On Interpersonal Utility Comparisons." *Social Choice and Welfare* 1: 165-175.

Karni, E. 1998. "Impartiality: Definition and Representation." *Econometrica* 66: 1405-1415.

Karni, E., and J.A. Weymark. 1998. "An Informationally Parsimonious Impartial Observer Theorem." *Social Choice and Welfare* 15: 321-332.

Kolm, S. 1996. *Modern Theories of Justice*. Cambridge, MA: MIT Press.

Kreps, D.M. 1988. *Notes on the Theory of Choice*. Boulder: Westview Press

MacKay, A.F. 1986. "Extended Sympathy and Interpersonal Utility Comparisons." *The Journal of Philosophy* 83: 305-322.

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 S. Barberà, P.J. Hammond, and C. Seidl, eds., *Handbook of Utility Theory*, vol. 1 (*Principles*), pp. 371-481. Dordrecht: Kluwer Academic.

Moreno-Ternero, J.D., and J.E. Roemer. 2008. "The Veil of Ignorance Violates Priority." *Economics and Philosophy* 24: 233-257.

Ng, Y. 1999. "Utility, Informed Preference, or Happiness: Following Harsanyi's Argument to Its Logical Conclusion." *Social Choice and Welfare* 16: 197-216.

Overvold, M.C. 1980. "Self-Interest and the Concept of Self-Sacrifice." *Canadian Journal of Philosophy* 10: 105-118.

———. 1982. "Self-Interest and Getting What You Want." In H.B. Miller and W.H. Williams, eds., *The Limits of Utilitarianism*, pp. 186-194. Minneapolis: University of Minnesota Press.

———. 1984. "Morality, Self-Interest, and Reasons for Being Moral." *Philosophy and Phenomenological Research* 44: 493-507.

Parfit, D. 1987. *Reasons and Persons*. Revised and corrected paperback edition. Oxford: Clarendon Press. First published in 1984.

Pattanaik, P.K. 1968. "Risk, Impersonality, and the Social Welfare Function." *Journal of Political Economy* 76: 1152-1169.

Risse, M. 2002. "Harsanyi's 'Utilitarian Theorem' and Utilitarianism." *Noûs* 36: 550-577

Roemer, J.E. 2008. "Harsanyi's Impartial Observer Is *Not* a Utilitarian." In M. Fleurbaey, M. Salles, and J.A. Weymark, eds., *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pp. 129-135. Cambridge: Cambridge University Press.

Safra, Z., and E. Weissengrin. 2003. "Harsanyi's Impartial Observer Theorem with a Restricted Domain." *Social Choice and Welfare* 20: 177-187.

Scanlon, T.M. 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.

Sen, A. 1976. "Welfare Inequalities and Rawlsian Axiomatics." *Theory and Decision* 7: 243-262.

———. 1986. "Social Choice Theory." In K.J. Arrow and M.D. Intriligator, eds., *Handbook of Mathematical Economics*, vol. 3, pp. 1073-1181. Amsterdam: North-Holland.

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 K.J. Arrow, A. Sen, and K. Suzumura, eds., *Social Choice Re-Examined: Proceedings of the IEA Conference held at Schloss Hernstein, Berndorf, Vienna, Austria*, vol. 2, pp. 202-229. Houndmills: Macmillan Press.

Vendler, Z. 1976. "A Note to the Paralogisms." In G. Ryle, ed., *Contemporary Aspects of Philosophy*, pp. 111-121. Stocksfield: Oriel Press.

Weymark, J.A. 1991. "A Reconsideration of the Harsanyi-Sen Debate on Utilitarianism." In J. Elster and J.E. Roemer, eds., *Interpersonal Comparisons of Well-Being*, pp. 255-320. Cambridge: Cambridge University Press.

———. 2005. "Measurement Theory and the Foundations of Utilitarianism." *Social Choice and Welfare* 25: 527-555.