# A reconsideration of the Harsanyi-Sen-Weymark debate on utilitarianism

HILARY GREAVES

*University of Oxford*

Harsanyi claimed that his Aggregation and Impartial Observer Theorems provide a justification for utilitarianism. This claim has been strongly resisted, notably by Sen and Weymark, who argue that while Harsanyi has perhaps shown that overall good is a linear sum of individuals' von Neumann-Morgenstern utilities, he has done nothing to establish any connection between the notion of von Neumann-Morgenstern utility and that of well-being, and hence that utilitarianism does not follow.

The present article defends Harsanyi against the Sen-Weymark critique. I argue that, far from being a term with precise and independent quantitative content whose relationship to von Neumann-Morgenstern utility is then a substantive question, terms such as 'well-being' suffer (or suffered) from indeterminacy regarding precisely which quantity they refer to. If so, then (on the issue that this article focuses on) Harsanyi has gone as far towards defending 'utilitarianism in the original sense' as could coherently be asked.

## I. INTRODUCTION

Textbook utilitarianism comprises three components: a particular account of individual well-being (hedonism), a particular account of the relationship between individual well-being and the overall goodness of the state of affairs (the additive method of aggregation), and a particular account of the relationship between goodness of states of affairs and what one ought to do (maximising consequentialism). Notoriously, the classical utilitarians had very little to say by way of justification for the second of these components: even granted the assumption ('welfarism') that overall good is a function of individual well-being, why suppose that the function in question must be straightforward summation, rather than (say), as prioritarians would have it, the sum of a concave transform of individual well-being levels? Thus Bentham enjoins us to

> *Sum up* the numbers expressive of the degrees of *good* tendency, which the act has, with respect to each individual, in regard to

> whom the tendency of it is *good* upon the whole: do this again with respect to each individual, in regard to whom the tendency of it is *bad* upon the whole. Take the *balance*.[1]

- but the reader will search in vain for any attempt to *justify* summing in particular. Mill, meanwhile, is not even explicit that he is *making* the summative claim, saying little more than that

> [E]ach person's happiness is a good to that person, and the general happiness, therefore, a good to the aggregate of all persons.[2]

From the 1950s, Harsanyi[3] presented two key theorems that, according to him, supply the missing justification for the utilitarian's summation: the *Aggregation Theorem* and the *Impartial Observer Theorem*. Samuelson[4] concurs, referring to Harsanyi's results as 'the resurrection for ethics of additive hedonism'. Quite aside from any concerns over the theorems themselves or the plausibility of their axioms, however, the ensuing discussion contains significant resistance to Harsanyi's claim that the conclusions of the theorems in question really amount to utilitarianism. The bone of contention is whether, granted (for the sake of argument) that Harsanyi's theorems establish that overall good is a sum of *von Neumann-Morgenstern (VNM) utilities*, we are warranted in concluding therefrom that overall good is the sum of *well-being* - might not VNM utility and well-being come apart?

---

[1] Bentham, J. (1879). *An introduction to the principles of morals and legislation.* Oxford: Clarendon Press, ch. 4, sec. 5.6; emphasis in original.

[2] Mill, J. S. (1962). Utilitarianism. In M. Warnock (Ed.), *Utilitarianism; On Liberty; Essay on Bentham.* London: Fontana, ch. IV.

[3] Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 61(5), 434-5; Harsanyi, J. (1955, August). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4), 309–321; Harsanyi, J. C. (1977b). *Rational behavior and bargaining equilibrium in games and social situations.* Cambridge University Press.

[4] Samuelson, P. A. (1981). Bergsonian welfare economics. In S. Rosefielde (Ed.), *Economic welfare and the economics of soviet socialism: Essays in honor of Abram Bergson* (pp. 223–66). Cambridge University Press. (Reprinted in Crowley (ed.), *The Collected Scientific Papers of Paul A. Samuelson*, Volume V, MIT Press (1986)), at p. 245.

This line of criticism has been pressed by, in particular, Sen[5] and Weymark.[6] The present article is concerned with one specific aspect of the criticism, concerning the means by which one progresses from (i) a mere *ordering* of outcomes in terms of how good they are for some particular individual to (ii) a quantitative - specifically, a *cardinal* - measure of the goodness of outcomes for the individual in question. The machinery of decision theory, and the associated notion of VNM utility, provides one means of making the transition from (i) to (ii). The core of the Sen-Weymark critique is a suggestion that the quantitative notion of well-being carries with it an independent way of making such a transition, and that, absent some reason (which has not been provided) to think that the two ways of adding cardinal structure lead to the same result, there is no reason to regard Harsanyi's conclusion as equivalent to *utilitarianism*. Of course (the criticism might continue) nothing prevents Harsanyi from *stipulating* that by 'well-being' he henceforth just means von Neumann-Morgenstern utility, and thereby hijacking the word 'utilitarianism' to describe the conclusions of his theorems, but this would merely be changing the subject: it would remain the case, and we should all be clear, that Harsanyi has not defended utilitarianism *in the original sense of that term*. Sen and Weymark therefore endorse Roemer's[7] conclusion that '[Harsanyi's] error lies in confusing a mathematical sum of VNM utilities with the substantive view of utilitarianism'.

This appears to be the dominant view of the interpretation of Harsanyi's theorems in the current literature. The purpose of the present article is to argue for an alternative account of the relationship between Harsanyi's theorems and the claims of the original utilitarians, and one that is more sympathetic to Harsanyi. According to this alternative account, terms such as 'well-being' (or 'welfare', or 'utility', or 'happiness') in the mouths of utilitarians prior to the advent of decision theory were subtly *indeterminate* over precisely which *quantitative* measure they referred to: they were indeterminate, that is, between various possible methods of (intrapersonally) 'cardinalizing' the ordinal

[5] Sen, A. (1976). Welfare inequalities and Rawlsian axiomatics. *Theory and decision*, 7, 243–62; Sen, A. (1977). Non-linear social welfare functions: A reply to Professor Harsanyi. In R. E. Butts & J. Hintikka (Eds.), *Foundational problems in the special sciences* (Vol. 2, 297–302). Springer.

[6] Weymark, J. A. (1991). A reconsideration of the Harsanyi–Sen debate on utilitarianism. In J. Elster & J. E. Roemer (Eds.), *Interpersonal comparisons of well-being* (pp. 255–320). Cambridge University Press, Cambridge; Weymark, J. A. (2005). Measurement theory and the foundations of utilitarianism. *Social Choice and Welfare*, 25(2-3), 527–555.

[7] Roemer, J. E. (2008). Harsanyi's impartial observer is not a utilitarian. In M. Fleurbaey, M. Salles, & J. Weymark (Eds.), *Justice, political liberalism, and utilitarianism: Themes from Harsanyi and Rawls.* Cambridge University Press (pp. 129-135).

well-being scale. It follows that there is no single 'substantive' view of utilitarianism to defend: rather, the content of utilitarianism itself is (at least prior to decision theory) somewhat indeterminate. In that case, Harsanyi's reaction to the situation he faced was entirely the appropriate one: first to suggest a means (in his case, VNM-based) of resolving the indeterminacy, and then to investigate whether or not given that means of *rendering* the thesis of utilitarianism a definite and substantive one, it turns out to be a true thesis. In particular, and *contra* Roemer et al., there is no independent (precise, quantitative, determinate) notion of well-being with which Harsanyi has 'confused' VNM utility; nor (crucially) is there any criticism in the vicinity that is more a matter of 'substance' and less one of semantics.

The article proceeds as follows. Section II outlines Harsanyi's Aggregation and Impartial Observer theorems, and the means by which Harsanyi proposes to conclude in favour of utilitarianism (the details of the more complicated Impartial Observer theorem are relegated to the Appendix). Section III distinguishes between the various ways in which one might dissent from Harsanyi's arguments; in particular, it separates issues related to the vexed question of *inter*personal utility comparisons from the *intra*personal issues of scale-fixing that will be my central concern. Section IV makes a preliminary attempt to diagnose the source of the disagreement between Harsanyi and his critics in terms of a rival pair of theses ('operationalism' and 'primitivism') regarding the means by which a notion such as that of well-being might or might not be equipped with determinate content: operationalism favours Harsanyi's position, primitivism that of his critics. Since, however, neither operationalism nor primitivism is ultimately a tenable account of the conditions under which a notion has genuine content, this observation will not settle the debate.

Section V explores in more detail the notion of *semantic indeterminacy* that is central to my positive account of the Harsanyi-Sen-Weymark debate. Section V.1 surveys some relatively uncontroversial examples of indeterminacy that serve to illustrate the general phenomenon. Section V.2, against the background of those examples, sets up some terminology for theorizing about such situations. Using this terminology, Section V.3 then lists the salient possibilities for the content (or lack of content) of 'well-being' talk - that is, various possible *semantic hypotheses* regarding the content of such talk. The significance of each of those possibilities for the Harsanyi-Sen-Weymark debate, *if* the possibility in question turned out (as a matter of correct semantics) to obtain, is discussed in Section V.4; in particular, I explicate the sense in which the indeterminacy thesis, if true, would vindicate Harsanyi. By this point the question of what the available alternatives (anyway) are to the VNM way of 'cardinalizing' the well-being scale will have become urgent; Section V.5 surveys the possibilities.

I do not, thus far, attempt to *resolve* the question of which of the hypotheses outlined in Section V.3 is correct. The important point for our purposes in any case concerns the *status of*, rather than the answer to, that question: namely, *it is a question of semantics*. That is: the question is not one of which properties some antecedently well-grasped quantity out there in the world has, but rather one of which (out-there-in-the-world) quantity or quantities, given the correct

account of how words get their meanings, we should take *words* such as 'well-being' (or 'welfare', or 'happiness', etc.), when used in the relatively theoretically undeveloped context of pre-1940s utilitarianism, to *refer to*.

Many readers are likely to be uninterested in such detailed issues of ('mere') semantics. I am sympathetic to this attitude; indeed, one of the central points of the present article is that the Sen-Weymark critique of Harsanyi, far from being the 'substantive' matter that its proponents take it to be, is a semantic storm in a teacup. I invite these readers to skip sections VI–VII; for them, the interesting question is only whether the dispute under discussion *is* (as I have suggested) merely one of semantics, rather than how to resolve it on condition that it is. Since even semantic questions have answers, however, the next two sections offer a preliminary investigation into which of the semantic hypotheses of section V.3 seems most likely to be correct. Section VI, building on section IV's observation that both operationalism and primitivism are inadequate, is a rough sketch of one common more mature approach to theorizing about the determination of word-referents. Section VII applies this general approach to the particular case of 'well-being': my tentative conclusion will be that, as a matter of semantic fact, the term 'well-being' was indeed (at least pre-1940s) somewhat indeterminate in reference. If so, then (as outlined in section V) Harsanyi has not committed even a semantic sin. Section VIII summarizes what Harsanyi should have said in response to the Sen-Weymark critique. Section IX is the conclusion.

# II. HARSANYI'S THEOREMS AND THEIR INTERPRETATION

We begin, then, by surveying Harsanyi's Aggregation Theorem and Impartial Observer Theorem. Both theorems crucially involve the evaluation of situations involving risk. Let I be a finite population, assumed fixed (that is, problems of *population* axiology are beyond the scope of the present discussion). Let X be a finite set of *outcomes*: states of affairs that are specific enough to pin down everything that matters to the well-being of any individual in I. A *lottery* over X is a probability distribution on X; let L(X) be the set of all such lotteries. Given a total ordering $\succeq$ of such a set $L$ of lotteries, say that a function $h : L \to \mathbb{R}$ *represents* $\succeq$ iff for all lotteries $p, q \in L$, $p \succeq q \Leftrightarrow h(p) \geq h(q)$; say that a function $k : X \to \mathbb{R}$ *expectationally represents* $\succeq$ iff for all lotteries $p, q \in L$, $p \succeq q \Leftrightarrow \sum_{x \in X} p(x)k(x) \geq \sum_{x \in X} q(x)k(x)$.

For the Aggregation Theorem, we consider a number of orderings of L(X): an 'overall' ordering $\succeq$, and, for each individual $i \in I$, an 'individual' ordering $\succeq_i$ for that individual. The intended interpretation is that $\succeq$ ranks lotteries in terms of (ex ante) better and worse overall, while $\succeq_i$ ranks lotteries in terms of (ex ante) better and worse *for the particular individual* $i$. Suppose that the structure $(\succeq, \{\succeq_i : i \in I\})$ obeys the following three conditions:

*AT1:* $\succeq$ obeys the axioms of expected utility theory.

*AT2:* For every $i \in I$, $\succeq_i$ obeys the axioms of expected utility theory.

*AT3 (Strong Ex Ante Pareto):* If $A, B$ are lotteries such that for every $i \in I$, $A \succeq_i B$, then $A \succeq B$; if, further, there exists $i \in I$ such that $A \succ_i B$, then $A \succ B$.

Then, the theorem establishes, the 'overall' ordering $\succeq$ of L(X) can be represented by an expression of the form

$$EU(p) = \sum_{i \in I} \sum_{x \in X} p(x) u_i^{VNM}(x), \tag{1}$$

where the individual von Neumann-Morgenstern utilities $u_i^{VNM} : X \to \Re$ expectationally represent the respective 'individual' orderings $\succeq_i$. ('Can be represented': provided that we select the representative VNM utility functions $u_i^{VNM}$appropriately. Of course, each individual's VNM utility function is defined only up to positive affine transformation - if $u$ is an adequate utility function for a given individual then so also is $au + b$, for any $a \in \Re^+, b \in Re$ - and if the expression (1) correctly represents a given ordering of L(X) relative to one choice of family of representative utility functions $(u_i^{VNM})$, in general it will not correctly represent L(X) relative to an arbitrary *different* family of representative utility functions. 'Expectationally represent': for every $i \in I$, the ordering $\succeq_i$ of $L(X)$ is ordinally represented by the expectation-value formula $\sum_{x \in X} p(x) u_i^{VNM}(x)$.) *Given* a further assumption that the individual von Neumann-Morgenstern utility scales in question coincide with the well-being scales for the corresponding individuals, this conclusion implies the utilitarian theory of the overall good.

(We note in passing that Harsanyi himself, and most of his commentators, interpret the orderings as (respectively) 'social' and individual *preference* orderings, rather than directly as betterness orderings. This is equivalent to the interpretation suggested above *on the assumption of a preference-satisfaction theory of betterness* (and is of course consistent with that assumption). The extra assumption, however, plays no central role in the argument, and for our purposes an insistence on translating all evaluative claims into preference-talk is a distraction: if the point is to defend utilitarianism, what we fundamentally seek is a representation of betterness. Readers who happen to be fans of a preference-satisfaction theory of well-being are free to effect such a translation for their own purposes.)

For the Impartial Observer argument, the key idea is that of an individual forming preferences about the state of the world while in a state of ignorance regarding which individual in society he is to be: specifically, in a state of facing an equal probability of being any given member of the society. Harsanyi,[8] argues that the 'moral point of view' is an impartial one giving equal and positive

[8] Harsanyi, 'Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking'; Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, 48-50.

weight to the interests of each person, and that the preferences corresponding to such a point of view coincide with those of a rational agent behind such a Veil of Ignorance. In that case, to determine moral matters, we should enquire into the preferences that an 'observer' would have in such a position of ignorance. The Impartial Observer argument similarly purports to establish, in the first instance, that the 'observer"s preferences would be representable by a sum of individuals' VNM utilities. (The details of this part of the argument are in the Appendix.) As a special case, the observer's preferences over *outcomes* are represented by a sum-over-individuals of individuals' von Neumann-Morgenstern utilities for those outcomes. But, again assuming that those individual von Neumann-Morgenstern utility scales are the individuals' well-being scales, this special case amounts to the statement that the 'observer' prefers one outcome to another when and only when the utilitarian deems the first outcome overall-better than the second. Given Harsanyi's further claim that the preferences of this 'observer' match the true overall-betterness relation, this again implies the utilitarian theory of overall good. (Later, Rawls would notoriously claim that a soul behind an appropriately specified Veil of Ignorance would prefer, not the outcome favoured by the utilitarian formula, but rather that recommended by a maximin formula.[9] In this context, Harsanyi's 'Impartial Observer' argument can be seen as taking on the same question as Rawls's appeal to the veil of ignorance, but proceeding on the assumption of standard decision theory instead of (Rawls's choice) the highly non-standard 'maximin' theory.)

# III. VOICES OF DISSENT

## III.1. *Three types of objection to Harsanyi's claims*

*Has* Harsanyi, then, supplied the missing defence of utilitarianism's summative formula for overall good? Voices of dissent can be grouped into three main categories.

In the first instance, the theorems themselves that form the core of Harsanyi's arguments are the subject of technical concern: how (precisely) the theorems are best formulated and which (originally implicit) supplementary assumptions they require,[10] and the availability or otherwise of similar results in alternative decision-theoretic frameworks.[11]

[9] Rawls, J. (1972). *A theory of justice*. Oxford University Press.

[10] E.g. Weymark, 'A Reconsideration of the Harsanyi-Sen Debate on Utilitarianism'; P. Mongin, 'Impartiality, Utilitarian Ethics, and Collective Bayesianism' (Ely Lectures delivered at John Hopkins University, 2002).

[11] E.g. Mongin, P. (1995). Consistent Bayesian aggregation. *Journal of Economic Theory*, 66(2), 313–351; Broome, J. (1990). Bolker-jeffrey expected utility theory and axiomatic utilitarianism. *The Review of Economic Studies*, 57(3), 477–502; see also Mongin 'Impartiality,

Second, while the evaluative assumptions that correspond to the axioms of Harsanyi's theorems at first blush seem eminently reasonable, on closer inspection one might have doubts. For example, while there is of course much discussion of the extent to which the axioms of one or another formulation of expected utility theory are requirements of rationality in general, it has particularly been questioned whether they are legitimate requirements to impose on an ordering of lotteries that is supposed to represent *overall* or *impartial* ex ante betterness.[12] One can similarly question the *ex ante* (as opposed to *ex post*) versions of Pareto principles: for instance, there is principled reason to reject them if one believes that interpersonal *equality* of well-being is intrinsically valuable;[13] Fleurbaey and Voorhoeve[14] point out that if a Pigou-Dalton condition is also assumed, then an ex ante Pareto principle conflicts with their 'Principle of minimum information'; Mongin and d'Aspremont[15] argue that ex ante Pareto principles are inappropriate on the ground that while individuals are sovereign regarding matters of taste, their empirical beliefs do not deserve such deference. A closely related line of thought might lead one to reject the identification of overall good with the preferences of a rational, self-interested but ignorant 'soul' that is required for the 'Impartial Observer' argument.[16]

The third line of dissent arises from the identification of individuals' *von Neumann-Morgenstern utility* levels with their *well-being* levels. This identification, we have seen, is essential if the conclusion of Harsanyi's arguments is to coincide with utilitarianism: the utilitarians' claim is that overall good corresponds to aggregate *well-being*, not to aggregate some-other-quantity-we-know-not-what. The issue is whether this identification can be defended.

---

Utilitarian Ethics, and Collective Bayesianism' and references therein.

[12] Diamond, P. A. (1967, October). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: Comment. *Journal of Political Economy*, 75(5), 765–6; Sen, 'Welfare Inequalities and Rawlsian Axiomatics'.

[13] E.g. Adler, M., & Sanchirico, C. (2006). Inequality and uncertainty: Theory and legal applications. *University of Pennsylvania Law Review*, 155, 279–377, at p. 323.

[14] Fleurbaey, M., & Voorhoeve, A. (2013). Decide as you would with full information! An argument against ex ante Pareto. In N. Eyal, S. Hurst, O. Norheim, & D. Wikler (Eds.), *Inequalities in health: Concepts, measures, and ethics* (pp. 113–128). Oxford University Press.

[15] Mongin, P., & d'Aspremont, C. (1998). Utility theory and ethics. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory, volume 1: Principles* (pp. 371–481). Kluwer Dordrecht, at p. 442.

[16] Roemer, J. E. (2002). Egalitarianism against the veil of ignorance. *The Journal of philosophy*, 167–184.

This third line of criticism actually bundles together two independent issues: one concerning the distinction between ordinal and cardinal representations of individual well-being, the other concerning interpersonal utility comparisons and the (related) distinction between weighted and unweighted utilitarianism. My concern in the present article will be exclusively with the first of these two issues. Let us begin, however, by distinguishing the two, and laying the second to rest; these are the tasks of the remainder of section III.

## III.2. *Some rudiments of measurement theory*

To clarify our discussion of well-being, von Neumann-Morgenstern utility and the relationship between them, some measurement-theoretic terminology is required.

Well-being and von Neumann-Morgenstern utility are both *quantities*. We begin by defining various quantity *types*, according to which aspects of internal structure the quantity in question does and does not possess. Our interest will be in quantities that measure some aspect of the state of each individual i in our population. Any such quantity consists in more or less structure defined on the outcome set $X$. An *individually ordinal* ('i-ordinal') quantity is given by $X$ together with, for each $i$, a total ordering $\succeq_i$. An *individually cardinal* ('i-cardinal') quantity is (further) equipped, for each $i$, with a quaternary function $C_i : X^4 \to \Re$ that obeys axioms corresponding to 'ratios of differences' (heuristically: supplying a cardinal scale for each $i$ *taken separately*).[17] A *co-cardinal* quantity has all the structure of an i-cardinal quantity, but in addition is equipped with a (single) equivalence relation $\sim$: for each $i, j \in I$, and all elements $a, b, e, f \in X$, there is a fact about whether or not $(a, b; i) \sim (e, f; j)$.[18] (Heuristically: we have $(a, b; i) \sim (e, f; j)$ iff the difference between $b$ and $a$ for

[17] *If* (but only if) the set of outcomes is sufficiently rich, one can instead take the i-cardinal structure to be given by an ordering of pairs ('the difference between $x_1$ and $x_2$ is at least as great as the difference between $x_3$ and $x_4$'), and still end up with numerical scales that are unique up to positive affine transformation; see the discussion of 'intrapersonal difference comparability' in Bossert, W., & Weymark, J. A. (2004). Utility in social choice. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory, Volume 2: Extensions* (pp. 1099–1177). Springer, at pp. 1127-1128, and references therein.

[18] We also need to impose requirements of mutual consistency between the i-ordinal and i-cardinal (resp., i-cardinal and co-cardinal) structures for a given quantity. Consistency between i-ordinal and i-cardinal structure: if $a \simeq_i b, c \simeq_i d, e \simeq_i f, g \simeq_i h \in X$ and $C_i(a, c, e, g) = r \in \Re$, then $C_i(b, d, f, h) = r$ also. Consistency between i-cardinal and co-cardinal structure: if $C_i(a, b, c, d) = C_j(e, f, g, h)$, and if in addition $(a, b; i) \sim (e, f; j)$, then

$i$ is the same size as the difference between $f$ and $e$ for $j$; that is, $\sim$ encodes a standard of interpersonal unit comparisons.)

Given any two such quantities for the same outcome set $X$ and population $I$, one can enquire whether or not the quantities themselves are *equivalent* to one another, in any of various senses of equivalence. Say that two quantities are *i-ordinally equivalent* (respectively: *i-cardinally equivalent, cocardinally equivalent*) if their i-ordinal (respectively: i-cardinal, cocardinal) structures are identical. The notion of i-cardinal (resp. cocardinal) equivalence is applicable, of course, only if both of the quantities in question possess at least i-cardinal (resp. cocardinal) structure. (We leave open the question, unimportant for our purposes, of whether or not there can be two quantities that are based on the same $X, I$, are of the same type (i-ordinal, i-cardinal etc.), and share all applicable structures (e.g. that have identical i-ordinal and i-cardinal structures, in the case of two i-cardinal quantities), but that are nevertheless distinct quantities - the question, that is, of whether or not full equivalence entails identity for quantities.)

It is sometimes fruitful to work, not directly with these quantities themselves, but rather with real-valued representations thereof: families of functions $\{(f_i : X_i \to \Re) : i \in I\}$. An i-ordinal quantity is *i-ordinally represented* by such a family $\{f_i\}$ iff, for each $i \in I$ and each $a, b \in X$, $f_i(a) \geq f_i(b)$ iff $a \succeq_i b$. An i-cardinal quantity is i-ordinally represented by such a family under the same condition; it is *i-cardinally represented* by $\{f_i\}$ iff, further, for every $i \in I$ and every $a, b, c, d \in X$, $\frac{f_i(a)-f_i(b)}{f_i(c)-f_i(d)} = r \Leftrightarrow C_i(a, b, c, d) = r$. A cocardinal quantity is i-ordinally or i-cardinally represented by $\{f_i\}$ again under the same conditions; it is *cocardinally represented* by $\{f_i\}$ iff, further, for every $i, j \in I$ and every $a, b, c, d \in X, f_i(a) - f_i(b) = f_j(c) - f_j(d)$ iff $(a, b; i) \sim (c, d; j)$. (We have, of course, no notion of cocardinal (respectively i-cardinal) representation for a quantity that is itself a merely i-ordinal or i-cardinal (resp., a merely i-ordinal) quantity.)

These different types of representation correspond to different equivalence relations among the class of real-valued representatives. If $\{f_i\}$ i-ordinally represents a given (i-ordinal, i-cardinal or cocardinal) quantity, then so also does any function-family related to $\{f_i\}$ by an *i-ordinal transformation*: that, is, a transformation $f_i \mapsto t_i \circ f_i$, where, for each $i$, $t_i : \Re \to \Re$ is a strictly increasing transformation. If $\{f_i\}$ i-cardinally represents a given quantity, then so also does any function-family related to $f_i$ by an *i-cardinal transformation*, i.e. a transformation of the form $f_i \mapsto t_i \circ f_i$,where, for each $i$, $t_i : \Re \to \Re$ is a positive affine transformation. If $\{f_i\}$ cocardinally represents a given quantity, then so also does any function-family related to $f_i$ by a *cocardinal transformation*, i.e. a transformation of the form $f_i \mapsto t \circ f_i$,where $t : \Re \to \Re$ is a (single) positive affine transformation.

---

$(c, d; i) \sim (g, h; j)$.

## III.3. *Objection 1: Harsanyi has done nothing to establish that well-being is i-cardinally equivalent to von Neumann-Morgenstern utility*

The central Sen-Weymark worry concerns the distinction between i-cardinal and merely i-ordinal agreement between well-being on the one hand, and von Neumann-Morgenstern utility on the other (or, equivalently, between i-ordinal and i-cardinal representations of well-being). The notion of an individual's *well-being* is explicitly evaluative, in that it is supposed to be a measure of how good things are - how well things are going - for the individual in question. Suppose then, to start with, that there are facts, for any given individual, about which states of affairs are better and worse *for that individual* than which others. The well-being of any given individual is, in that case, at least an i-ordinal quantity. Suppose further that individual well-being is i-cardinal, i.e. that there are facts, for states of affairs A, B, C, D, as to whether or not the *difference* between how good A is and how good B is for a given individual is (say) *twice* the *difference* between how good C is and how good D is for that individual. Suppose now that we try to represent the well-being scale, for a given individual, by means of a particular assignment $f$ of numbers to states of affairs; as above (section III.2), we might work with merely i-ordinal representations, or with representations that are, in addition, i-cardinal.

What of von Neumann-Morgenstern utilities? For our present purposes, the signal achievement of decision theory is to show how a cardinal notion (viz. the von Neumann-Morgenstern measure of the relative goodnesses of *outcomes*) can be derived from an ordinal one, provided that the ordering from which one starts is an ordering not only of some set of outcomes, but also of lotteries over that outcome set (i.e. assignments of probabilities to outcomes). Specifically, the representation theorems of decision theory establish that if an ordering $\succeq_i$ of lotteries over an outcome set $X$ obeys a set of constraints (the axioms of expected utility theory) that are arguably reasonable under the intended interpretation of $\succeq_i$, then that ordering can be represented 'expectationally' by a utility function $u_i : X \to \Re$, i.e. for any lotteries $a, b$, we have $a \succeq_i b$ iff $\sum_{x \in X} p_a(x) \cdot u_i(x) \geq \sum_{x \in X} p_b(x) \cdot u_i(x)$, where $p_a(x)$ (resp. $p_b(x)$) is the probability of outcome $x$ under lottery $a$ (resp. $b$) - and, crucially for our purposes, that (given the requirement of *expectational* representation) this function $u_i$ is unique up to positive affine transformation. This last means that von Neumann-Morgenstern utility is a 'cardinal' quantity, since affine transformations preserve ratios of differences. If, as is normal,[19] this decision-theoretic machinery is applied to rankings of outcomes and lotteries in terms of how good they are for

[19] But not inevitable: cf. the 'extended preferences' approach to grounding interpersonal comparisons, discussed in e.g. Harsanyi, *Rational Behavior*, secs. 4.2-4.4; Broome, J. (1998). Extended preferences. In C. Fehiga & U. Wessels (Eds.), *Preferences* (pp. 271–287); Adler, M. (2012). *Well-being and fair distribution: Beyond cost-benefit analysis.* Oxford University

each individual *treated separately* (when, that is, the input to the machine is a *separate* ranking $\succeq_i$ of outcomes/lotteries for each individual $i$), the result is a notion of (individual) von Neumann-Morgenstern utility that is i-cardinal, but of course not cocardinal. (When it is necessary to emphasize this latter point (when required), we will refer to the quantity in question as *individual* von Neumann-Morgenstern utility (iVNM utility), rather than von Neumann-Morgenstern (VNM) utility *simpliciter*.)

The question now is the relationship between von Neumann-Morgenstern utility on the one hand, and well-being on the other. Given our policy of interpreting the orderings $\succeq_i$ directly as *betterness*-for-the-individual orderings rather than necessarily the individuals' *preference* orderings, and since expected von Neumann-Morgenstern utility reduces to von Neumann-Morgenstern utility itself in the case of riskless lotteries, it is automatic that VNM utilities, if they exist at all (that is, if the individual-betterness ordering of prospects satisfies the axioms of expected utility theory), are *ordinally* equivalent to individual well-being. But - Sen and Weymark's basic point - nothing in the representation theorems of decision theory, or elsewhere in Harsanyi's work, guarantees that they are also *i-cardinally* equivalent to well-being, i.e. that ratios of well-being differences and ratios of VNM utility differences are identical.

The point can be made vivid by consideration of the variety of possible numerical representations that equally well ordinally represent a given ordering $\succeq_i$. What the standard results of expected utility theory establish is that, provided an ordering $\succeq_i$ of lotteries obeys the theory's axioms, then $\succeq_i$ is ordinally represented by a formula of the form

$$\sum_{x \in X} p(x) \cdot u_i(x) \tag{2}$$

in which the function $u_i : X \to \Re$ ordinally represents $\succeq_i$ on riskless lotteries (i.e. lotteries in which the same outcome is obtained in every state of nature), and is unique up to positive affine transformation. Trivially, though, if an ordering of outcomes (such as $\geq_i |_X$) is ordinally represented by such a function $u_i$, then for any strictly increasing transformation $f : \Re \to \Re$, the same ordering is also ordinally represented by $f(u_i)$; and, if f is non-affine, it cannot generally be that both $u_i$ and $f(u_i)$ i-cardinally represent the same underlying structure. To be sure, unlike $u_i$ itself, $f(u_i)$ does not extend to any expectational representation of $\geq_i$ on $L(X)$; but (the objection urges) nothing justifies insisting that an i-cardinal representation of well-being must have that particular property. The Sen-Weymark objection is then that since nothing has therefore been said that might justify treating $u_i^{VNM}$ rather than some increasing transform thereof as (i-cardinally equivalent to) well-being, Harsanyi's claim that the Aggregation and/or Impartial Observer Theorems justify the utilitarian method of aggregation of *well-being* has not been justified.

---

Press, ch. 3; Greaves, H., & Lederman, H. (n.d.). *Extended preferences.* (Manuscript).

## III.4. *Objection 2: Harsanyi has established at most weighted utilitarianism, not utilitarianism* simpliciter

The second aspect of the 'who says that VNM utility is well-being?' criticism concerns the fact that even if the well-being scale for a given individual and the iVNM utility scale for that same individual are i-cardinally equivalent, it does not follow that the sum of well-being across a population is identical to the sum of iVNM utility across that same population. In fact, the latter would-be sum is not even well-defined, because, as we noted above, iVNM utility, being derived from entirely *separate* rankings of lotteries in terms of goodness for each individual in turn (via individual preference orderings or otherwise), is not a quantity that is equipped with a standard of interpersonal comparison: it is a merely i-cardinal, not a cocardinal, quantity. If well-being is an interpersonally comparable quantity, so that utilitarianism is a coherent thesis, the quantity *well-being* cannot be *identical* to the quantity *iVNM utility* (at most, in the terminology of section III.2, the two can be 'i-cardinally equivalent'). Harsanyi therefore needs to be careful over the statement of the assumption that he needs *in the vicinity of* 'VNM utility is well-being', and relatedly over the description of his conclusion as 'utilitarian'.

Once the dust over this matter settles, the following picture emerges. Suppose, for the sake of argument, that well-being, unlike individual VNM utility, is a co-cardinal quantity. In that case, both of the following theses are both coherent and non-trivial: the utilitarian thesis that overall good is represented by the sum-over-individuals of well-being ($\sum_i w_i$), and the variant *weighted* utilitarian thesis that overall good is represented by a *weighted* sum-over-individuals of well-being ($\sum_i a_i w_i$, for some coefficients $a_i \in \Re^+$). Given the further (Harsanyian) assumption discussed in section III.3 above, viz. that well-being and individual VNM utility are i-cardinally equivalent, the conclusion of Harsanyi's theorems is *weighted* utilitarianism, rather than (as Harsanyi himself seemed to claim) utilitarianism *simpliciter* - notwithstanding Harsanyi's having shown (equation (1) above) that overall good *can* be represented by an unweighted sum of VNM utilities. (On the alternative assumption that well-being itself is a merely i-cardinal (rather than cocardinal) quantity, the situation is subtly different, but not in ways that are ultimately very important for our purposes. If well-being is merely i-cardinal then, in the first instance, utilitarianism *simpliciter* is not coherent. There remains, on the other hand, a coherent thesis deserving of the name 'weighted utilitarianism', but that thesis must in this case be stated slightly differently: it is the thesis that overall good is represented by some function of individuals' well-being levels that is positive affine w.r.t. each individual's well-being. Harsanyi's theorems then establish weighted utilitarianism in this sense.)

This criticism is well taken.[20] For the remainder of this article, let us accept

[20] For further discussion of weighted utilitarianism and interpersonal comparisons of utility

this conclusion, but set it aside: my concern is with Objection 1 above.

# IV. OPERATIONALISM, PRIMITIVISM AND THE NEED FOR A MIDDLE GROUND

Harsanyi has published several replies to Sen.[21] Specifically on the issue of the i-cardinal equivalence of VNM utility and well-being, however, Harsanyi appears unable to grasp Sen's concern. Witness, for instance, Sen's objection

> Obviously, the von Neumann-Morgenstern values - let us call them the V-values - of social welfare will be a linear combination of the V-values of individual welfares. But when someone talks about social welfare being a non-linear function of individual welfares, the reference need not necessarily be to the V-values at all. The V-values are of obvious importance for predicting individual or social choice under uncertainty, but there is no obligation to talk about V-values only whenever one is talking about individual or social welfare.[22]

and Harsanyi's (non-)reply:

> [Sen] proposes that, if both individual choices and social-policy choices did follow the Bayesain rationality axioms, then we should *act* as good utilitarians, by always choosing the social policy maximizing a given specific linear combination of all individuals' VNM utility functions. But, at the same time, we should use a terminology which amounts to carefully *disguising* the fact that we are utilitarians . . . We should refuse to *call* an individual's VNM utility function our measure for his personal welfare, even though, in the mathematical

in the context of Harsanyi's theorems, see, e.g. Mongin and d'Aspremont, 'Utility Theory and Ethics', sec. 5.2. For Harsanyi-style theorems that aim to establish unweighted utilitarianism via the imposition of an additional axiom of 'anonymity', see Mongin and d'Aspremont, 'Utility Theory and Ethics', Proposition 5.3; d'Aspremont, C., & Mongin, P. (2008). A welfarist version of Harsanyi's aggregation theorem. In *Justice, political liberalism, and utilitarianism* (p. 184- 197). Cambridge University Press, Theorem 7.2.

[21] Harsanyi, J. C. (1975). Nonlinear social welfare functions: Do welfare economists have a special exemption from Bayesian rationality? *Theory and Decision*, 6(3), 311–332; Harsanyi, J. C. (1977a). Nonlinear social welfare functions: A rejoinder to Professor Sen. In R. E. Butts & J. Hintikka (Eds.), *Foundational problems in the special sciences* (Vol. 2, pp. 293–296). Springer.

[22] Sen, 'Welfare Inequalities and Rawlsian Axiomatics', p. 248.

> expression used to evaluate alternative social policies, we would go on representing this individual's interests by his VNM utility function. No doubt, we *could*, if we wanted to, persuade in this very peculiar manner. But it is hard to see what we could gain by following Sen in this rather elaborate and cumbersome camouflage operation.[23]

What is going on in these exchanges? An interpretation uncharitable to Harsanyi would be that the latter has *simply* failed to grasp Sen's point. Clearly Harsanyi *has* failed to grasp the point; the suggestion I wish to investigate in the remainder of the article, however, is that there are principled reasons why one might so fail. Further (the suggestion continues), for related reasons of principle, Harsanyi is *correct* not to grasp it: Sen's criticism itself is, in a sense I will explicate, devoid of content.

In the first instance, the Harsanyi-Sen failure to make contact would be easily understood on the hypothesis that the respective authors subscribed to differing methodological commitments regarding what is required in order to give content to a question (here, the question of whether von Neumann-Morgenstern utility really *is well-being*): Harsanyi is perhaps an operationalist, Sen what I will call a 'primitivist'.

*Operationalists* insist that in order for a concept to have determinate content, it must be clearly associated with a measurement procedure that determines whether or not it is to apply in any particular case. Witness, for example, De Finetti:

> In order to give an effective meaning to a notion, and not merely an appearance of such in a metaphysical-verbalistic sense, an operational definition is required. By this we mean a definition based on a criterion that allows us to measure it.[24]

Precisely what passes muster according to this criterion depends, of course, on precisely what counts as a 'measurement'. The general idea is, however, clear enough for our purposes, and has obvious affinities with the idea of identifying well-being with VNM utility. In the behaviourist/operationalist atmosphere of the early twentieth century, there was deep suspicion about the ascription of anything like 'numbers in the head' to individuals; such ascriptions were regarded as legitimate in so far as, but only in so far as, they encoded some more-or-less *observable* aspect of the individual's psychology. Ordinal notions arguably pass this test, since an 'ordering-in-the-head' could, given also a preference-satisfaction theory of well-being, consist in the individual's *preference ordering* over various possibilities, while the latter in turn can be cashed out in terms of dispositions to choice behaviour. Cardinal notions, on the other hand, are at first sight much more dubious by operationalist lights: no such measurement seems available for intrapersonal comparisons that go beyond the merely ordinal. This is a context in which the representation theorems of expected utility

[23] Harsanyi, 'Nonlinear Social Welfare Functions: A Rejoinder to Professor Sen', p. 294.

[24] De Finetti, B. (1974). *Theory of probability, volume I.* London: Wiley, p. 76.

theory appear highly significant: as noted in section III, they identify *a* cardinal notion (at least) that can plausibly be measured via observation of choice behaviour under conditions of uncertainty, and that therefore plausibly passes operationalist muster. We had better mean von Neumann-Morgenstern utility by our cardinal talk of well-being, this line of thought urges, because only in that way are we assured of having a notion that really has an 'effective meaning, and not merely an appearance of such in a metaphysical-verbalistic sense'. Such operationalism, then, whether or not it was in fact Harsanyi's position, would rationalize Harsanyi's refusal to acknowledge the alleged VNM-independent notion of well-being whose existence and importance Sen is trying to urge.

*Primitivists* go to the other extreme. According to them, a concept can simply be *primitive*, and such that we have intuitive access to when it does and does not apply; further, provided we find ourselves willing to claim such intuitive access, no additional checks are required to ensure that it genuinely has content. In contrast to Harsanyi's refusal to discuss the issue at all, Sen is explicit in his rejection of operationalism with respect to the (i- and co-)cardinal notion of well-being, and in suggesting that his dispute with Harsanyi results from this difference. This is illustrated most clearly in the following conversation that Sen imagines between two fictional characters. Although Sen names the characters '1' and '2', the reader may easily infer that '1' is supposed to speak for Harsanyi, '2' for Sen:

> 1: '. . .[W]hat do these cardinal welfare numbers stand for? What meaning can we attach to them since they are not von Neumann-Morgenstern numbers?'
>
> 2: 'They reflect my views of the welfare levels and gaps . . .'
>
> 1: 'But I can't relate them to your observed behaviour.'
>
> 2: 'I should think not. Nor can I relate your von Neumann-Morgenstern numbers over *interpersonal* choices to your observed behaviour. . . No, these numbers reflect my introspection on the subject as do yours, I presume.'[25]

We have, then, a possible diagnosis of the Sen-Harsanyi dispute in terms of differing background methodological commitments. According to this diagnosis, the primitivist Sen takes himself to have a determinate (at least i-)cardinal notion of well-being, furnishing (intrapersonal) unit comparisons that are altogether independent of the von Neumann-Morgenstern utility scale. He therefore complains that Harsanyi has done nothing to justify the (on his view, substantive) claim that the von Neumann-Morgenstern utility scale cardinally coincides with this independent well-being scale. The perhaps-operationalist Harsanyi, sceptical of Sen's alleged independent cardinal notion of well-being, takes himself literally not to know what Sen is talking about, and for that reason refuses directly to address Sen's question. Sen, in turn, rejects both operationalism and its resulting demand for further explication of the cardinal structure of well-being,

[25] Sen, 'Welfare Inequalities and Rawlsian Axiomatics', pp. 249-50; emphasis in original.

via an apparent insistence that it is unproblematic simply to take the notion in question as primitive, and to claim access to it via intuition (or 'introspection').

If that is the dispute, who is in the right? *The immediate answer is 'neither'*, in the sense that neither operationalism nor primitivism is a tenable account of necessary conditions for a term to have determinate content: operationalism is vastly too restrictive, primitivism too permissive.

That operationalism is too restrictive has long been recognized. For one thing, its insistence on providing a measurement procedure for theoretical terms *one by one* means that it inadvertently rules out even paradigm cases of meaningful theoretical terms as being 'devoid of content'. One cannot garner from the theory of Newtonian mechanics, for instance, an operational definition of 'force', or 'mass', or 'inertial frame' *alone* in terms of measurement procedures: any attempt to define 'mass' (say) inevitably requires use of the notions of force and/or inertial frame, in addition to more 'observational' terms. It is also clear, though, that even a more holistic variant of the operationalist criterion cannot be the full story, since it addresses only the question of how one might confer meaning on a new 'theoretical' term *assuming a background of 'observational' terms that are already - somehow - understood.* This is in danger of merely shifting the question: a fundamental theory must explain also how those 'observational' terms get their meanings. *Perhaps* this line of thought is on the road to the complete and correct theory of content-determination, but it is as yet *only* on the road, and it is not yet sufficiently clear where the road leads (a question to which we will return in section VI).

Turning, on the other hand, to primitivism: the point here is that the operationalists' worries about pseudo-notions that have 'only a metaphysical-verbalistic appearance' of having genuine content were not groundless. Faced with an interlocutor whose concerns fail to move one, scepticism about whether or not those concerns really have any content sometimes is the correct diagnosis of the opponent's mistake. For an artificial example of this, consider an interlocutor who expresses the greatest degree of curiosity regarding the [apparent] question, 'Is the Prime Minister's house blorg?' 'Blorg' being a new word that has never been used except in this particular question, there is simply no content (or insufficient content) to the question; should the interlocutor insist that the question *must* have a determinate answer, even if we have no idea of what might constitute evidence for or against any particular answer, the correct response is to suspect him of minor insanity. Nor will it help to claim introspective ability to detect the answer, absent some other means, *themselves independent of introspection,* of fixing the content of the question. Nobody suggests that the Harsanyi-Sen discussion of 'well-being' is quite like this, but the possibility of it having something crucial in common with this case cannot be ruled out *ab initio.* The proposal I wish to investigate is that in some such way, the Sen-Weymark question of whether or not von Neumann utility *really is well-being* fails to have (sufficient) content, and that Harsanyi was therefore correct not to grasp it. Section V fleshes out this proposal in more detail.

# V. SEMANTIC INDETERMINACY

The problem with questions involving the word 'blorg' is that that word suffers from *semantic indeterminacy*: while nothing physically prevents one from uttering the word, there are nowhere near enough matters of fact regarding *what it means* for questions in which it appears to have determinate content.

The purpose of the present section is to explore the relationship of this phenomenon of semantic indeterminacy to the Harsanyi-Sen-Weymark debate. Section V.1 sets out a variety of further examples, to illustrate the generality of the phenomenon, and the variety of ways in which it can arise. Section V.2 introduces some key terminology for theorizing about such cases. Section V.3 makes use of this terminology to articulate several rival hypotheses regarding the semantics of well-being talk. We will not yet try to *evaluate* those hypotheses (a task we consider later, in section VII); instead, Section V.4 motivates that later discussion by investigating the *significance* of such semantic questions for the evaluation of the Harsanyi-Sen-Weymark debate. This discussion will have raised the question of how else, besides via pegging to a VNM scale, the notion of well-being might be (i-)cardinalized; Section V.5 investigates the possibilities. The key message from the present section will be that *if* 'well-being' is semantically indeterminate then, *contra* Sen and Weymark, Harsanyi's claim to have defended utilitarianism is entirely appropriate.

## V.1. *Further examples*

It will be helpful first to appreciate the generality of the phenomenon we have observed in our first, very artificial, example ('blorg'), in order to allay any concern that positing semantic indeterminacy for 'well-being' involves an *ad hoc* appeal to an exotic, unusual phenomenon. Further examples, from the equally artificial to the entirely natural, are easy to find.

We will survey four further cases. Our first case illustrates most cleanly that semantic indeterminacy can arise as a result of existing usage being insufficiently rich to secure determinacy:

*Partial stipulative definition.* Let us stipulate that the predicate 'is a luper' is to be true of lupins, and false of non-flowers. Nothing in this (limited) usage of 'luper' suffices to determine whether or not the sentence 'daffodils are lupers' is true. The question of whether or not it is true suffers from semantic indeterminacy.

The case of partial stipulative definition, however, is thoroughly artificial. Our second and third examples are taken from real life, and illustrate that the same thing can happen in natural contexts, whether 'ordinary language' or 'scientific':

*Vague boundary on an underlying continuum.* Many predicates of natural language attempt to impose binary 'yes'/'no' judgements on an underlying continuum or quasi-continuum of physical states of affairs. Notoriously,

they generally fail to do so with exact precision: precisely how many hairs does someone have to have on his head before the predicate 'bald' ceases to apply to him? Precisely where, around the base of Snowdon, is the boundary such that all and only people whose centres of gravity are physically located inside that boundary count as being 'on Snowdon'? The existence of this phenomenon is easy to understand in terms of semantic indeterminacy: it is *indeterminate* whether the predicate 'is bald' picks out the property of having fewer than 1,057 hairs, that of having fewer than 1,058 hairs, etc., and similarly indeterminate precisely which region of space corresponds to 'is on Snowdon'.

*Term in a partially incorrect theory.* We turn now to a scientific example. Special relativity (SR) holds that an object's momentum $p$ is the product of its rest mass $m_0$, its velocity $v$ and a factor $\gamma$, where $\gamma$ itself depends on $v$ (tending to unity in the limit $v \to 0$). Newtonian theory, meanwhile, holds that momentum is given simply by the product of mass and velocity, with mass independent of velocity. To what does the term 'mass', as used in the Newtonians' mouths, refer? If it refers to rest mass, then the Newtonian claim $p = mv$ is at best approximately true, and then only at low velocities; if, on the other hand, it refers to the ratio of momentum and velocity (i.e. to relativistic mass), then the Newtonian claim that mass is independent of velocity is false. According to special relativity, there is no quantity that would make *all* of the claims Newtonians made using the term 'mass' come out true. Field[26] has argued persuasively that in this case, there is *no fact of the matter* as to which of the two SR-sanctioned quantities the old term referred: 'mass' as used by Newtonians was indeterminate in reference.

We turn finally to a case that is again somewhat more artificial than the previous two, but that is in one key sense more closely parallel to the state of play concerning well-being, since it relates specifically to the fact that cardinal structure goes beyond ordinal structure:

*Ordinality without cardinality.* Suppose that one has received a large shipment of cubic packages, of various sizes. One needs, with one's teammates, to store these packages in a room that is accessible via a number of doorways; the doorways too are of various sizes. In a strategy discussion aiming to identify the most efficient packing process, one of the team members raises the following would-be question: 'If my package is twice as big as yours, do I need a doorway that is twice as wide as the smallest doorway your package would fit through?' A moment's reflection shows that this question has no determinate answer, because its very content is indeterminate. While our existing usage of size-talk ('big') amply suffices to determine the conditions under which one cubic package counts as being 'bigger' than

[26] Field, H. (1973). Theory change and the indeterminacy of reference. *Journal of Philosophy*, 70(14), 462–481.

another, nothing in existing usage determines whether 'twice as big' means having *twice the side length*, or *having twice the face area*, or *having twice the volume*. Our would-be questioner expresses a question whose answer is positive under the first of these candidate meanings, and a question whose answer is negative under the second and third candidate meanings; there is therefore no fact of the matter as to the correct answer to 'the' question he *did* ask.

## V.2. *Partial denotation and semantic indeterminacy*

We have thus far described our cases of (actual or apparent) indeterminacy at a relatively informal level. To make progress, we need to become more rigorous in our semantic terminology: what, *precisely*, is to be said about the meanings of the problematic words in such cases, and of the truth-values of sentences in which they appear?

In the cut-and-dried examples that are the staples of first-year-undergraduate predicate logic courses - perhaps, 'John is over 1.5m tall' - we can explicate the semantics of a sentence by assigning a particular object (here, John) as the *denotation* of the subject term 'John', assigning a particular property (here, that of being over 1.5 m tall) as the denotation of the predicate term 'is over 1.5 m tall', and taking the sentence as a whole to be true just in case the denoted object has the denoted property. The indeterminacy-diagnosis of the cases discussed in V.1 is then that the ordinary-language term 'bald' (for instance), unlike 'is over 1.5 m tall', is *indeterminate* in denotation between a large number of precise properties (having fewer than 1,000 hairs on one's head, having fewer than 1,001 hairs on one's head ...). Furthermore, this line of thought continues, if the denoted person has some but not all of the properties that are reasonable candidates for being 'the' denotation of the predicate 'is bald', there is no fact of the matter as to whether or not the sentence 'Tom is bald' is true. We are right to become evasive if someone presses us on the question of whether or not a person with precisely 2,539 hairs on his head is bald - we recognize that the sentence is neither determinately true nor determinately false, and that there would therefore be something misleading (at best) about answering either in the affirmative or in the negative.

To regiment talk of such cases, it will be helpful to introduce some further terminology. Given any fragment of language - say, a collection of names and predicates - an *admissible interpretation* of that language-fragment is a completely precise assignment of entities in the world to items of language, in such a way as to do no violence either to the existing meanings of the words in question taken individually, or to systematic connections between them. An admissible interpretation is, therefore, a *precisification* or *sharpening* of our actual, somewhat indeterminate, language. Relative to any such admissible interpretation, names and predicates have determinate denotations, and sentences have determinate truth-values. In cases of semantic indeterminacy there are, however, many equally admissible interpretations, and nothing (in our actual

usage of language or otherwise) to privilege any one over the others. In non-interpretation-relative terms, therefore, rather than taking the basic semantic relation (between items of language, such as names and predicates, on the one hand, and entities in the world, such as objects and properties, on the other) to be that of denotation *simpliciter*, we might take it instead to be one of *partial denotation*:[27] say that an item T of language *partially denotes* an entity X in the world iff X is assigned to T by one or more of the admissible interpretations of our language. (Thus, for example, 'is bald' partially denotes the property of having fewer than 1,000 hairs on one's head, partially denotes the property of having fewer than 1,001 hairs on one's head, partially denotes the property of having fewer than 2,059 hairs on one's head, etc.) Say that T *fully denotes* X if T partially denotes X, and does not partially denote anything else; say that T is *denotationally indeterminate* iff there is more than one entity that T partially denotes. Our cases of interest are cases of denotational indeterminacy. Somewhat similarly for the notion of truth: say that a sentence S of natural (vague) language is *true* iff it is true relative to all admissible interpretations, false iff it is false relative to all admissible interpretations, and indeterminate in truth-value (neither true nor false) iff it is true on some admissible interpretations, false on others. (For more in-depth exposition and discussion of this 'supervaluationist' apparatus, see e.g. Fine[28] and Williamson.)[29]

## V.3. *Rival semantic hypotheses about 'well-being'*

Cases of baldness, mass and so forth are useful to our case of interest only by way of analogy. The relevant semantic question in the context of the Harsanyi-Sen-Weymark debate is: when pre-1940s theorists discussed 'well-being' (or 'welfare'), how determinate was the content of their discussion?

Let us grant for the sake of argument that all partial denotations of 'well-being' agree with one another at least on the *ordering* of possible lives for a given individual; on, that is, the facts about which possible lives are better than which others for the individuals living them. This yet leaves open various possibilities on cardinal matters. The following list is not exhaustive, but includes the semantic hypotheses whose rivalry is interesting from the point of view of the Harsanyi-Sen-Weymark debate.

*(HC-VNM)* 'Well-being' fully denotes some cocardinal quantity that is i-cardinally equivalent to iVNM utility.

*(HC-¬VNM)* 'Well-being' fully denotes some cocardinal quantity that is i-ordinally, but not i-cardinally, equivalent to iVNM utility.

*(HC-Indet)* 'Well-being' partially denotes a cocardinal scale that is i-cardinally equivalent to iVNM utility, and partially denotes at least one cocardinal

[27] The terminology follows Field, 'Theory Change and the Indeterminacy of Reference'.

[28] Fine, K. (1975). Vagueness, truth and logic. Synthese, 30(3), 265–300.

[29] Williamson, T. (2002). *Vagueness*. Routledge, ch. 5.

scale that is equivalent to iVNM utility i-ordinally but not i-cardinally. (All partial denotations of 'well-being', however, are cocardinal quantities.)

## V.4. *The significance of the semantics of 'well-being' for the Harsanyi-Sen-Weymark debate*

Let us defer (until section VII) the question of which (if any) of the above semantic hypotheses is correct, and first investigate the conditionals: *on the assumption of* any of the above hypotheses, what would follow for the Harsanyi-Sen-Weymark debate?

If (*contra* each of the three semantic hypotheses listed above) 'well-being' failed to denote *any* cocardinal scale (either because it failed to denote anything at all, or because it denoted some scale that was not cocardinal - for example, the (merely i-cardinal) iVNM scale), then the utilitarian's assertion would suffer from reference failure: either because the term 'well-being' itself so suffers, or because 'the sum-over-individuals of well-being' does. (If (HI-VNM) is true then there is a sense in which *weighted* (but not unweighted) utilitarianism is coherent, albeit, as we noted earlier, in a slightly different sense; again Harsanyi has proved the thesis in question correct.)

The more interesting hypotheses for our purposes are the listed three. If either (HC-VNM) or (HC-¬VNM) is true then utilitarianism, and its weighted weakening, are at least *coherent*; in the first case (subject to the soundness of the theorems) Harsanyi has proved weighted utilitarianism true, whereas in the second he has proved weighted utilitarianism (i.e. again, the claim that overall goodness is represented by a weighted sum of individuals' *well-being* levels) *false*. On the face of it, therefore, Harsanyi appears to be dogmatically insisting that (HC-VNM) is true, and refusing to offer any justification for that insistence; Sen and Weymark worry that it might turn out to be (HC-¬VNM) instead that is true.

There is, however, another possibility. If (HC-Indet) is true then utilitarianism is *coherent* - it is at least determinately the case that 'well-being' refers to some cocardinal quantity, as required for the utilitarian's summative claim to make sense - but it is (at best) *indeterminate* whether or not utilitarianism is *true*. (It might in principle be determinately *false*: this would be the case if the correct 'social welfare' (overall betterness) function failed to exhibit separability of persons.) This indeterminacy in turn arises because the utilitarian claim is indeterminate in content: indeterminate between claims that overall goodness is the sum of each of the various partial denotations of 'well-being'.

I claim that *in this last case too, Harsanyi's behaviour is vindicated*, but in a different way. To see this, consider again the example of the term 'mass' in Newtonian mechanics and in special relativity. As Field's case study reports, modern textbooks on special relativity often apparently take a stand on whether the term 'mass' is to refer to rest mass, or to relativistic mass: they each offer *a* definition of 'mass' as either one or the other quantity. One *might*

take the authors of these textbooks, in so doing, to be defending conflicting semantic hypotheses, viz. that the term 'mass' *already in pre-relativistic times* determinately referred to the quantity singled out in their respective preferred definitions. But this is not a plausible reading of the situation. More plausibly, the author recognizes that (or simply does not consider the question of whether) *pre-relativistically* the denotation of 'mass' was indeterminate, but, in the service of clarity, is now making a *stipulation* that henceforth the term 'mass' is to refer to rest mass (respectively relativistic mass). If his stipulation is successful - if, that is, it catches on sufficiently widely in the relevant linguistic community - then it amounts to a self-fulfilling statement: it *causes* it to be the case that 'mass' henceforth refers to what this author says it refers to, resolving the unhelpful semantic indeterminacy that previously existed. The case is therefore one of language evolving in response to advances in theory, rather than semantic analysis of the status of historic language-use during some period of inferior theoretical understanding. Of course there are rules even to this game: it would be unhelpful to stipulate that 'the eradication of poverty' is to refer to the state of having a particle accelerator capable of operating at energies of 14 TeV, and then to announce 'we have achieved the eradication of poverty'; some respect for pre-existing usage is required. But cases like our 'mass' example are not these gratuitous changes of the subject, provided that the new usage is a *precisification of*, rather than a *deviation from*, the old one (provided, that is, that the class of interpretations that are admissible given the new usage is a proper *subclass* of those that were admissible relative to the old usage, so that new usage does not count as true anything that would be *false* relative to old usage, but only renders determinate some cases of previous indeterminacy). The unfortunate feature of the 'mass' case is not that authors attempted precisifications of the pre-existing and denotationally indeterminate term, but only that they failed to agree on a unique such precisification, thereby generating unnecessary linguistic confusion (in, for example, the unfortunate physics undergraduate diligent enough to peruse more than one such textbook). Stipulative precisification is clearly the right reaction to discoveries of semantic indeterminacy in such cases: cases, that is, in which (unlike that of baldness) new and better theory wishes to make use of distinctions that older theory failed to draw. What else is one to do - retain slavish adherence to the old usage and thereby deny oneself the now-required greater expressive power that precisification would offer?

Returning now to the case of immediate interest: if Harsanyi had realized that 'well-being' is denotationally indeterminate between some quantity/ies agreeing i-cardinally with iVNM utility on the one hand, and some quanty/ies not so agreeing on the other, what should he have done? Had he either asserted or denied that the (weighted-)utilitarian claim *in the mouths of the 19th century utilitarians* would have (determinately) expressed a truth, he would himself have said something false. Had he simply dropped the subject, he would have missed out on the important discovery that on one particular precisification, that claim is true. Far better, then, to make a stipulative precisification of 'well-being' that succeeds in rendering the claim in question either determinately true or determinately false, and then establishing which is the case. On the hypothesis

that (HC-Indet) is true, Harsanyi can be castigated for not having *explained* that his business was one of making such a stipulative precisification, rather than making any claim about the semantics of pre-existing usage, and perhaps also for not having noted that any claim that this stance does no violence to pre-existing usage depends on the assumption that (HC-Indet) is indeed true. But that is all.

(Something like the hypothesis (HC-Indet) has been proposed before. The following suggestion (with respect to 'utility' rather than 'well-being') was made, by means of an analogy to temperature, already by von Neumann and Morgenstern:

> Given a physical quantity, the system of transformations by which it is described by numbers may vary in time, i.e. with the stage of development of the subject. Thus temperature was originally a number only up to any monotone transformation. With the development of thermometry . . . the transformations were restricted to the linear ones . . .
>
> For utility the situation seems to be of a similar nature.[30]

It is a little more explicit, although with an epistemological (as opposed to semantic) gloss, in the following passage by Broome:

> I doubt we have [an intuitive grasp of what is good] that is adequate for the purposes of utilitarianism. Utilitarianism requires good to be quantitative... It is not enough for utilitarianism that things should be ordered by their goodness, so we have concepts of better and worse. We also need a concept of how much better one thing is than another. I doubt we have a clear intuitive concept of good that is quantitative in this sense.[31]

These remarks (especially the latter) are perhaps more suggestive of the hypothesis that 'well-being' initially (but determinately) denotes a merely *ordinal* quantity than of (HC-Indet) , but if so, I conjecture that their authors would be happy enough to accept (HC-Indet) as a friendly amendment. Note that in contrast to (HC-Indet) , if the 'ordinalist' hypothesis is true then Harsanyi's stipulation that 'well-being' is henceforth to refer to some cocardinal quantity i-cardinally equivalent to VNM utility is indeed a *revision*, not a precisification, of prior usage.)

---

[30] von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behaviour.* Princeton University Press, p. 23.

[31] Broome, J. (2008). Can there be a preference-based utilitarianism? In M. Fleurbaey, M. Salles, & J. Weymark (Eds.), *Justice, political liberalism and utilitarianism: Themes from Harsanyi and Rawls* (pp. 221–238), at p. 222.

## V.5. *Alternative candidate meanings*

In specifying the (Sen-Weymark) hypothesis (HC-¬VNM) above, we alluded to 'some cardinal scale that agrees with VNM utility i-ordinally but not i-cardinally', but we have not attempted any direct specification of such a cardinal scale. It will be helpful to now to make the latter attempt: what can be said to fix some particular i-cardinal scale for discussion?

This question is actually somewhat more difficult to answer than the analogous question concerning the candidate semantic values for the examples of 'blorg', 'luper', 'bald', 'on Snowdon', 'mass' or 'twice as large as' discussed in section V.1. One can specify non-VNM i-cardinal quantities *derivatively*, via the notion of iVNM utility: as noted in section III, for any increasing but non-affine function $f$, $f\left(u^{VNM}\right)$ is such a quantity. If we insist on a specification of the i-cardinalization of the well-being scale that is not in this way derivative on the VNM utility scale, however, our task becomes somewhat harder. In fact, I am aware of only three attempts to address this task; and of those, only one has (as far as I am aware) been defended in print.[32]

First suggestion: assume hedonism. In that case, a person's well-being level at a given time, being identical to her degree of happiness, supervenes on the physical state of her brain, or at least of her body, at that time. Presumably, though, not every difference in physical state of the brain makes a difference to well-being level. Define the space of *hedonic states* to be the coarse-graining of brain-state-space that retains all information that is relevant to amount of happiness. (Thus, an hedonic state is a set of happiness-equivalent brain states.) Assume that the space of hedonic states is equipped with a natural total ordering, corresponding to the 'more happy than' relation. Assume further that this ordering is locally finite, i.e. that for every pair of hedonic states, the number of hedonic states lying strictly between them is finite. We can then *define* the well-being difference between hedonic states A and B to be the number of steps required to proceed from A to B through the ordering (i.e. the number of hedonic states lying strictly between A and B, plus one). The assumptions required for this to well-define a quantitative notion of well-being difference - in particular, the assumption of local finitude, never mind the assumption of hedonism itself - may be highly doubtful. However, granting the required assumptions for the sake of argument, the procedure sketched here does at least provide *a* means of (at least i-)cardinalizing the well-being scale, and one that does not piggy-back on the notion of von Neumann-Morgenstern utility.[33]

[32] The difficulty of the question has often been noted in the literature on prioritarianism: see, e.g. Broome, J. (1991). *Weighing goods.* Oxford: Blackwell; Parfit, D. (2012, September). Another defence of the priority view. *Utilitas*, 24, 399-440; Greaves, H. (2015). Antiprioritarianism. *Utilitas*, 27(01), 1–42.

[33] Variations on this theme are explored by Edgeworth, F. Y. (1881). *Mathematical psychics: An essay on the application of mathematics to the moral sciences.* C. Keagann Paul,

Second suggestion: again assuming hedonism, whatever turns out to be the physical/neurophysiological basis for happiness is likely to come with its own, physically privileged cardinalization; this physical cardinalization then supplies a candidate denotation for i-cardinally committed well-being talk. If, for example, degree of happiness turned out to be determined by the number of neurons of a given type firing per second, well-being level could simply be equated with those numbers; if degree of happiness turned out to be determined (instead) by the concentration of some particular hormone in the brain, then well-being level could be equated with the concentration of that hormone. The question of whether or not any such i-cardinalization really is available, clearly, is hostage to the neurophysiological basis of happiness, and as far as I know is as yet unsettled.

Third suggestion: heuristically, the reason consideration of betterness orderings of lotteries is able to supply an i-cardinalization is that an assumption of *separability* with respect to states of nature (for example, Savage's sure-thing principle) is plausible for such orderings. The same mathematical ideas can be used in other contexts, whenever a betterness ordering obeys a formal separability condition with respect to some partition of the determinants of betterness (together with auxiliary conditions that will not concern us). If, for instance, the betterness ordering for possible lives is determined by 'consumption' along a number of dimensions (e.g. health, pollution level, house size, education level, food quality) and *if* a separability condition obtained with respect to those dimensions, then the betterness ordering of possible lives would be represented by a utility function that is a sum of 'subutility' functions, one subutility for each dimension of consumption, and (crucially for our purposes) the overall utility function is determined up to positive affine transformation by this representation. In that case, the requirement that a numerical representation of the betterness ordering for lives be a sum of subutilities for individual dimensions of consumption picks out an i-cardinalization.[34] The challenge for this approach is finding additional interpretations under which the separability condition really

pp. 7ff., 60ff., 98ff.; Ng, Y.-K. (1975). Bentham or Bergson? Finite sensibility, utility functions and social welfare functions. *The Review of Economic Studies*, 545–569; and Tännsjö, T. (n.d.). *Utilitarianism or prioritarianism?* (Unpublished). It is also the basic idea behind the Borda count.

[34] For discussion of various results connecting separability conditions to additive representations and of the range of possible applications of those results, see. e.g. Blackorby, C., Primont, D., & Russell, R. R. (1998). Separability: A survey. In *Handbook of utility theory: Volume 1: Principles.* Springer, sec. 5; Von Winterfeldt, D., Edwards, W., et al. (1986). *Decision analysis and behavioral research* (Vol. 604). Cambridge University Press Cambridge, pp. 331-4; Broome, *Weighing Goods.*

is plausible; this is not the case in the example just suggested.

Setting aside this difficulty of actually singling out a determinate cardinalization in a manner that is *independent* of the notion of VNM utility, however, the fact remains - as the Sen-Weymark critique highlighted, and the related piggy-backing method of scale-specification constructively demonstrates - that there *are* quantities that are i-ordinally but not i-cardinally equivalent to iVNM utility. The question is whether any of them is (i-cardinally equivalent to) well-being.

Note well that this question is a *semantic* one. In sections VI and VII, I will make a preliminary attempt to answer it, and will suggest that the only plausible hypotheses are (HC-VNM) and (HC-Indet) . This will, however, require tangling with the details of metasemantic theory: we will need to take on the question of what (given the failure of section IV's operationalist and primitivist attempts) are the rules governing the association of words with meanings (section VI), and the application of those rules to the present case (section VII). Readers who are uninterested in such matters of (mere) semantics can skip to the section VIII.

# VI. ASSESSING RIVAL SEMANTIC HYPOTHESES, PART I

## VI.1. *A rudimentary metasemantic theory*

To take on the question of which of the hypotheses outlined in section V.3 is, as a matter of semantics, correct, we require a systematic understanding of the conditions under which a term has determinate content - and of what that content then is - that is neither as restrictive (and restricted) as operationalism, nor as permissive as primitivism.

Let us take a step back. We take it as a given that the semantic facts - which items of language (nouns, predicates and so forth) refer to which entities (objects, properties and so forth) in the non-linguistic world - must supervene, somehow, on the non-semantic facts. (That is: there cannot be two possible worlds, exactly alike in their physical descriptions, including the descriptions of which agents use which words in which ways in which circumstances, but such that the referent of (say) 'tree' in the language of the speakers in one of the possible worlds is distinct from (and not even qualitatively identical to) the referent of 'tree' in the language of the speakers in the other possible world. To put it another way: there cannot be 'magic semantic strings' connecting words to meanings in a manner that is entirely disconnected from, and inexplicable in terms of, a complete physical description of the world.) This much is a platitude; the open question is only *how* the semantic supervenes on the non-semantic.

In slightly more detail: Let an *interpretation* of a pattern of linguistic behaviour be a function that assigns, to each item of language, a corresponding non-linguistic entity as its referent. (For example, the standard interpretation of the linguistic behaviour of the French assigns the city of London to the term

'Londres', the property of being a tree to the term 'arbre', and so on.) In principle there is, for any pattern of linguistic behaviour, an infinite number of possible interpretations. Some, however, will be *incorrect*. What determines the actual referent of 'arbre' must be something about the way the community of language-users under consideration uses the word in question. (It is determinately the case, for instance, that the referent of the French word 'arbre' is *not* the property of being a cat, nor is it the portion of space less than 10,000 km from the centre of the Earth. It is equally clear, however, that had the French used the word 'arbre' differently, these alternative entities could have been the referent of the linguistic item 'arbre'.) We seek, then, a set of principles that captures the way in which language-use determines the correctness and incorrectness of interpretations.

From this point on in particular, there is no universal consensus on the details. It will be helpful, though, to have on the table an example of the sort of thing that might be attempted by way of answer. A standard approach takes the semantic facts to be fixed, roughly speaking, jointly by a principles of *charity* and *eligibility*. (Causal principles are also presumably some part of the story: the fact that 'blue' means *blue* rather than *green* (respectively, that 'David Cameron' denotes Cameron rather than Miliband) presumably has something to do with the causal connections between the existence and presence of blue objects (respectively of Cameron) and some utterances of the now-corresponding words. Since such causal principles appear to be less important for more deeply theoretical than for more observational terms, and 'well-being' falls in the 'more theoretical' category, however, I will largely set causal considerations aside.)[35]

A principle of charity favours either interpretations that render the majority of utterances of sentences by the linguistic community *true*, or (a different type of charity-principle) interpretations that assign meanings to those sentences in such a way that the meanings in question are things that the speakers could *reasonably have believed*, given their evidence and cognitive resources. Thus, 'green' means *green* partly in virtue of the fact, perhaps, that many speakers have uttered sentences like 'grass is green' and 'trees have green leaves', and (holding fixed the interpretation of the other components of these sentences) these utterances would express falsehoods if 'green' meant, say, *blue* or *heavy*.

[35] For overviews of attempts to articulate a causal principle as whole or part of metasemantic theory, see Neander, K. (2012). Teleological theories of mental content. *The Stanford Encyclopaedia of Philosophy (Spring 2012 edition)*. (Online at http://plato.stanford.edu/archives/spr2012/entries/content-teleological/); Rupert, R. D. (2008). Causal theories of mental content. *Philosophy Compass*, 3(2), 353–380. On the 'charity plus eligibility' programme, see especially Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4), 343–377; Lewis, D. (1984). Putnam's paradox. *Australasian Journal of Philosophy*, 62(3), 221–236.

Considerations of eligibility are required to rule out gerrymandered meanings that would enable the principle of charity perfectly to be satisfied, but only because they are 'cooked up' to do precisely that: thus, taking the truth-based principle of charity by way of example, if I mistakenly assert 'my front door is blue', I succeed in asserting a falsehood (the principle of charity notwithstanding) in part because the interpretation that assigns to 'blue' *the set of all blue things together with my (yellow) front door* is a more gerrymandered, hence less eligible, candidate denotation for a predicate term than is *the set of all blue things*. (A rationality-based principle of charity equally requires appeal to eligibility, since it requires some claim that it is more rational to project natural than 'gruesome' predicates.[36]) The claim - whichever particular set of principles the metasemantic theorist eventually settles on - is then that the matter of which interpretation of the language in question is correct is *determined* (ontologically, not merely epistemically) by the matter of which interpretation best fulfils this set of principles; if one interpretation better satisfies some of the principles while another better satisfies others, tradeoffs must be made.

Against this background, the phenomenon of semantic indeterminacy is easy to understand. The world is equipped with an enormous number of reasonably natural (eligible) objects, properties and so forth, any of which could in principle be the denotation of a given term in our language, *if* the language-use facts fell in such a way as to make that so. According to the present approach to metasemantics, we have seen, an interpretation - an assignment of semantic contents to terms of language - is the *correct* interpretation of a given language iff that interpretation optimizes the joint satisfaction of the principles of the correct metasemantic theory. But in some cases - in particular, cases in which the actual use of the item of language in question is relatively impoverished - there will plausibly be *many* interpretations that are equally good by the lights of the principles of metasemantics, and nothing in our actual use of language to settle the choice between those interpretations. In such a case, it will be indeterminate what the denotation of the term in question is.

## VI.2. *Application to the examples: how indeterminacy might arise*

It is straightforward to explain the above cases of (relatively uncontroversial) indeterminacy within this framework.

*Indeterminacy in cases of sudden introduction.* In our 'blorg' example: actual usage of the term 'blorg' is far too thin to have settled any more than that its denotation can only be some property or other, and in fact (since a singular term in the place 'Is the prime minister's house ...' would be at least grammatically correct) probably not even that much. At most, a

[36] See Weatherson, B. (2013). The role of naturalness in Lewis's theory of meaning. *Journal for the History of Analytical Philosophy*, 1(10).

rationality-based principle of charity applied to the asker of the question 'Is the prime minister's house blorg?' might determine that the only entities partially denoted by 'blorg' are properties F such that it could plausibly be thought to be (i) of interest and (ii) uncertain whether or not the prime minister's house has property F. This, though, leaves an enormous number of partial denotations (modern, in the south-east, expensive, white ...).

*Indeterminacy in cases of partial stipulative definition.* Our 'luper' example is similar. Appeals to use alone are unlikely, in this case, to settle anything other than what is contained in the explicit stipulations. Our best hope, in search of determinacy, is to break the tie between the large number of candidate meanings by appeal to eligibility. But nor does it seem likely that this is possible: the candidate meanings *lupin* and *flower*, for a start, seem equally natural, as does *member of the family fabaceae*. The term 'luper' is denotationally indeterminate between at least these; that is why the sentence 'a daffodil is a luper' has no determinate truth-value.

*Indeterminacy in Sorites cases.* The analysis of these cases in terms of indeterminacy is well known. Our term 'bald', for example, has been applied to numerous people with *very few* hairs on their heads, and explicitly withheld from many people who have *very many* hairs on their heads. But there is a significant range of borderline cases, such that in practice we generally refrain from either affirming *or denying* that the predicate 'bald' applies to those cases. (Similar remarks apply to 'is on Snowdon': one might well assert 'he is on Snowdon' of a climber known to be halfway along Crib Goch, and one might well deny it of someone known to be safely ensconced in Pete's Eats some 6 miles away; given an adventurer just setting out from the car park at the base of the mountain, though, one would tend to evade any question as to whether or not he is, right now, 'on Snowdon' ('well, he's just setting out from the car park'). The candidate denotations, then, are the precise locations one might draw a boundary somewhere between the clear positive and the clear negative cases: *having fewer than 1,000 hairs on one's head/having fewer than 1,001 hairs on one's head/*..., or any of the infinitely many precise physical boundaries one might draw to indicate the precise limits of 'being on Snowdon'. Given the sort of usage pattern that occurs in such cases, appeals to charity seem unlikely to favour one such candidate denotation over another; and it is particularly clear in these cases that no such candidate is likely to be more natural than any other.

*Indeterminacy in cases of theory change.* From the standpoint of special relativity, neither of the quantities *rest mass* or *relativistic mass* is obviously more natural than the other. Nor does anything in Newtonian usage of the term 'mass' seem to tell between these two candidate denotations: relative to an interpretation that assigns either to the term 'mass', some of the Newtonians' claims are false, but those that are false relative to one structure do not seem to have been in any sense more firmly held or

central to Newtonian theory than those that are false relative to the other structure. We might well conclude (with Field), then, that the Newtonians' term 'mass' is denotationally indeterminate, partially denoting rest mass and partially denoting relativistic mass.

*Indeterminacy in cases of ordinality without cardinality.* In our cube-size example, there seem to be three equally natural candidate meanings for the 'twice as big as' relation between cubes: having twice the side length, having twice the face area, and having twice the volume. Not only do considerations of naturalness fail to break the tie, but neither does any of our general size-talk seem to favour one of these candidates over the other in terms of fit to use. If so, then our question 'does a cube that is twice as big need a doorway that is twice as wide?' is at least threefold indeterminate in content, and as a consequence has no determinately correct answer.

I say only that indeterminacy-diagnoses of the cases discussed in this section are *plausible*; analysing these cases in terms of denotational indeterminacy is not *inevitable*. A sizeable minority programme ('epistemicism'[37]) holds instead that in such cases, one of the candidate denotations is in fact singled out by patterns of usage, facts of naturalness, and/or whatever else constitutes the input to the correct principles of metasemantics. This is not a crazy thesis, at least with respect to our above examples. Perhaps the Newtonian term 'mass' determinately referred to rest mass on account of that quantity's being marginally *more natural* than relativistic mass (rest mass is, after all, independent of frame of reference, while the same cannot be said of relativistic mass). Perhaps our invented term 'luper' determinately denotes, not the genus *Lupinus* or the order *Fabales*, but the family *Fabaceae*, this being (in the final analysis, but unbeknownst to us now) the most natural kind consistent with our linguistic stipulations. Perhaps (again unbeknownst to us) the term 'bald' determinately denotes, say, the property of having fewer than 2,394 hairs on one's head, this turning out to be the candidate meaning that marginally trumps all others in terms of fit to actual usage, given the way of measuring such 'fit to usage' mandated by the correct theory of metasemantics.

If not denotational indeterminacy, what unites such cases? The analysis that epistemicism offers, as suggested by the programme's name, is epistemic: while there is in each case of (let us say) *apparent* indeterminacy a fact of the matter as to what the denotation of the problematic term is, due (perhaps) to the delicate nature of the tradeoffs that settle which is the denotation and/or our ignorance of some of the facts of usage and naturalness (etc.) that form the input to some of the metasemantic principles, we are *not in a position to know* which it is, and for that reason are not in a position to know the (objectively determinate) truth-values of such sentences as 'a man with 5,679 hairs on his head is bald'. Cases of apparent indeterminacy, then, are cases in which, in a literal and clear sense, we don't know what we are talking about; there is, though, a fact of the matter

[37] See especially Williamson, *Vagueness*, esp. chs. 7,8.

as to what we are talking about. A move from an indeterminacy-framework to an epistemicist-framework affects the logic of some of the finer points of discussion of such cases, but does not challenge the observation that there is *something distinctive* about cases of apparent indeterminacy. I conjecture that the discussion of the present article would be affected only in formulation, not in substance, by a shift to an epistemicist account of what that distinctiveness consists in; investigation of whether or not this is the case, however, lies beyond the scope of this article.

# VII. ASSESSING THE SEMANTIC INDETERMINACY HYPOTHESIS, PART II: THE CASE OF 'WELL-BEING'

In thus appealing to vaguely stated criteria of charity and eligibility, we have sketched only a very vague account of what the right metasemantic theory might look like. That vague account will suffice, however, to mount a preliminary attack on our central question*:* *has* Harsanyi (as Sen and Weymark charge) changed the subject, hijacking the terms 'well-being' and 'utilitarianism' to refer to something quite distinct from their usual referents, so that his description of his theorems' outputs as 'utilitarianism' is mere obfuscation? Or has he established the truth of utilitarianism in the original sense of that term? Or something else?

This - as argued in Section 5.4 - boils down to the question of which (if any) of the semantic hypotheses outlined in Section V.3 is correct.

Assuming that these are the only candidates on the table, the question of whether (HC-Indet) is true is the question of whether either of the other two candidates performs better than the other; let us therefore consider the latter question, addressing considerations of naturalness and charity in turn. (In principle, it might of course be that *none* of (HC-VNM) , (HC-$\neg$VNM) or (HC-Indet) is true: for example, because 'well-being' instead determinately denotes a merely ordinal quantity. Such alternative hypotheses, however, will tend to perform very poorly in terms of charity, since we do in fact make (platitudinous) intra- and inter-personal unit comparisons. Intrapersonal case: it is a platitude, for instance, that a hundred dollars generates a *greater improvement in one's well-being* if one is on the breadline than if one is already comfortably off; this too makes sense only if well-being is i-cardinal. Interpersonal case: it is similarly a platitude that some people *gain more*, in well-being terms, from viewing great works of art than others do, and that our hundred dollars usually generates a greater well-being improvement for a poor person than it does for a *different* rich person.)

On naturalness: two of the candidate interpretations we have considered have some prima facie plausible claim to be reasonably natural. First: an i-cardinalization of well-being that is based on natural features of the physical supervenience basis is reasonably natural by anyone's lights. Second: the VNM

i-cardinalization, arising as it does in a natural way from a plausible evaluative theory, is also arguably a natural one. ('Arguably': whether or not this is the case depends on what the right account is of 'higher-order naturalness', i.e. of the natural/unnatural distinction between properties arising at a level of theorizing that is 'higher-level' than the level of basic microphysical theory. The best-developed account of higher-order naturalness is in terms of length of definition in terms of the natural microphysical predicates.[38] On this account, a cardinalization based closely on the physical supervenience basis would count as vastly more natural than a VNM cardinalization. That account of higher-order naturalness is, however, wildly implausible, notwithstanding the absence of any remotely well-worked-out alternative. This state of the debate leaves us with little idea how to assess the relative naturalness of physical-based and VNM cardinalizations.[39]) On the other hand, the cardinalizations to which Sen and Weymark draw our attention, viz. those specifiable by performing some arbitrary increasing transformation on the iVNM scale, are (in the absence of any simpler, non-piggybacking specification that turns out to pick out the same scale) presumably less natural than the iVNM scale itself.

Charity, meanwhile, is likely to count significantly in favour of the VNM scale over any physically natural rival. The point here is[40] that the VNM scale, since it is derived from *evaluative* rankings of lotteries, is guaranteed to be at least in some reasonable way connected to matters of how *important* a transition between two outcomes (or centred outcomes) is - to *how much difference* that transition makes, in an evaluative sense. There is, in contrast, no particular reason to think that any *physically* natural cardinalization will amount to a reasonable measure of importance: no particular reason, to take the hedonic-state-count proposal by way of example, to think that there might not be many more intermediate states between two hedonic states that ordinary usage would judge 10 units apart high up the well-being scale than between two hedonic states that ordinary usage would judge to be roughly the same well-being-distance apart lower down the well-being scale.

We appear to have, then, charity-based considerations counting in favour of (HC-VNM) over (HC-¬VNM) , and naturalness-based considerations either

[38] Lewis, D. K. (1986). *On the plurality of worlds*. Cambridge Univ Press, p. 61.

[39] For further discussion of the general issue of how to understand higher-order naturalness, see Sider, T. (2011). *Writing the book of the world*. Oxford University Press, ch. 7.11.1; Williams, J. R. G. (2007). Eligibility and inscrutability. *The Philosophical Review*, 361–399; Hawthorne, J. (2007). Craziness and metasemantics. *The Philosophical Review*, 427–440; Hawthorne, J. (2006). Quantity in Lewisian metaphysics. In J. Hawthorne (Ed.), *Metaphysical essays*. Oxford University Press, pp. 236-7.

[40] As noted already by Vickrey Vickrey, W. (1945). Measuring marginal utility by reactions to risk. *Econometrica*, 319–333.

being roughly neutral or counting in favour of (HC-¬VNM) over (HC-VNM) . This discussion has been sketchy, not least because it skirts over many details of what, precisely, the right metasemantic principles are - details that may make a difference to the evaluation of the present case. It is therefore necessarily somewhat inconclusive. It should, however, have served to make the case that neither hypothesis (HC-VNM) nor hypothesis (HC-Indet) (at least) is completely outlandish, and that there are reasons to regard each as more plausible than (HC-¬VNM) .

# VIII. WHAT HARSANYI SHOULD HAVE SAID

This, then, is the fall-back position that I offer to Harsanyi. The question of whether utilitarianism is true or false is initially ill-formulated because prior to the advent of decision theory (at least), cocardinally committed talk of 'well-being' was indeterminate in denotation between several inequivalent cocardinal quantities, not all of which are even i-cardinally equivalent to one another. Taking on an ill-formulated question *directly* is not a route to progress: one must first improve the formulation of the question, by a judicious choice of stipulation if necessary. Responding to this situation, Harsanyi *stipulates* that the denotation of 'well-being' is henceforth to agree i-cardinally with von Neumann-Morgenstern utility. This is a stipulation that is consistent with, but deliberately goes beyond, prior usage of the term: like a stipulation that 'cube twice as big' is to mean cube of twice the *side length,* or that 'mass' is to mean *rest* mass, it does no violence to the sentences involving 'well-being' that were already (determinately) true, but it introduces some further determinacy in regions previously plagued by indeterminacy. (It is, in that sense, a *precisification* rather than an *alteration* of our pre-existing concept of well-being; cf. Section V.2.) In particular, once this stipulation has been made, (weighted-)utilitarian theory has determinate content - determinate enough to be either determinately true or determinately false. The Aggregation and Impartial Observer theorems then show - subject, of course, to the correctness of their substantive assumptions - that the content in question is in fact true rather than false.

This interpretation, and the preceding discussion, throws an unfavourable light on the aspect of the Sen-Weymark critique of Harsanyi that we have been discussing. Harsanyi has done nothing, the critics complain, to establish that von Neumann-Morgenstern utility is i-cardinally equivalent to *well-being*, and without that latter claim his theorems do nothing to establish (even weighted) *utilitarianism*. I have argued that this criticism is itself devoid of content (it has 'only a metaphysical-verbalistic appearance' of having content). The question of whether VNM utility is (i-cardinally equivalent to) *well-being* is identical to the question of whether the denotation of 'well-being' is some quantity that is i-cardinally equivalent to iVNMU, or rather to one that is not i-cardinally equivalent to iVNMU. This is in any case a (merely) semantic question, but (I have suggested) turns out itself to suffer from reference failure: there is no such thing as *the* denotation of 'well-being'. There is, in that case, no determinate

thing in the ball-park of the Sen-Weymark demand that Harsanyi has failed to do - no determinate VNM-independent referent of 'well-being' to show that VNM utility is i-cardinally equivalent to, and no determinate VNM-independent utilitarianian thesis to defend. Harsanyi has, rather, done us all a twofold service: suggested a way of resolving the semantic indeterminacy that previously plagued attempts to articulate a (weighted) utilitarian theory of the good, and provided a cogent (if not knock-down) argument for the truth of the resulting determinate thesis. For this he can only be thanked.

The suggested interpretation also, and for the same reasons, throws an unfavourable light on the suggestions of Mongin[41] and of Fleurbaey and Mongin, [42] viz. that the way to recover utilitarian relevance for Harsanyi's theorems in the light of the Sen-Weymark critique is to bolster those theorems with additional *formal* machinery. These authors introduce additional formal primitives to correspond to a cardinal notion of preference or utility/well-being ('true well-being', let us call it) that is in principle independent of individual VNM utility, but then to impose axioms that (more or less directly) require the individual VNM utilities and the corresponding in-principle-independent 'true well-beings' to be i-cardinally equivalent to one another after all, so that the Sen-Weymark critique is answered. This approach, of course, effectively assumes that we *do* have a VNM-independent but nonetheless determinate referent of 'well-being' to theorize about (and to represent with our additional formal primitive), so that it is a substantive matter - the sort of thing an axiom can assert - what the relationship is of that independent thing to VNM utility. I have argued that this is mistaken. If I am correct, then the additional formal primitives introduced for these theorems are just so much uninterpreted mathematics; to call them 'cardinal preferences' or 'utilities' is not to resolve the semantic indeterminacy.

# IX. CONCLUSION

The Sen-Weymark critique of Harsanyi centres on the point that in order for the conclusions of Harsanyi's theorems to amout to utilitarianism, it must be assumed that well-being and von Neumann-Morgenstern utility agree, cardinally, with one another (in the terminology of section Section III.2, that they are i-cardinally equivalent). Sen and Weymark, along with many others, regard this as a substantive assumption for which Harsanyi has offered no justification, and therefore as being the locus of a significant gap in Harsanyi's argument. The critics would be correct *if* 'well-being' determinately picked out some method of cardinal-scale-fixing that is distinct from that of decision theory, for example the hedonic-state-count method described above: for there *is* a substantive (if not enormously important) question of whether the hedonic-state-count i-cardinalization, if it exists at all, turns out to coincide with the

[41] Mongin, 'Impartiality, Utilitarian Ethics, and Collective Bayesianism'.

[42] Fleurbaey, M., & Mongin, P. (n.d.). *The utiltarian relevance of the aggregation theorem.* (Unpublished manuscript).

VNM i-cardinalization. But I have argued that, instead, the term 'well-being' is at worst denotationally indeterminate between some quantity or quantities i-cardinally equivalent to von Neumann-Morgenstern well-being, and some that are not. If so, then Harsanyi has neither made a dubious substantive assumption nor illicitly changed the subject, but merely made a helpful terminological precisication. There is in any case no disagreement here to be had that is 'substantive' in the sense of going beyond matters of semantics.[43]

*hilary.greaves@philosophy.ox.ac.uk*

# APPENDIX. THE IMPARTIAL OBSERVER ARGUMENT

The theorem at the heart of the Impartial Observer argument can then be formulated[44] as follows. Let $Ext \coloneqq \mathrm{X} \times \mathrm{I}$ be the set of *extended alternatives* $(A, i)$, where $A \in X, i \in I$: $(A, i)$ is the situation of being individual $i$ while the objective description of the world is given by $A$. (In philosophers' terminology, then: these extended alternatives are *centred* outcomes.) Let $L^{Ext}$ be the set of lotteries over $Ext$. Define two subsets of $L^{Ext}$, as follows: (i) for each $i \in I$, let $L_i^{Ext}$ contain just those lotteries all of whose non-null outcomes are centred on the individual $i$; (ii) let $L_{imp}^{Ext}$ contain the 'impartial' lotteries, i.e. those lotteries $\pi \in L^{Ext}$ that are generated[45] from separate probability distributions $p$ on $X$ and $z$ on I, where, in addition, $\pi_I$ assigns equal probability to each individual $i \in I$. Let $\succeq$ be an ordering of $L^{Ext}$ corresponding to the preferences of our 'observer'; as for the Aggregation Theorem, let $\succeq_i$ be the betterness-for-i ranking of $L(X)$. Suppose that the structure $(\succeq, \{\succeq_i : i \in I\})$ obeys the following three axioms:

*IOT1:* $\succeq$ obeys the axioms of expected utility theory.

*IOT2:* For all $i \in I$, $\succeq_i$ obeys the axioms of expected utility theory.

[43] For valuable discussions, I am grateful to Ted Sider, Robbie Williams, and participants in the 2014 Conference on Rational Choice and Philosophy at Vanderbilt University, especially Christian List and John Weymark. Thanks also to an anonymous referee for extremely helpful comments and suggestions.

[44] Harsanyi's own presentations of (especially) the Impartial Observer result are rather informal. The formulation outlined here is close to that provided by Weymark, 'A Reconsideration of the Harsanyi-Sen debate on Utilitarianism'.

[45] I.e. there exist probability distributions $\pi_X$ on X and $\pi_I$ on I such that the probability that $\pi$ assigns to extended alternative $(A, i)$ is given by the product $\pi_X(A) \cdot \pi_I(i)$.

*IOT3 (Principle of Acceptance):* For all $i \in I$ and for all $a, b \in L_i^{Ext}$, $a \succeq b$ if and only if $a \succeq_i b$.

Then, $\succeq$ can be represented on $L_{imp}^{Ext}$ by an expression of the form $V(\pi) = \sum_{i \in I} V_i(\pi_X)$, where, for each $i \in I$, $V_i$ expectationally represents $\succeq_i$ on $L_i^{Ext}$. That is, the 'observer"s preferences over impartial lotteries correspond to maximizing a sum-over-individuals of individuals' expected von Neumann-Morgenstern utilities for outcomes.