Philos Stud
DOI 10.1007/s11098-016-0748-6

Aggregating extended preferences

Hilary Greaves[1] · **Harvey Lederman**[2]

Abstract An important objection to preference-satisfaction theories of well-being is that they cannot make sense of interpersonal comparisons. A tradition dating back to Harsanyi (J Political Econ 61(5):434, 1953) attempts to solve this problem by appeal to people's so-called *extended* preferences. This paper presents a new problem for the extended preferences program, related to Arrow's celebrated impossibility theorem. We consider three ways in which the extended-preference theorist might avoid this problem, and recommend that she pursue one: developing aggregation rules (for extended preferences) that violate Arrow's Independence of Irrelevant Alternatives condition.

Keywords Interpersonal well-being comparisons · Extended preferences · Preference-satisfaction theory · Theories of well-being

1 Introduction

Queen Victoria was better off than an average Roman slave. Further, the degree to which Queen Victoria was better off than an average Roman slave is greater than the degree to which a child who gets an extra scoop of ice-cream is better off than one who does not. The first of these facts concerns the *levels* of well-being of different individuals; the second concerns the *units* of different individuals' well-being, the degree to which one is better off than another. Any theory of well-being must make sense of each of these kinds of interpersonal well-being comparisons.

Hilary Greaves and Harvey Lederman have contributed equally to this paper.

✉ Harvey Lederman
harveyslederman@gmail.com

[1] University of Oxford, Oxford, UK

[2] University of Pittsburgh, Pittsburgh, PA, USA

Published online: 02 September 2016

A preference-satisfaction theory of well-being holds that a person's well-being is in some sense determined by the satisfaction of her preferences. Standard theories of this kind face a challenge in making sense of interpersonal comparisons of well-being. On a standard conception of preferences, the objects of a person's preferences are simply states of affairs or possible worlds. Each individual's preferences induce a ranking of states of affairs with respect to that individual's well-being: the individual is better off if a higher-ranked state of affairs obtains than she would be if a lower-ranked state of affairs obtained. An individual's preference ranking determines facts about his or her own well-being, but not about anyone else's. Agnes's preferences, for example, determine the ranking of states of affairs which is relevant to considerations of her well-being; Brandon's preferences determine a ranking of states of affairs which is relevant to considerations of his well-being. But neither Agnes's preference ordering, nor Brandon's, nor the pair of them taken together obviously provide the resources for comparisons between Agnes's and Brandon's well-being. The fact that Agnes is better off in one state than she would be in another (together with similar facts for Brandon) is not enough to determine whether Agnes is better off in that state of affairs than Brandon.

One popular line of response to this problem invokes *extended preferences*. According to this response, people do not just have 'ordinary preferences', preferences over states of affairs; they also have 'extended preferences', over more fine-grained objects, called 'extended alternatives'. For example, Agnes may prefer *being Agnes while Agnes eats meat* to *being Brandon while Brandon eats meat*. 'Extended preferences' such as this one—between being Agnes in one situation and being Brandon in another—have the right structure to give rise to interpersonal comparisons of well-being: even a single such ordering already compares possible predicaments of distinct individuals. So the preference-satisfaction theorist may hope to invoke such extended preferences to make sense of interpersonal comparisons of well-being. From the perspective of proponents of extended preferences, interpersonal well-being comparisons only seemed to pose a problem for the preference-satisfaction theory because we mistakenly took all preferences to be ordinary preferences, preferences over states of affairs. Once we take account of extended preferences, over the more fine-grained 'extended alternatives', we see that individuals' preferences have the right form to generate interpersonal comparisons after all.

The appeal to extended preferences may seem to be a step in the right direction, but even if it is, it is only a first step.[1] Assuming that one does not wish to retreat to expressivism or subjectivism (and we will assume, following proponents of the

[1] In fact, the extended preferences program faces a number of challenges, and the program may not be a step in the right direction at all. Some important challenges concern (1) the precise nature and coherence of extended preferences, once we probe beyond the cursory sketch given above, and (2) the relationship between one individual's ordinary preferences, (say) Agnes's preference between eating meat and eating fish on the one hand, and, on the other, different individuals' extended preferences over affairs which concern that individual, (say) Brandon's preference between *being Agnes and eating meat* and *being Agnes and eating fish*. We discuss these issues in detail in a separate paper (Greaves and Lederman, forthcoming). In fact, we ourselves think that some of these other challenges are fatal to the extended preferences program. Solely by way of division of labour, the present paper focuses exclusively on the problem of aggregation.

extended preferences program, that we do not), it remains to be said how individuals' extended preference orderings will combine to determine facts about well-being.[2] If all individuals had the same extended preferences—as John Harsanyi and other early proponents of the extended preferences program claimed—then there would be no problem: the objective well-being ordering could simply be identified with the extended preference relation shared by all individuals. But increasingly authors in this area have recognised that individuals may not have the same extended preference ordering. And this means that the approach faces an important and pressing question. If individuals have different extended preferences, then there must be some way of producing an objective well-being ordering from individuals' diverse extended preferences: in other words, there must be a way of *aggregating* people's extended preference relations into a single ordering. But we do not yet know what this aggregation rule might be.

This paper studies how, given that extended preferences are not all the same, they might be aggregated to generate well-defined comparisons of well-being. We relate this problem to Arrow's impossibility theorem, and explore ways of avoiding analogues of Arrow's result in this context.[3]

Section 2 presents the problem of aggregation more precisely. Section 3 gives a first pass at why the problem is difficult, by showing how a recent proposal, due to Matthew Adler, leads to trouble. Section 4 recalls Arrow's theorem. Section 5 describes a variant on Arrow's theorem, based on assumptions that are weaker and more plausible than Arrow's in the context of the extended preferences program. Section 6 considers whether one might respond to this result by imposing a kind of domain restriction on the preferences which are aggregated. Section 7 considers whether one might respond by claiming that the 'ordering' of well-being levels fails the formal property of Quasi-Transitivity. We suggest that each of these responses is unattractive, for different reasons. Section 8 then suggests a different and more promising 'way out' for proponents of extended preferences: denying that the aggregation rule satisfies the formal condition known as 'Independence of Irrelevant Alternatives'. Section 9 is the conclusion.

2 Setup

We can state the problem of aggregation formally in a simple, abstract setting. There is a finite population of individuals $N = \{1, 2, \ldots, |N|\}$, and a finite set of

[2] In fact, it is not obvious that either individuals' extended preference relations, or the objective 'better off than' relation, need to have the formal properties of an *ordering*—they may, for example, fail to be transitive and/or complete. We will return to this later.

[3] This question, with the same motivation, has been raised by Adler (2014, 156; forthcoming, 26), who flags it as an important topic for future research. Voorhoeve (2014) discusses the incomparability problem that we focus on in Sect. 3. In Adler (2016), Adler presents an independent investigation of the problem of aggregation. We learned of this last paper after submitting the present paper but before it was published. There is significant overlap between the two papers: those interested in an alternative presentation of the material in sections 2–5 of the present paper, in particular, may wish to consult Adler's paper.

extended alternatives X, where $|X| \geqslant 3$.[4] In a moment we'll say something about the structure of these 'extended alternatives', but any proponent of extended preferences will believe individuals' extended preference relations are binary relations over some alternatives. We represent binary relations as sets of ordered pairs; a binary relation over X is thus a subset of $X \times X$, an element of $\mathcal{P}(X \times X)$; we henceforth denote the latter as $\mathcal{R}$. A *profile* of extended preferences is a specification of each individual's extended preferences, that is, a function from N into $\mathcal{R}$, an element of $\mathcal{R}^N$. We use R (plain font) as a variable over preference profiles, elements of $\mathcal{R}^N$; when a value for R is clear from the context we use R_i to refer to the ith component of R, the preference relation assigned to the individual i.

The aim of the extended preferences program is to define an *aggregation rule* $f : D \rightarrow \mathcal{R}$, where $D \subseteq \mathcal{R}^N$. Such an aggregation rule takes in profiles of extended preferences ($R \in D$), and outputs a relation over extended alternatives comparing them in terms of objective well-being ($f(R) \in \mathcal{R}$). To distinguish this kind of aggregation rule from others we will consider later, we often call it a *relation aggregation rule* (RAR). The symbols R_i and $f(R)$ stand for the 'weak' relations, 'at least as preferred by i' on the one hand, and 'at least as well-off as' on the other. When R is clear from context, we will use P_i and $f^P(R)$ to represent the 'strict', i.e. asymmetric, parts of these weak relations: thus xP_iy just in case xR_iy and $\neg yR_ix$, and similarly for $f^P(R)$.

That, then, is the formal abstract setting for the problem of aggregation. The problem itself arises from constraints which are motivated by features of the project of producing interpersonal well-being comparisons from extended preferences. Before we move on to these constraints, it will be helpful to have in place a slightly more concrete framework than the one we've presented to this point. We ourselves find this framework more conceptually illuminating than the completely abstract one, although as far as we're aware, nothing important in what follows will hinge on the details of these more concrete structures.

[4] For simplicity, we assume throughout the paper that both N and X are finite. As far as we know, nothing essential turns on the finiteness of the set of alternatives. But assuming that the population is finite is not entirely idle. As is well-known, Arrow's theorem in its original form does not hold for infinite populations (Fishburn 1970); analogues of the 'oligarchy theorems' on which our Spinelessness Theorem is based can similarly fail in the setting of an infinite population. Essentially, the problem is this: Arrow's conditions only imply that the set of 'decisive' groups (for a precise definition, see the Appendix) forms an ultrafilter in the powerset algebra of N, while the conditions of the oligarchy theorems only imply that the set of decisive groups forms a filter in this algebra. Since in the powerset algebra based on an infinite set there can be non-principal ultrafilters (and thus filters with infinitely descending chains under subset), Arrow's theorem no longer implies the existence of an individual dictator, and the oligarchy theorems no longer imply the existence of an oligarchy. But analogous, equally troubling results can still be proven in the infinite setting: for example, if there is a well-behaved, σ-additive measure on the infinite population, it can be shown that under conditions analogous to Arrow's, for any positive ϵ, no matter how small, then for some $\delta < \epsilon$, there will be a group of measure δ whose unanimous preferences are sufficient to determine the overall ordering (Kirman and Sondermann 1972). (An analogous modification of the oligarchy theorems can also be proven (see, e.g. Weymark 1984, Section 4).) The existence of such 'invisible dictators' is just as problematic as the dictator of Arrow's original theorem (and *mutatis mutandis* for the oligarchy theorems). In short: while stating related results for infinite populations requires technical machinery we won't introduce here, related conceptual points could be made in the infinite setting as well, so that the main line of argument does not depend on our simplifying assumption that the population is finite.

Let a *choice-situation* be a structure $\langle W, N, E \rangle$, where W is a nonempty set of possible worlds and N is a finite set of individuals. We identify the set of extended alternatives X with the set of centred worlds, obtained by taking the Cartesian product of the worlds and individuals: $X = W \times N$. These centred worlds specify not just what the world is like, but which individual is the 'centre' of the world. The property of being i in world w is associated with the centred world (w,i); if $i \neq j$, then this property differs from the property of being j in w, which is associated instead with (w,j). The final component of the structure, $E : N \rightarrow (\mathcal{R}^W)$, assigns each individual a function from worlds to preference relations over extended alternatives. In the general case, E might assign different relations to an individual at different elements of W. But for the purposes of this paper, we assume for simplicity that individuals' preferences do not vary across the worlds we are considering, so that E may be thought of as constant on W; in symbols, $E : N \rightarrow \mathcal{R}$. One can see that this class of more concrete models is contained in the class of more abstract models introduced above, by letting each R_i in the earlier formalism be given by the value of $E(i)$ in this more concrete one.

For the moment, we impose no constraints on which binary relations may count as preference-relations—that is, which relations may be inputs to this function. We also impose no constraints as yet on the output relation. In particular, nothing in the formal framework itself even requires that either the input (preference) relations or the output (well-being) relation be *orderings*. This gives our formal framework the expressive power to countenance, in particular, failures of transitivity and/or completeness in the input and/or output relations (as we shall do at some points in the paper), or alternatively to impose those constraints as additional substantive axioms (as we shall do at other points).

Some of these additional constraints are of importance to comparisons of well-being. In particular, as is well known to decision theorists, if one takes preferences to be defined over *lotteries* as well as over outcomes themselves, and if in addition various constraints (including, but not restricted to, transitivity and completeness) are imposed on these preferences over lotteries, then the preferences in question can be represented by utility functions on outcomes, unique up to positive affine transformation. On some views of the representation theorems which describe this relationship, the resulting utility functions determine unit comparisons. This machinery is most familiar in the context of preferences over ordinary alternatives and lotteries thereon (where, on some views of the representation theorems, the result is a well-defined notion of *intra*personal unit comparison). But the structural points apply equally in the context of *extended* preferences. If extended preferences are defined not only over extended alternatives themselves, but also over lotteries on extended alternatives ('extended lotteries'), and if in addition the individuals' preferences on extended lotteries satisfy the constraints in question (the axioms of decision theory), then each individual's extended preferences can be represented by a utility function on extended alternatives, again unique up to positive affine transformation. Of course, in general individuals have different extended preference orderings of these lotteries, and hence different utility functions on extended alternatives; thus an aggregation rule is still needed. But if the same constraints

(axioms of decision theory) are also imposed on the output of the aggregation rule, then that output relation too can be represented by an 'overall' utility function on extended alternatives, again unique up to positive affine transformation; and one could take this output utility function to determine (intra- and) interpersonal *unit* comparisons, in addition to the (intra- and) interpersonal *level* comparisons that one has as soon as the output relation is so much as an ordering (of extended alternatives).

For most of the rest of the paper, we will focus solely on interpersonal *level* comparisons. As we will see, this simple case will already be enough to impose tight constraints on the aggregation rules available to the extended preferences program. Considering unit comparisons would add additional structural constraints, and accordingly would make things even harder for the extended preferences theorist.[5] But some arguments later in the paper will rely on the possibility of moving from preference-orderings to utility functions, and that is why we have mentioned the relationship between preferences and utilities here.

3 The problem of spinelessness

So far, we have stated what kind of function an aggregation rule is, but we have not said why defining a function of this kind poses a problem for the extended preferences program. In this section we will introduce the problem, by considering a particular aggregation rule, the Strong Pareto Rule, which has recently been advocated by Adler (2012, 53) as a rule for aggregating extended preferences. The rule is defined as follows:

Strong Pareto Rule: For all $R \in \mathcal{R}^N$, and all $x, y \in X$, $xf(R)y$ if and only if for all $i \in N$, xR_iy.

[5] We note in passing that Harsanyi's famed 'aggregation theorem' (Harsanyi 1955) describes one further structural constraint which emerges if we consider unit comparisons. Informally, the theorem says that if (1) each agent's preferences are representable by a von Neumann–Morgenstern (vNM) utility function, (2) the output ordering is also representable by a vNM utility function, and (3) the output ordering satisfies an Ex Ante Strong Pareto condition, then the output ordering will be representable by a weighted sum of the individual vNM utility functions. Both the Strong Pareto condition and the claim that well-being should be representable by a vNM utility function are extremely plausible in the context of the extended preferences program. So Harsanyi's theorem shows that any acceptable aggregation rule for extended preferences must have a particular functional form: it must be representable by a vector of weights on those individual utilities. This 'single-profile' version of Harsanyi's theorem does not, as far as we are aware, have any implausible consequences; it simply exhibits a convenient way of expressing the family of functions to which the aggregation rule used in the extended preferences program must belong. 'Multi-profile' extensions of Harsanyi's theorem (e.g. Mongin 1994), by contrast, appear more problematic for the extended preferences program. But they rely on a condition similar to Independence of Irrelevant Alternatives (IIA), a condition which, we will argue later, the theorist of extended preferences must reject even if she is to make sense of level-comparisons. More could be said here, but we won't consider such problems further in the sequel, since given the solution we recommend they don't pose a challenge to the theorist of extended preferences distinct from the one we will develop in Sects. 4–8.

This rule states that one extended alternative is (weakly) better-for-the-individual than another if and only if *all* individuals' extended preferences rank the first (weakly) above the second.

Before we state our problem, note that on the standard understanding of the extended preferences program, as on the standard understanding of the preference-satisfaction theory of well-being more generally, only *rational* preferences are taken to be relevant to determining well-being facts. Thus, in the case of extended preferences in particular: it may be that some actual individuals have irrational extended preferences, but if so we are to assume that the input to the aggregation rule consists only of idealised versions of these preferences. For the remainder of the paper, when we speak about preferences which are inputs to the aggregation rule, we will be assuming these are rational preferences.

Now we turn to the problem. Given diversity in individuals' preferences, the Strong Pareto Rule leads to massive incomparability in well-being. To see this, suppose first that the profile of extended-preference relations to be aggregated included all rationally permissible extended-preference relations; we will call this the 'possibilist' version of the extended preferences program.[6] It is natural to

[6] Adler, for instance, sometimes suggests that the input to the aggregation rule should include all the extended preferences that any (actual) individual *could* have, or could have had, *at any time* (2012, 226–227). This is presumably extensionally equivalent to including all rationally permissible extended preferences. The resulting 'possibilist' version of the extended-preferences program would probably have to deviate from the formal framework as we have sketched it so far. In the first instance, it would have no special place for an assignment of preference relations *to individuals*, since individuals could have different relations in different possibilities. And if we did try to maintain some place for such an assignment in the possibilist setting, we would face problems arising from considerations of cardinality, which are likely to prevent there from being any surjective function from the set of individuals to the set of *all* rationally permissible extended-preference relations. To see this, recall, in particular, that we can identify the set of extended alternatives with the product $N \times W$, where N is the set of individuals—so the cardinality of the set of extended alternatives is at least as great as the set of individuals—and that extended preference relations are binary relations on this set of extended alternatives; since binary relations are elements of the powerset of this product, Cantor's theorem shows that there can be no surjection from the set of individuals onto the set of binary relations over extended alternatives. If rationally permissible preferences have the same cardinality as the whole powerset, then there can also be no surjection onto the set of rationally permissible preferences. In fact, a similar line of thought might be thought to show further that there can be no set of 'all rational extended preference relations' at all, and thus perhaps that this version of extended preference theory is itself incoherent: if (1) the specification of every possible world $w \in W$ includes a specification of which extended preference relations are held by which individuals, (2) there is one extended alternative corresponding to each element of the product $W \times N$, (3) there are more rationally permissible extended preference relations than extended alternatives (as is presumably the case), and (4) for every rationally permissible extended preference relation r and every individual i, there is some possible world in which i holds r, then the set of rationally permissible extended preference relations would have to have greater cardinality than itself, which is obviously impossible. However, this is not a special problem with (extended) preferences: a related argument can be used to show that if the objects of belief are sets of possible worlds, and belief states are represented by one set of possible worlds, then there is no set of all possible belief-states (Kaplan discovered this problem in the late 1970s, but it was not published until Kaplan (1995); a related puzzle is presented by Kripke (2011), who also discusses some of the history. A similar argument was independently discovered by Brandenburger (2003).) Since this mathematical fact presumably does not show that there is no interesting notion of rational belief, the corresponding argument does not show that there is no interesting notion of rational preference (extended or otherwise). In any event our argument will not rely on the assumption that one can make sense of the set of all rational preferences; see the next note.

suppose in the extended preferences setting that the only constraints on rational preferences are 'purely structural' ones: for example, for any rationally permissible extended-preference relation $R \subseteq X \times X$, the precisely 'reversed' extended-preference relation $R^{-1} \subseteq X \times X$ (such that for all alternatives $x, y \in X, xRy$ iff $yR^{-1}x$) is also rationally permissible.[7] It follows that if there is any rationally permissible extended-preference relation which ranks x strictly above y, there is another one which ranks y strictly above x. But this means that if any rational preference strictly prefers x over y, then x and y will be incomparable in the output ordering generated by the Strong Pareto Rule: since it is not the case that for every $i \in N\, xR_iy$ and also not the case that for every $i \in N\, yR_ix$, it follows that it is not the case that $xf(R)y$ and also not the case that $yf(R)x$.

An obvious and natural response is for the preference-satisfaction theorist to retreat and say instead that the extended preference relations to be aggregated include, not all rationally permissible extended-preference relations, but only the extended preferences that are the rational version of preferences which are actually possessed by some individual. We will call this the 'actualist' extended preferences theory, in contrast to the 'possibilist' theory of the previous paragraph. An actualist approach is anyway much more in keeping with the spirit of the ordinary preference-satisfaction theory of well-being, according to which the facts about which possible worlds are better or worse for Jane depend in some sense on Jane's actual preferences (or a suitable idealisation thereof): no (ordinary) preference-satisfaction theorist takes those facts to be determined purely by questions of which preference relations it *could have been rationally permissible* for Jane to have. And the simple argument of the preceding paragraph will not affect the actualist extended preferences theory: once we consider actual preferences, there is no longer any reason to think that for *every* extended-preference relation exhibited by some member of the constituency, its reversal will also be exhibited by some member of the constituency.

A moment's reflection, however, shows that the situation is unlikely to be much better in this (actualist) case. For any extended alternatives x, y, the output of the Strong Pareto Rule refrains from ranking x as being even *weakly* better than y whenever there is *any* person whose rational, fully informed preferences strictly prefer y to x. It only takes one person to have rational, fully informed preferences which regard education as a bane, for instance, for the Strong Pareto Rule to deliver the verdict that a life with greater education is neither better, nor even equally as good as, a life that involves lesser education but in which other relevant things are equal. Similarly for material consumption, hedonic pleasure, achievement, health and so forth. If the individuals whose extended preferences are aggregated are, say, all the inhabitants of any medium-sized country, and if the idealisation which produces rational preferences invokes only structural constraints on preferences, then it is overwhelmingly plausible that for almost any pair of extended alternatives,

[7] Note that this closure condition on its own does not generate any of the cardinality difficulties mentioned in the previous note. Thus one can equally well state the 'possibilist' position by replacing 'all rationally permissible preferences' with 'a very rich set of rational preferences, which includes non-actual ones'; our argument turns only on this set being closed under inverses.

there is some pair of individuals whose preferences disagree on them. Thus, according to the Strong Pareto Rule, again, almost every pair of extended alternatives is incomparable in terms of well-being.

Allowing such massive incomparability, though, amounts to denying the data with which we started. For the extended preferences program to have the implication that essentially no two individuals' well-being is comparable is for it to end in failure.

The Strong Pareto Rule leads to this catastrophic result for a simple reason. Say that individual $i \in N$ *has a veto* over alternatives $x, y \in X$ under some aggregation rule f just in case for all preference profiles R in the domain of f, if xP_iy then it is not the case that $yf(R)x$. An aggregation rule is *Spineless* on some subset of alternatives $Z \subseteq X$ if and only if every individual $i \in N$ has a veto on every pair of alternatives $x, y \in Z$. Any rule which is Spineless on the set of all alternatives will deliver the result that if there is even relatively modest variability in the preferences of the constituency, there will be massive incomparability in the overall well-being ordering. The problem with the Strong Pareto Rule is that it is Spineless on the set of all extended alternatives.

A natural response to the problem of massive incomparability is to blame the Strong Pareto Rule, and seek an alternative rule which is not Spineless (either on the set of all extended alternatives, or on any worrying large subset thereof). But, as we will show, this is more easily said than done. In the next section, we recall Arrow's theorem, which shows that any aggregation rule satisfying certain conditions will have a dictator: a single individual whose preference relation trumps all others' in deciding facts about well-being. Although this result is powerful, the assumptions used in it are plausibly not applicable to the aggregation of extended preferences. But in fact one can show that any aggregation rule which satisfies much weaker conditions will be Spineless, even if it does not have a dictator. The weaker assumptions used in this result are much more compelling than the assumptions of Arrow's original theorem in the setting of extended preferences. The theorem thus presents a serious challenge to the extended preferences program (Sect. 5).

4 Arrow's theorem

Arrow's result can be formulated in the setting introduced in Sect. 2. In addition to those definitions, recall that a binary relation $r \subseteq X \times X$ is an *ordering* iff it is reflexive, transitive and complete; we will denote the set of orderings by $\mathcal{O}$. Given a preference profile $R \in (\mathcal{P}(X \times X))^N$, and a subset $Y \subseteq X$, we write $R|_Y$ to denote the restriction of R to Y, that is: $\langle \{\langle x, y\rangle \in R_i : x, y \in Y\}\rangle_{i \in N}$.

For Arrow's Theorem, we impose axioms to the effect that both the relations in the input profile and the output relation have the formal properties of orderings. The complete list of axioms (considered as constraints on an aggregation rule f with domain D) is as follows:

UD (Unrestricted Domain): $D = \mathcal{O}^N$.

RF (Reflexivity): $\forall R \in D, f(R)$ is reflexive.

T (Transitivity): $\forall R \in D, f(R)$ is transitive.

C (Completeness): $\forall R \in D, f(R)$ is complete.[8]

WP (Weak Pareto): $\forall R \in D, \forall x, y \in X, ((\forall i \in N\ xP_i y) \rightarrow xf^P(R)y)$.

IIA (Independence of Irrelevant Alternatives): $\forall x, y \in X, \forall R, R' \in D, \left(\left(R|_{\{x,y\}} = R'|_{\{x,y\}}\right) \rightarrow (xf(R)y \leftrightarrow xf(R')y)\right)$.

ND (Non-dictatorship): $\neg\exists i \in N, \forall R \in D, \forall x, y \in X, (xf(R)y \leftrightarrow xR_i y)$.

Arrow's celebrated result shows that these conditions cannot be jointly satisfied:

Theorem 1 (Arrow's impossibility theorem; Arrow 1963). *There is no RAR satisfying UD, RF, T, C, WP, IIA and ND.*

5 The Spinelessness theorem

Arrow's impossibility theorem has been most discussed in the context of aggregation of ordinary preferences: cases, for example, in which X is interpreted as a set of possible income distributions, candidates for the presidency, or something similar, and we seek a 'social choice' among these *un*centred alternatives on the basis of individuals' diverse *ordinary* preferences. But a theorem is a theorem, and cares not how we interpret it. If (as in the notation introduced in Sect. 2) we take X instead to be the set of *extended* alternatives then, *insofar as* the Arrow conditions are conditions of acceptability for an extended-preference aggregation rule, Arrow's theorem shows that there is no acceptable aggregation rule.

This immediately raises the question of the extent to which Arrow's conditions *are* conditions of acceptability for an extended-preference aggregation rule. We will not question the Non-Dictatorship condition, as that seems unassailable (given any remotely diverse domain). It is also difficult to see how any aggregation rule that violated the Weak Pareto condition would fit with the motivations of the extended preferences program. For the spirit of the extended preferences program requires betterness facts not merely to *supervene somehow* on individual preferences: it further requires betterness facts to *respect* individuals' preferences. And while this leaves open a nontrivial question about what the betterness facts are when individuals' preferences fail to coincide, proponents of the program tend to agree that the betterness facts should match individuals' unanimous judgments when such unanimity exists. The Reflexivity condition, too, is difficult to question, since it just

[8] Thus *RF*, *T* and *C* together are equivalent to the condition that $\forall R \in D, f(R) \in \mathcal{O}$. We separate the conditions here because in later sections we consider weakening or dropping some of these conditions independently of others.

follows from the fact that we are discussing 'weak' rather than 'strict' betterness relations.

Most of the remaining conditions, however, are inappropriate in the extended preferences context.

First, it might seem that the framework itself is inappropriate: it requires an output ordering to be determined only on the basis of individuals' input preference *relations* (which may be orderings), but we might well further have individuals' preferences over *lotteries over* extended alternatives. If so, then (via the usual decision theoretic machinery, as noted in Sect. 2) the input could consist of profiles of individuals' *utility functions* on extended alternatives, rather than merely orderings of extended alternatives. The question therefore arises of whether, even in the absence of any acceptable aggregation rule on relations, there might nonetheless be an acceptable rule that instead takes profiles of *utility functions* as its input.

Second, the requirement of Universal Domain is questionable: an aggregation rule for the purposes of the extended preferences program needs to be able to aggregate *rational* extended preferences, but it may well be that some orderings of X (that is, elements of $\mathcal{O}$) are such that it is rationally impermissible to hold the associated extended-preference ordering.[9]

Third, the Completeness requirement on the output relation is too strong. We complained, in the context of the Strong Pareto Rule, that *massive* incomparability is implausible, but it is highly plausible that for at least *some* pairs of extended alternatives, neither is better than the other, and nor are the two equally good.

These considerations do suffice for a response to Arrow's original theorem. But they do not allow for an escape from a closely related result. As we will now show, even if one weakens Arrow's conditions in all of the above ways simultaneously, one can still prove that any rule which satisfies much weaker assumptions will be Spineless.

5.1 Sen's lemma

In response to the first reply to Arrow's theorem, it is easily shown that, given Independence of Irrelevant Alternatives (IIA), *the limited kind of utility information* that is available in the extended preferences setting does not suffice to make available any essentially new aggregation rules. The first objection just described—that an aggregation rule on preference relations misses out on important information which is preserved in the utility function—thus cannot help to avoid an impossibility theorem, unless in addition IIA is jettisoned.

[9] Binary relations which are not elements of $\mathcal{O}$ may also be such that it is rationally permissible to hold the associated extended preference-ordering: for example, it is at the very least arguable that rational preferences need not be complete (in other words, the *inputs* to the aggregation rule need not be complete). But assuming that the domain is as stated in the condition *UD rather than some larger domain* if anything makes it easier to find an acceptable aggregation rule: insofar as we can argue that there is no acceptable aggregation rule for a domain $D \subseteq \mathcal{O}^N$, a fortiori there is no acceptable aggregation rule for a larger domain.

The basic point is that in this setting, we have only the utility information that is recoverable from individuals' preferences over lotteries (on extended alternatives). That information amounts to a positive affine family of utility functions for each individual *taken separately*. While we might *represent* individuals' preferences over extended lotteries using a particular profile of utility functions, our aggregation rule had better deliver the same ordering of extended alternatives for any of the other profiles that would equally well have represented the same extended-preference information.

Formally: an *extended utility function* is a function $u : X \rightarrow \mathbb{R}$. Let U be the set of all such utility functions; an element of U^N is an $|N|$-tuple of utility functions. A *utility aggregation rule (UAR)* is a function $\overline{f} : \overline{D} \rightarrow \mathcal{R}$, where $\overline{D} \subseteq U^N$.[10] We write u for a typical element of U^N, u_i for the ith component of u, and $u(x)$ for the vector of real numbers $\langle u_1(x), \ldots u_{|N|}(x)\rangle$. The formal expression of the requirement that the output of the utility aggregation rule not depend on arbitrary aspects of our choice of representation is:

CNC (Cardinal non-comparability) Let $\pi_{CNC} : U^N \rightarrow U^N$ be any permutation of U^N of the form $u_i \mapsto a_i u_i + b_i$ where, for each $i \in N$, $a_i > 0$ and $b_i \in \mathbb{R}$. Then $\overline{f}(u) = \overline{f}(\pi_{CNC} u)$.

Restrictions analogous to those stated above in the presentation of Arrow's theorem can be imposed on UARs, as well as on RARs, but need to be restated slightly in order to apply in the UAR framework. Particularly important for our immediate purposes is Independence of Irrelevant Alternatives, which, in the UAR framework, can only be:

IIA* (Independence of irrelevant alternatives, utility version)
$\forall x, y \in X, \forall u, u' \in \overline{D}, ((u(x) = u'(x) \wedge u(y) = u'(y)) \rightarrow (xf(u)y \leftrightarrow xf(u')y))$.

It is easy to show that, provided that conditions CNC and IIA* are satisfied, a move from preference profiles to utility profiles does not make available any new aggregation rules, in the following precise sense:

Definition 1. A UAR $\overline{f}$ *reduces to the RAR* f iff

$$\forall x, y \in X, \forall u \in U^N, \left(x\overline{f}(u)y \leftrightarrow xf(R)y\right),$$

where R is the preference profile that is ordinally represented by the utility profile u (that is, for all $i \in N$ and all $x, y \in X$, xR_iy if and only if $u_i(x) \geq u_i(y)$).

[10] In the utility-function context, it is arguably natural, if the input to an aggregation rule is a profile of *utility functions* rather than merely orderings, for the output also to be a utility function (or a positive affine family of such functions) rather than merely an ordering. Any such output utility function, however, certainly induces an output ordering; thus an impossibility theorem formulated in terms of 'utility aggregation rules' in our sense (where the output relation is merely required to be some relation or other) applies a fortiori to these richer objects: we lose no generality in considering only UARs in our sense.

Lemma 1 (Sen 1970). *Let $\overline{f}$ be a UAR satisfying IIA* and CNC. Then, there exists an RAR f such that $\overline{f}$ reduces to f.*

Proof. This is part of the proof of Sen's Theorem 8*2. □

Thus the first objection to Arrow's framework is inconsequential: even if we work in the standard framework of *relation* aggregation rules, we will not risk missing any otherwise-available aggregation rules, if IIA is imposed.

5.2 The impossibility theorem

What of the two remaining responses mentioned above: denying Universal Domain and denying Completeness? We next show that even weakening Universal Domain considerably and dropping Completeness altogether, one can still prove that any rule which satisfies a fairly weak set of conditions will be Spineless. Although Arrow's original result no longer applies—we cannot show that there is a dictator—we argued in Sect. 3 that this property of Spinelessness already leads to an unacceptable degree of incomparability. If the assumptions of the theorem are true, the result is damning for the extended preferences program.

5.2.1 Dropping Completeness

The most striking feature of the Spinelessness Theorem is that we will need no condition in place of Completeness. Completeness will thus be conspicuous by its absence.

5.2.2 Replacing Universal Domain with Sufficient Diversity

The situation with Universal Domain is slightly different. While we can relax Universal Domain substantially, we will still require a version of the condition. But the new condition is fairly weak: it only requires a comparatively minimal degree of diversity among possible extended preferences. In particular, we will work with a set $Z \subseteq X$ of extended alternatives with respect to which the following constraint is true of the domain D:

SD (Sufficient Diversity): For any quadruple of distinct extended alternatives $x, y, u, v \in Z$ and any $|N|$-tuple $\mathbf{r}$ of transitive, reflexive and complete relations on$\{x, y, u, v\}$, there exists a profile $R \in D$ whose restriction to $\{x, y, u, v\}$ is $\mathbf{r}$.

As suggested above, we reject the Universal Domain condition itself, on the grounds that the domain need only contain profiles of *rational* preference relations. But, we claim, even this smaller domain (whatever exactly its boundaries are) will still be large enough for Sufficient Diversity to hold of some subset Z of extended alternatives *that also has the following property*: Z is large enough that if the aggregation rule were Spineless on Z, that would result in a problematic amount of incomparability.

5.2.3 Quasi-transitivity

We will also not need to assume the full transitivity condition. Recall that given a profile R, $f^P(R)$ is the asymmetric portion of $f(R)$. Then for this result we need only

QT (Quasi-transitivity) For all $R \in D$, $f^P(R)$ is transitive.[11]

5.2.4 Anonymity

The Anonymity principle is an innocuous strengthening of Arrow's Non-Dictatorship condition. Its purpose is to capture the idea that in moving from a profile of individuals' preference relations to a single output preference relation, the aggregation rule should 'treat all individuals equally'. It should not arbitrarily privilege some individuals over others: by allowing one individual to act as dictator, by entirely discounting the preferences of some individuals but not others, or by treating the 'bare identities' of individuals (as opposed to: features of their preferences) as in any other way relevant. It seems clear that this more general idea is just as compelling as the Non-Dictatorship requirement itself.

In the extended preferences context, however, we need to take care with the formulation of the Anonymity principle. Let π be a permutation of the set N of individuals. *One* natural action of π on the space of preference profiles $\mathcal{R}^N$ is given by simply by permuting the preference relations R_i: thus one might define $\pi(R) = (R_{\pi^{-1}(1)}, \ldots, R_{\pi^{-1}(|N|)})$.[12] Further, given any action of such permutations on $\mathcal{R}^N$, *one* very natural corresponding notion of what it is for an aggregation rule f to be invariant under π is for it to be the case that for all $R \in \mathcal{R}^N, f(R) = f(\pi(R))$. It might seem natural, then, to impose an Anonymity condition requiring the aggregation rule to be invariant under permutations of individuals in this sense. And

[11] In our view, the distinction between Transitivity and Quasi-Transitivity is mainly of technical interest: we are not aware of any plausible reasons for thinking that rational preferences need not be transitive, but (at the same time) must be quasi-transitive. (Here is a purported reason that we regard as *im*plausible. Consider three alternatives x, y, z that are arranged in close succession along some continuum: for example, shades of red, or amounts of sugar. It is sometimes claimed that such alternatives can have the property that both the difference between x and y and the difference between y and z are imperceptible, while (however) the difference between x and z is perceptible; further, that this might justify being indifferent between x and y, and being indifferent between y and z, while having a strict preference for x over z. This pattern of preferences satisfies Quasi-Transitivity, but not full Transitivity, since, here, strict preferences but not indifferences are transitive. We each reject this argument, but for different reasons. One of us thinks the argument goes wrong in its first step: there can be no such pattern of 'imperceptible' differences in the sense of 'perceptible' relevant to well-being. One of us thinks it goes wrong in its second step: granting the suggested pattern of perceptibility/imperceptibility exists, it does not justify this pattern of indifference and strict preference.) But in any case, even granting that the distinction is of more than technical interest, the main observation for present purposes is that our result would apply to quasi-transitive preferences as well.

[12] Why not $\pi(R) = (R_{\pi(1)}, \ldots, R_{\pi(|N|)})$? Because we seek throughout, for notational convenience, to be defining left- rather than right-actions. That is, the action of permutations π on profiles R must be such that $(\pi_2\pi_1)(R) = \pi_2(\pi_1(R))$ (rather than that $(\pi_2\pi_1)(R) = \pi_1(\pi_2(R))$). In the present case (and assuming the action of permutations π on N is itself a left-action), this condition is met by the definition given in the main text, but not by the alternative.

indeed this would look very much like the Anonymity principle that is often appealed to in social choice theory, in the context of ordinary rather than extended preferences (compare, for example, Weymark 1984, 238).

This Anonymity principle, however, is inappropriate in the context of extended preferences. We can see this by thinking in terms of our concrete model of extended preferences, in which the set of extended alternatives is constructed as a product $X = W \times N$, where W is a set of (uncentred) possible worlds. For any extended alternative $x = (w, i)$, there is some particular individual i who is the 'centre' of that extended alternative. But in that case, the aggregation rule might very well assign some significance to the special connection that obtains between each individual and the extended alternatives of which she herself is the centre. It seems eminently reasonable, for instance, for Annie's extended preferences to count more heavily in the question of intrapersonal comparisons among Annie-centred alternatives than Ben's extended preferences do. And then the naive Anonymity principle stated above will fail: the aggregation rule might very well, for example, deliver a different output when *Annie*'s extended preferences rank $(w_1, Annie)$ above $(w_2, Annie)$ than it does when *Ben's* extended preferences rank $(w_1, Annie)$ above $(w_2, Annie)$ (and *mutatis mutandis*). This would not amount to the aggregation rule's 'arbitrarily assigning different roles to different individuals' in any objectionable sense.

However, it *would* go against the spirit of Anonymity if the aggregation rule assigned significance to the special connection between Annie's extended preferences and Annie-centred extended alternatives *without also assigning the same significance* to the special connection between Ben's extended preferences and *Ben*-centred extended alternatives. This observation suggests an alternative Anonymity principle that, unlike the naive one stated above, does seem defensible in the context of extended preferences. To capture this idea, we allow a permutation π of N to act not only on the space N of individuals, but also on the space X of extended alternatives. In terms of our concrete model, in which an extended alternative is identified with a pair (w, i), this comes about because there is a natural action of π on the space W of possible worlds, namely the action that permutes the identities of individuals (relative to the qualitative facts). The action of π on X is then given by: $\forall (w, i) \in X, \pi(w, i) = (\pi(w), \pi(i))$.[13] And given actions of π on both N and X, there are corresponding natural ways to define actions of π on the space $\mathcal{R}$ of binary relations on X, on the space $\mathcal{R}^N$ of N-tuples of such relations, and on the space of aggregation rules f:

Action of π on relations $r \in \mathcal{R}$: $\forall x, y \in X, x(\pi(r))y \leftrightarrow (\pi^{-1}(x))r(\pi^{-1}(y))$;
Action of π on profiles $R \in \mathcal{R}^N$: $\pi(R) = \left(\pi\left(R_{\pi^{-1}(1)}\right), \ldots, \pi\left(R_{\pi^{-1}(|N|)}\right)\right)$;
Action of π on aggregation rules f: $\forall R \in \pi D, (\pi f)(R) = \pi(f(\pi^{-1}R))$.

[13] As stated this definition does depend on the details of the 'more concrete' framework introduced in Sect. 2. But all that we require here is an idea which can be stated independently of that framework, namely: that every permutation of individuals induces some corresponding permutation on the (abstract) set of alternatives. This induced permutation can then be used for the definitions which follow.

We require, then, that the aggregation rule be 'invariant under permutations of individuals' in the (different) sense that first allowing π to act on the input profile R and then applying the aggregation rule yields the same result as first applying the aggregation rule to R and then allowing π to act on the the resulting binary relation:

A (Anonymity): For all permutations π of N, $f = \pi f$.[14]

This of course entails that the domain D of f is invariant under π; it also entails that X is invariant under π. This latter assumption is natural for X itself, but it is perhaps more substantial when we consider a subset Z of X, as we will below.

5.2.5 The Spinelessness Theorem

Our objection to the Strong Pareto Rule was that it refused to deliver a strict betterness relation between any two alternatives whenever even a single individual had the opposite strict preference; this 'Spinelessness' was the feature of that rule which led to excessive incomparability. Recall our definitions:

Definition 2 ***(veto).*** Individual $i \in N$ *has a veto for* the pair $x, y \in X$ iff for all profiles $R \in D$, if xP_iy then not $yf^P(R)x$.

Definition 3 *(Spinelessness).* A RAR f is *Spineless with respect to* $Z \subseteq X$ iff every individual has a veto for every pair of alternatives in Z.

To state the Spinelessness Theorem, let $\mathcal{R}_{rat} \subseteq \mathcal{R}$ be the set of rational preference relations, whichever these are. (The following theorem does not itself require any assumptions about the extent of $\mathcal{R}_{rat}$: it is neutral, for instance, over whether or not rationality requires preferences to be transitive and/or complete.) Note that given that a RAR f with domain D satisfies IIA, it always has a well-defined restriction $f|_{D'}$ to an arbitrary subdomain $D' \subseteq D$. We then have

Theorem 2 (Spinelessness Theorem)**.** *Let f be a RAR with domain $D = \mathcal{R}_{rat}^N$. Suppose that f satisfies RF, WP, QT and IIA. Let $Z \subseteq X$ be any set of extended alternatives with respect to which $(\mathcal{R}_{rat})^N$ satisfies SD, and such that $f|_{(\mathcal{R}_{rat}|_Z)^N}$ satisfies Anonymity. Then f is Spineless with respect to Z.*

The proof is in the Appendix.[15]

We suggested above that there will always be worryingly large subsets $Z \subseteq X$ with respect to which the Sufficient Diversity condition holds. It is also clear that some of these will have the further property that Anonymity is a reasonable

[14] The intuitive explanation just given suggests what might seem to be a different formal condition: for all R and π, $\pi(f(R)) = f(\pi(R))$. But it's easy to show that these are equivalent. Since the conditions hold for all permutations, they must hold for π^{-1}. But $\pi^{-1}(f(R)) = f(\pi^{-1}(R))$ iff $\pi(\pi^{-1}(f(R))) = \pi(f(\pi^{-1}R))$. Moreover, $\pi(\pi^{-1}(f(R))) = f(R)$ and, by definition, $\pi(f(\pi^{-1}(R))) = (\pi(f))(R)$. Since this holds for all R, the intuitive condition holds for all π iff the mathematically simpler one in the main text does.

[15] Mathematically, there is nothing very original in this theorem: the key aspects of the proof are contained in the work of Sen (1970) and Weymark (1984).

condition on $f\big|_{(\mathcal{R}_{rat}|_Z)^N}$ (and not merely on f; that is, for any permutation π of N, the action of π on X leaves $Z \subseteq X$ invariant). For—recalling our justification of Anonymity in Sect. 5.2—this will be the case whenever Z has the property that for all $(w, i) \in Z$ and all $j \in N$, Z also contains some j-centred extended alternative (w', j) such that j's situation in w' is just like i's situation in w.

But if an aggregation rule is Spineless on a sufficiently large set of alternatives Z, it will yield an unacceptable degree of incomparability. The extended preferences theorist must escape the Spinelessness Theorem. But how?

The next three sections consider different motivations for denying assumptions used in the Spinelessness Theorem. We argue that the first two of these are not particularly promising; we recommend that the extended preferences theorist explore the third.

6 Rejecting Sufficient Diversity

We first consider the possibility that the problem could be met by restricting the domain, going beyond the denial of Universal Domain to deny also the weaker assumption that the Sufficient Diversity condition holds of any interestingly large sets $Z \subseteq X$ of extended alternatives. We take in turn the 'possibilist' and 'actualist' versions of the extended preferences program (in the senses discussed in Sect. 3).

The motivation for the Sufficient Diversity condition is dubious at best in the possibilist version of the extended preferences program: a version, that is, according to which the profile of preference relations to be aggregated includes all rationally permissible preference relations. This is because in that version, arguably the domain need only include *one* profile: the single profile that consists of one copy of each rationally permissible preference relation.[16] But as we have already noted, the 'possibilist' version of the extended preferences program is in tension with the ordinary preference-satisfaction theory: Jane's well-being is typically thought to depend only on (idealised versions of) the preferences she has, not on what preferences she might have. But in any event, if the set of rational preference relations is closed under inverses, then there is a more basic reason (independent of Arrow's theorem) for thinking that the resulting well-being comparisons must involve massive incomparability (or indifference) at least on different-centred alternatives. For it is hard to see, for any given different-centred extended alternatives $x, y \in X$ on which at least two individuals' preferences disagree, in virtue of what x could be strictly better than y or vice versa—unless the program appeals to substantive, non-preference-based considerations that again violate the spirit of preference-satisfaction theory. The prospects for rescuing the preference-

[16] More carefully: Whether there is in fact only one profile with this property (and Sufficient Diversity fails), or instead many (in such a way that Sufficient Diversity is met), depends on the apparently merely technical issue of whether the 'individuals' are in that setting individuated by their preference relation (so that there is exactly one profile to consider), or independently of that relation (so that permuting preference relations among individuals gives rise to a distinct profile).

satisfaction theory of well-being via a possibilist version of the extended preferences therefore seem dim.

Next we turn to the actualist version of the program—where the profile of preference relations to be aggregated includes only (the rational idealisations of) actual preferences. In this version of the program it is significantly more difficult to resist Sufficient Diversity. Note first that while it may (or may not) be the case that the rational versions of actual living individuals' preferences are diverse in the sense that different individuals *in fact* rank the extended alternatives in Z differently, this is not the type of diversity that is relevant to our Sufficient Diversity condition. That condition instead concerns the diversity of the *domain of the aggregation rule f*: a matter of variation of preference *profiles* across *possible worlds*, rather than variation of preference *relations* across *individuals*. And for any element $R \in \mathcal{R}^N_{rat}$, there is presumably some possible world in which individuals' idealised extended preferences are as given by R. Assuming that the domain of the aggregation rule does correspond to all possible profiles of individuals' idealised extended preferences, therefore, the Sufficient Diversity condition can be resisted for a given set of alternatives $Z \subseteq X$ only insofar as it can be argued that some orderings of Z can only arise in rationally impermissible preference orderings. But, barring an appeal to the kind of deeply 'substantive' notion of rationality that is inconsistent with the spirit of preference-satisfaction theory, this will still leave many problematically large subsets $Z \subseteq X$ for which the Sufficient Diversity condition does hold.

Might one, though, deny that the domain has to consist of the preference profiles that correspond to individuals' idealised preferences *across all possible worlds*? If, instead, the aggregation rule for determining objective well-being comparisons in w only needs to be defined for preference profiles that are instantiated in worlds close to w (or, in the limit, only the preference profile that is instantiated in w itself), then again the condition of Sufficient Diversity may not hold for any problematically large sets of extended alternatives, much as in the possibilist case.

This move, however, affords only a superficial appearance of a solution. For there still remains a question of what is the correct mapping from possible worlds w to objective well-being comparisons true in w, and this mapping from worlds to well-being comparisons corresponds to an aggregation rule (let us call it the 'master aggregation rule'). Since the domain of this master aggregation rule must be Sufficiently Diverse for problematically large sets of extended alternatives, the Spinelessness Theorem still poses a serious challenge in the search for an acceptable such rule. To put the point more formally: suppose, in concession to the previous paragraph, that we allow some sense in which each possible world w corresponds to an aggregation rule f_w, where, for each world w, the domain of f_w need not include many or perhaps any preference profiles other than the profile R^w that is instantiated at w itself. Nonetheless, we can still construct and discuss the mapping $w \mapsto f_w(R^w)$. Assuming only that $f_{w_1}(R^{w_1}) = f_{w_2}(R^{w_2})$ whenever $R^{w_1} = R^{w_2}$—which is surely required by the idea that well-being facts are determined by preferences—this allows us to construct a mapping $R^w \mapsto f_w(R^w)$. But the latter—the 'master' aggregation rule—just is the aggregation rule f that we

discussed in the previous paragraph but one, and for which we argued Sufficient Diversity holds on problematically large sets of extended alternatives.

Might a similar line of thought, though, provide resources for denying that some of the *other* axioms of the Spinelessness Theorem—besides Sufficient Diversity—need hold in the relevant sense? The thought here would be as follows: the axioms of that theorem (Weak Pareto, Independence of Irrelevant Alternatives, and so forth) are true *of the actual aggregation rule* $f_@$, but need not be true of non-actual aggregation rules f_w where $w \neq @$. Thus, in particular, on this view they need not hold of the 'master aggregation rule' f that we construct by 'gluing together' elements of all the more primitive aggregation rules f_w. If so, then the Spinelessness Theorem fails to present any problem: it fails to present any problem for the actual primitive aggregation rule $f_@$ because the domain of that rule is not Sufficiently Diverse with respect to any interestingly large set of extended alternatives, and it fails to present any problem for the master aggregation rule f because that aggregation rule need not satisfy the conditions of e.g. Reflexivity/Weak Pareto/IIA/Quasi-Transitivity.

But this move too is hopeless. Whatever reasons we have for thinking that the betterness relation on extended alternatives must be reflexive or quasi-transitive, or that it must respect unanimous preferences, are reasons to think that *necessarily* that must be the case, and whatever reasons we have for imposing IIA are reasons for imposing IIA *on the master aggregation rule f*. In fact, not all of these reasons are entirely compelling: we will shortly suggest, for instance, that the IIA condition, and perhaps also the Quasi-Transitivity condition, might very well be resisted. But these other reasons for resisting some of the axioms have nothing to do with postulating different aggregation rules for different possible worlds. The appeal to a reduced domain for 'the actual' aggregation rule gets the would-be defender of extended preferences nowhere.

7 Aggregation rules that violate Quasi-Transitivity

We have now argued for conditions RF, WP, SD and A, and against Spinelessness. Given Theorem 2, this leaves two possibilities for the extended preference theorist: she could seek an aggregation rule that violates Quasi-Transitivity, and/or she could seek an aggregation rule that violates Independence of Irrelevant Alternatives. The next two sections consider these remaining possibilities, in that order.

It is not beyond question that the weak betterness relation must be 'quasi-transitive': that is, that the *strict* betterness relation must be transitive. We share the near-unanimous view (but *pace* Temkin (1987, 2014) and Rachels (1998)) that the strict betterness relation cannot involve any 'cycles'. But arguably this is all that is required for the betterness relation to do useful work in normative theory. So long as the strict betterness relation is *acyclic*—that is, there are no sequences of alternatives $x_1, \ldots, x_n$ such that $x_1 \succ x_2 \succ \cdots \succ x_n \succ x_1$ (where $\succ$ stands for strict betterness)—it is arguable that it could still do the work it is needed for. (This is all that is required, for example, for the purpose of invulnerability to money pumps, or for guaranteeing that in any finite set of options, there always exists at least one such that no other available option is strictly better.)

In fact, impossibility theorems do also exist based on the Acyclicity condition in place of Quasi-Transitivity (e.g., Austen-Smith and Banks (2000, Theorem 2.5, p. 46), Brown (1973, Theorem 13, p. 18)). Those impossibility theorems, however, differ in a crucial respect from theorems that are based on a Transitivity or Quasi-Transitivity condition; in particular, they introduce a further condition of 'Neutrality', which can be informally stated as follows:

> N (Neutrality) For any permutation ρ, of the set of alternatives, first permuting the input preference profile by ρ and then applying the aggregation rule produces the same result as first aggregating and then permuting the output betterness relation by ρ.

At first sight, Neutrality might seem to be a natural expression of the idea that preference-satisfaction theory cannot discriminate ab initio between any pair of extended alternatives: that it must defer entirely to what individuals' extended preferences have to say about those alternatives. But in fact Neutrality may be positively inappropriate in the present context, for the same reason that our initial, 'naive' statement of an Anonymity condition in Sect. 5.2 proved inappropriate. For, as noted there, it is arguably reasonable for an aggregation rule to treat the relationship between Annie's extended preferences on the one hand, and extended alternatives that are centred on Annie, as special; yet an arbitrary permutation of extended alternatives will not preserve Annie-centredness, or even same-centredness for pairs of extended alternatives. Further, in this case—unlike the case of Anonymity that we discussed above—it is unclear that we will be able to find any relevantly similar variant of the Neutrality condition by adding a relevant permutation of individuals to the mix, because while every permutation of individuals corresponds naturally to some permutation of extended alternatives, the reverse is not true in general (it is true only for permutations of extended alternatives that preserve 'same-centredness'). Thus, there may be an aggregation rule that satisfies Acyclicity, violates Neutrality, satisfies the axioms of Theorem 2 except for Quasi-Transitivity, and is not Spineless with respect to any problematic subset of extended alternatives. This is another avenue that the extended preferences theorist could pursue, *insofar as* she is happy for the output ranking to violate the full Transitivity condition.

But while we concede that this approach is *possible*, it is unattractive. Would-be 'betterness relations' that satisfy Acyclicity but not Quasi-Transitivity are, in our view, strange indeed. We suspect that most would reject the preference-satisfaction theory of well-being before accepting such a surprising view about the structure of the betterness relation. So we regard this way of replying to the Spinelessness Theorem as unpromising. The next section turns to a final possible way out, which we think is much more promising.

8 Aggregation rules that violate Independence of Irrelevant Alternatives

While IIA perhaps has a superficial air of plausibility, we are not aware of any positive argument for it that is applicable in the setting of the extended preferences program.[17] It is formally very natural, and there are no clear arguments against it, but in the absence of a positive argument for the principle, we think that denying IIA is the natural route for the proponent of extended preferences to take. We therefore recommend this avenue of investigation to the extended-preference theorist.

The crucial question, then, is whether there in fact exist any aggregation rules that violate IIA, that thereby escape the Spinelessness problem, *and that are otherwise plausible in the context of extended preferences*. In this section, we will narrow down the options by pointing out some rules that strike us as *un*promising in the context of the extended preferences program. We will then indicate the direction in which we think more positive progress is most likely to be made.

Firstly: perhaps the best-known example of a relation aggregation rule that violates IIA is the *Borda rule*. The rule can be described as follows: given that the number of alternatives is finite, we assign a nonnegative integer $n(i,x)$ to each pair consisting of an individual i and an alternative x, such that for each individual i, the most-preferred alternative is assigned the highest integer $n(i,x) = |X|$, while the least-preferred alternative is assigned $n(i,x) = 1$. The overall Borda score for a given alternative is given by summing these scores across all individuals: $B(x) = \sum_{i \in N} n(i,x)$. The Borda rule then ranks one alternative above another just in case the first has a higher Borda score $B(x)$ than the second.

The problem with appealing to the Borda rule to aggregate extended preferences is that doing so threatens to undermine the motivation for appealing to extended preferences in the first place. The goal, recall, was to solve the problem of interpersonal comparisons within a preference-satisfaction theory of well-being. The appeal to extended preferences, however, is not the only possible way of doing this. An alternative approach, which we call *structuralism*, seeks to define interpersonal comparisons on the basis of structure that is already present in a profile of preference orderings, without expanding the objects of ordinary preferences. The problem with the Borda rule as a tool for the proponent of extended preferences is that it is unclear why one would regard the appeal to extended preferences, together with the Borda rule, as superior to invoking structuralism in the first place.

To illustrate this point, consider perhaps the most straightforward structuralist proposal (for the definition of interpersonal comparisons, in an ordinary-preference setting). Suppose that the number of ordinary alternatives (members of W in our more concrete model) is finite. Then, for each individual i and each ordinary alternative $x \in W$, there is an integer $n(i,x)$, representing the position of alternative x in i's preference ordering. The structuralist can then define interpersonal well-

[17] In voting theory, it has been argued that aggregation rules that violate *IIA* are open to manipulation. However, no concept of manipulability is applicable in the extended preferences context: our question concerns how the *facts* about individuals' extended preferences determine the facts about overall betterness, not how any choice should be based on individuals' *reports* of their own preferences.

being comparisons as follows. Interpersonal level comparisons: state of affairs x is as good for person i as state of affairs y is for person j iff $n(i,x) = n(j,y)$. Interpersonal unit comparisons: the ratio of the difference between x and y for i to the difference between v and w for j is given by $\frac{n(i,x)-n(i,y)}{n(j,v)-n(j,w)}$.[18]

[18] This particular proposal is problematic in a setting in which the data we start from is not merely a profile of ordinary preference *orderings* (one ordinary preference ordering for each individual), but rather a profile of ordinary *utility functions*. Any stipulation for fixing *inter*personal unit comparisons of course automatically induces a standard of *intra*personal unit comparisons: the intrapersonal ratio (difference between x and y for i)/(difference between s and t for i), for example, must be equal to the product of the two interpersonal ratios (difference between x and y for i)/(difference between v and w for j), (difference between v and w for j)/(difference between s and t for i). The problem is that the intrapersonal unit comparisons that are induced by the above analogue of the Borda rule will not in general be consistent with the pre-existing intrapersonal unit comparisons already given in the profile of utility functions, since the Borda rule pays no attention to any features of individuals' preferences or utilities that go beyond the induced ordinal ranking. The would-be structuralist therefore needs some other prescription, one that respects the existing cardinal information that is already present in individuals' (ordinary) utility functions.

There are various ways in which this can be done. The basic task is to select, from the positive affine family of utility functions that cardinally represent each individual's ordinary preferences, one privileged representative utility function; the profile of representative utility functions across individuals then well-defines a standard of interpersonal comparisons. The best-known such selection rule, the 'zero-one' rule, is available in any situation in which every individual's utility is bounded above and below: one can then select, for each individual, the utility function whose greatest lower bound is zero, and whose least upper bound is one. (This rule is employed, if not argued for, by Isbell (1959) and Schick (1971).) There are, of course, other possibilities: for example, one could equalise the greatest lower bound (setting this to zero for each individual) and the sum of the utilities of all other alternatives, or one could equalise the mean and the variance.

These proposals suffice to recover consistency with existing interpersonal comparisons. Like the simpler model discussed in the main text, however, these proposals all require interpersonal comparisons to supervene on the profiles of individuals' *relative* judgments on alternatives, and for that reason are open to conceptually similar objections. There are three main objections, which we record here for completeness. (We will state them using the zero-one rule just discussed.)

Firstly: the zero-one rule leads to arguably counterintuitive verdicts in particular cases, where intuition seems to hold that there might simply be more at stake for one person than there is for another. There are two ways of interpreting the dictates of the rule, which we will call 'narrow' and 'broad'; these different interpretations of the rule are subject to different versions of the problem. A narrow interpretation of the rule is one according to which we select the most- and least-preferred alternatives, for the purpose of calibrating the utility functions of distinct agents, *only from among the options in play in a given choice situation*. The narrow interpretation is subject to obvious problems: clearly Kate and John can be such that Kate's well-being is affected far more by choice of ice-cream flavour than John's is; if ice-cream flavour choices are all that is relevant to the case at hand, the narrow interpretation of the rule is committed to denying this datum. This motivates moving to a broad interpretation of the rule, according to which we select the most- and least-preferred alternatives, for each agent, from among all conceivable options. Here the intuition is less clear: it is not obvious that there are pairs of people who exhibit differences in how much the realization of their most preferred and least preferred options matters to them. But insofar as there is some intuition that this could happen, that is an intuition that the proponent of the zero-one rule has to deny.

Secondly: given the diversity of possible structuralist proposals, in the absence of any argument for one particular such proposal over the others, the postulation of any particular one would be unacceptably arbitrary. (This worry is pressed, in a discussion of interpersonal well-being comparisons, by Sen (1970, 98).) The worry obviously dissolves if it can be argued that one particular structuralist proposal is better than the others; for such an argument for the superiority of mean-variance normalisation, albeit in a different context, see Cotton-Barratt et al. (2014). (Cotton-Barratt et al. are actually arguing only for

This brief discussion of structuralism is enough for us to state the problem with the Borda rule as a tool for the proponent of extended preferences. Using the Borda rule ultimately requires giving serious conceptual weight to the *position* of an alternative in an individual's preference ranking. But if one is happy with this emphasis on position after all, a much simpler option is to skip the appeal to extended preferences altogether and go structuralist from the start. This isn't to say there's *no* position which preserves a conceptual distinction between structuralism on the one hand, and the use of the Borda rule in the extended preferences setting on the other. But it is hard to make out a position where the use of the Borda rule in the extended preference setting would have some advantage over the far simpler theory which is structuralist through and through. The use of the Borda rule thus undermines some of the appeal of extended preferences insofar as it ultimately requires making significant use of structural properties of agents' preference orderings. And there is a general lesson here: the proponent of extended-preferences must take care, in seeking an escape route from the Arrow-like theorem of Sect. 5 via violation of IIA, that she is not thereby invoking features which would on their own be enough to solve the problem with which we began.

A second approach to rejecting IIA is also unpromising, but for a quite different reason. In an interesting recent series of contributions to the literature on social choice theory, Fleurbaey (2007) and Fleurbaey and Maniquet (2008a, 2008b) investigate aggregation rules according to which the output ranking of alternatives x and y depends, not only on the input profile of preferences regarding x and y, but also on the properties of the alternatives that each individual ranks as being indifferent to each of x and y. Working in economic contexts in which the alternatives are assignments of consumption bundles to individuals, their rules assign a privileged status to the alternative in which each individual receives an equal share of each resource. This second type of rule, however, is clearly of no help in the extended preferences context, because we do not have any extended alternatives that are plausible candidates for having this privileged status. (Fleurbaey and Maniquet assign a privileged status to the equal-split alternatives on grounds of fairness, but, whatever their role in a context of distributive justice might be, considerations of fairness do not have the same place in extended-preference theory.) The general lesson is that not every rule that is available in the

Footnote 18 continued
equalisation of variances, rather than of means and variances, since level comparisons are irrelevant in the context they focus on.)

Thirdly: the verdicts that the zero-one rule yields on questions of interpersonal comparisons depend on some things that arguably they should not depend upon. Most obviously, on the 'narrow' interpretation, questions of interpersonal level comparisons regarding state of affairs x, or interpersonal unit comparisons regarding states of affairs x and y, can depend on whether or not some particular third state of affairs z is also included in the set $\mathcal{S}$ relative to which the zero-one rule is specified (the 'set of states of affairs under consideration'). A natural response to this is to stipulate that $\mathcal{S}$ is to include *all possible* states of affairs—that the rule is to be interpreted broadly—but it is also unclear whether there is any privileged sense of 'possible' with boundaries that are sufficiently determinate for present purposes.

For further discussion of the structuralist program, see e.g. Cotton-Barratt et al. (2014), Griffin (1986), Hammond (1991, 216), Hausman (1995), Jeffrey (1971, 655), Rawls (1999, 283–284).

social-choice context will even be definable in the extended preferences context, owing to the relative lack of structure in the set of extended alternatives.

The question then is whether there are other IIA-violating rules: ones that both (1) unlike the Borda rules, can be appealed to without undermining some of the motivation for the extended preferences program and (2) unlike the rules investigated by Fleurbaey and Maniquet, can be defined in the extended preferences context.

In this connection, we regard the following avenue as worthy of further investigation. The *Kendall tau distance* between two binary relations R, R' is the number of ordered pairs $\langle x, y \rangle$ such that xPy but $yP'x$ (where P and P' as usual stand for the strict relations derived from R and R' respectively). Relative to an input profile $R \in \mathcal{R}^N$, the *Kemeny score* of a candidate output relation $f(R)$ is the sum of the Kendall tau distances between $f(R)$ and the input relations R_i of each individual $i \in N$. The *Kemeny-Young rule* selects, for any input profile, that output relation[19] that has the lowest corresponding Kemeny score. This rule satisfies all of the axioms of our impossibility theorem except for IIA, and there is no reason to think that it will lead to Spinelessness on an overly large set of extended alternatives.

The extended-preference theorist will not want to use the Kemeny-Young rule itself, if only for the reason that this rule, like the others discussed in this section, is a relation aggregation rule, not a *utility* aggregation rule. If individuals are supposed to have extended utility functions, and not merely extended preference relations on X, to aggregate extended preferences by means of a relation aggregation rule would be to throw away relevant information; further, since we ultimately want interpersonal unit- as well as level-comparisons, the proponent of extended preferences should seek an aggregation rule whose *output*, too, is a utility function rather than merely an ordering. (We noted in Sect. 5.1, following Sen, that every utility aggregation rule *that satisfies an analogue of IIA* reduces to a relation aggregation rule; but no such reducibility holds if, as here, the independence condition is jettisoned.) Our suggestion is therefore that the extended-preference theorist explore utility-aggregation analogues of the Kemeny-Young rule, and investigate the acceptability of these analogues for the purpose of connecting profiles of individual extended utility functions to betterness-for-the-individual facts. (A related consideration is that the Kemeny-Young rule, as it stands, applies only when the set of alternatives is finite, which is plausibly not true of the extended preferences context; any UAR variant, however, will presumably have no difficulty with infinite sets of alternatives.)

9 Conclusion

The extended preferences program is a *prima facie* promising approach for preference-satisfaction theorists to resolve the problem of interpersonal well-being comparisons. The founders of the extended preferences program believed that all

[19] Or relations; some prescription will be needed to deal with ties.

individuals would have the same extended preferences. It was thus easy to see how well-being would be determined by extended preferences: one could simply identify the 'better-off-than' relation with the unique extended preference relation shared by all individuals.

But more recently a consensus has emerged that extended preferences are not shared by all individuals. If extended preferences do differ, the program faces a difficult challenge: to come up with a way of aggregating extended preferences into a single well-being ranking. This problem is formally isomorphic to the problem identified by Arrow in his celebrated impossibility theorem, but there are important conceptual differences between the two settings. In Arrow's theorem, for example, the assumption that the output ordering must be complete can be justified by the need for policy makers to come up with a plan for every contingency. There is no similar requirement that comparisons of well-being form a complete ordering; there may well be living individuals whose *actual* well-being levels are incomparable, and it is still more plausible that some *possible* pairs of lives are incomparable in well-being terms.

But even if we relax those Arrovian assumptions that are obviously inappropriate in the extended preferences setting, we can still prove a powerful result. The Spinelessness Theorem shows that any aggregation rule defined on a Sufficiently Diverse domain of preferences will be guaranteed to be Spineless (on those subsets of alternatives for which the Sufficient Diversity condition holds). Spinelessness of this kind should be unacceptable to the proponent of extended preferences, since it would result in a problematic degree of interpersonal incomparability in well-being. We considered three responses to this problem on behalf of the extended preference theorist. The first—attempting to deny Sufficient Diversity—proved hopeless; we could at most justify domain restrictions for various more 'local' aggregation rules each of which was defined only for relatively few preference profiles, but nothing in this discussion was able to prevent the 'global' aggregation rule we had previously been focussing on from existing, or to supply any resources for denying Sufficient Diversity with respect to that more global rule. The second—denying the quasi-transitivity of 'better-off-than'—was perhaps not hopeless, but is nonetheless deeply unattractive. It seems eminently plausible—even if it is not uncontroversial—that well-being comparisons are not just acyclic, they are also quasi-transitive (and indeed transitive). The third response—denying IIA—seems to us much more plausible; although we do not know of a concrete solution along these lines, we have sketched one line of investigation that may be worthy of further work (as well as two others that we regard as *un*promising).

Many seem to be attracted to the preference satisfaction theory of well-being without taking seriously the difficulty of the problem of interpersonal well-being comparisons which plagues the theory. We ourselves think that the extended preferences program will not yield the answer to this problem, for reasons independent of the problem of aggregation (see Greaves and Lederman, forthcoming). But we hope the discussion in the present paper will help those who are more sanguine about the prospects of the program to isolate aggregation rules which will be useful for their purposes.

Acknowledgments We thank Christian List, Teruji Thomas and John Weymark for useful discussions and correspondence.

Appendix: Proof of Theorem 2

The bulk of our proof is contained in the Field Expansion Lemma that forms the core of the proof of Arrow's theorem, and in Weymark's proof that this lemma in turn implies his Theorem 1 (Weymark 1984).

We use the following definitions. As before, let N be a finite set of individuals, and let X be a finite set of alternatives. Let $G \subseteq N$ be an arbitrary set of individuals. Let $x, y \in X$ be any alternatives. Let f be an arbitrary relation aggregation rule with domain $D \subseteq (\mathcal{P}(X \times X))^N$. Then, relative to f,

- G is *semidecisive w.r.t.* (x, y) iff $\forall R \in D, (((\forall i \in G\ xP_iy) \wedge (\forall i \notin G\ yP_ix)) \rightarrow xf^P(R)y)$.
- G *is decisive w.r.t.* (x, y) iff $\forall R \in D, (\forall i \in G\ xP_iy \rightarrow xf^P(R)y)$.
- G is *decisive* iff G is decisive w.r.t. every pair of alternatives.
- i *has a veto w.r.t.* (x, y) iff $\forall R \in D, (yP_ix \rightarrow \neg xf^P(R)y)$.
- G is an *oligarchy for* $Z \subseteq X$ iff for all $x, y \in Z$, (i) G is decisive w.r.t. (x, y), and (ii) every member of G has a veto w.r.t. (x, y).
- G is an *oligarchy* iff G is an oligarchy for X.

We recall also the following definitions, where π is a permutation of N, which induces a corresponding permutation of X:

Action of π on relations $r \in \mathcal{R}$: $\forall x, y \in X, (x(\pi(r))y \leftrightarrow (\pi^{-1}(x))r(\pi^{-1}(y)))$;
Action of π on profiles $R \in \mathcal{R}^N$: $\pi(R) = \left(\pi\left(R_{\pi^{-1}(1)}\right), \ldots, \pi\left(R_{\pi^{-1}(|N|)}\right)\right)$;
Action of π on aggregation rules f: $\forall R \in \pi D, (\pi f)(R) = \pi(f(\pi^{-1}R))$.
A (Anonymity): For all permutations π of N, $f = \pi f$.

Our claim (recall) is

Theorem 2. *Let f be a RAR with domain $D = \mathcal{R}_{rat}^N$. Suppose that f satisfies RF, WP, QT and IIA. Let $Z \subseteq X$ be any set of extended alternatives with respect to which $(\mathcal{R}_{rat})^N$ satisfies SD, and such that $f|_{(\mathcal{R}_{rat}|_Z)^N}$ satisfies Anonymity. Then f is Spineless with respect to Z.*

The proof uses the following lemmas.

Lemma 2 (Field Expansion Lemma). *Let f be an RAR that satisfies QT, WP and IIA, and whose domain D satisfies SD w.r.t. X. If a subpopulation $G \subseteq N$ is semidecisive over any pair of alternatives, then G is decisive.*

Proof. See e.g. Arrow (1963, 98-100), Sen (1986, 1080). (Arrow and Sen officially assume Universal Domain, but in fact their proofs of this Lemma only require the far weaker condition SD.) □

Lemma 3. *Let f be an RAR whose domain D satisfies SD w.r.t. X. Then there is at most one oligarchy relative to f.*

Proof. See Weymark (1984, Lemma 2); again, the proof only requires SD rather than the UD condition officially assumed by Weymark. □

Given Lemmas 2 and 3, we can establish the following:

Lemma 4 (Weymark's oligarchy theorem). *Let f be an RAR that satisfies RF, QT, WP and IIA, and whose domain D satisfies SD w.r.t. X. Then there exists a unique oligarchy relative to f.*

Proof. Weymark (1984), Theorem 1. □

The proof of our theorem is then as follows.

Proof. If f satisfies IIA, then f naturally induces a RAR $f|_{(\mathcal{R}_{rat}|_Z)^N}$ for the aggregation of preferences over $Z \subseteq X$, by restriction. Given that $\mathcal{R}_{rat}^N$ satisfies SD w.r.t. Z, so does $(\mathcal{R}_{rat}|_Z)^N$. Thus, applying Lemma 4 to the RAR $f|_{(\mathcal{R}_{rat}|_Z)^N}$ establishes that there exists a unique oligarchy G (for Z) with respect to $f|_{(\mathcal{R}_{rat}|_Z)^N}$.

We next show that $G = N$. Suppose, for contradiction, that $G \subsetneq N$. Let π be any permutation of N that maps one or more members of $N \setminus G$ to members of G (since $G \subsetneq N$, such a permutation exists). It is straightforward to check that if G is an oligarchy (for Z) relative to $f|_{(\mathcal{R}_{rat}|_Z)^N}$, then, for any permutation π of N such that $(\mathcal{R}_{rat}|_Z)^N$ is closed under π, πG is an oligarchy (for $\pi(Z)$, and thus since $\pi(Z) = Z$, for Z) relative to $\pi f|_{(\mathcal{R}_{rat}|_Z)^N}$. Since (by assumption) $f|_{(\mathcal{R}_{rat}|_Z)^N}$ satisfies Anonymity, however, we have $f_{(\mathcal{R}_{rat}|_Z)^N} = \pi f_{(\mathcal{R}_{rat}|_Z)^N}$. Since we have chosen π such that $\pi G \neq G$, this contradicts Lemma 3.

In case $G = N$, every individual has a veto for every pair of alternatives in Z, relative to $f|_{(\mathcal{R}_{rat}|_Z)^N}$. But if this is true relative to $f|_{(\mathcal{R}_{rat}|_Z)^N}$, then by *IIA* it is also true relative to f. That is, f is Spineless with respect to Z, as claimed. □

References

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

Adler, M. (2012). *Well-being and fair distribution: Beyond cost-benefit analysis*. Oxford: Oxford University Press.

Adler, M. (forthcoming). Extended preferences. In: M. Adler & M. Fleurbaey (Eds.), *Oxford handbook of well-being and public policy*. New York: Oxford University Press.

Adler, M. D. (2015). Aggregating moral preferences. *Economics and philosophy, 32*(2), 283–321.

Arrow, K. (1963). *Social choice and individual values* (2nd ed.). New York: Wiley.

Austen-Smith, D., & Banks, J. (2000). *Positive political theory I*. Michigan: University of Michigan Press.

Brandenburger, A. (2003). On the existence of a "complete" possibility structure. *Cognitive Processes and Economic Behavior*, 30–34.

Brown, D. (1973). *Acyclic choice*. Cowles Foundation Discussion Papers 360. Cowles Foundation for Research in Economics, Yale University.

Cotton-Barratt, O., MacAskill, W., & Ord, T. (2014). *Normative uncertainty, intertheoretic comparisons, and variance normalisation*. Unpublished manuscript.
Fishburn, P. C. (1970). Arrow's impossibility theorem: Concise proof and infinite voters. *Journal of Economic Theory*, *2*(1), 103–106.
Fleurbaey, M., & Maniquet, F. (2008a). Fair social orderings. *Economic Theory*, *34*(1), 25–45.
Fleurbaey, M. (2007). Social choice and just institutions: New perspectives. *Economics and Philosophy*, *23*(1), 15–43.
Fleurbaey, M., & Maniquet, F. (2008b). Utilitarianism versus fairness in welfare economics. In M. Fleurbaey, M. Salles, & J. Weymark (Eds.), *Justice, political liberalism, and utilitarianism* (pp. 263–280). Cambridge: Cambridge University Press.
Greaves, H, & Lederman, H. Forthcoming. Extended preferences and interpersonal comparisons of well-being. *Philosophy and Phenomenological Research*.
Griffin, J. (1986). *Well-being: Its meaning, measurement, and moral importance*. Oxford: Oxford University Press.
Hammond, P. (1991). Interpersonal comparisons of utility: Why and how they are and should be made. In J. Elster & J. E. Roemer (Eds.), *Interpersonal comparisons of well-being* (pp. 200–254). Cambridge: Cambridge University Press.
Harsanyi, J. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, *63*(4), 309–321.
Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, *61*(5), 434.
Hausman, D. (1995). The impossibility of interpersonal utility comparisons. *Mind*, *104*(415), 473–490.
Isbell, J. R. (1959). Absolute games. In A. W. Tucker & R. D. Luce (Eds.), *Contributions to the theory of games* (Vol. 4, p. 357). Princeton: Princeton University Press.
Jeffrey, R. (1971). On interpersonal utility theory. *Journal of Philosophy*, *68*(20), 647–656.
Kaplan, D. (1995). A problem in possible-world semantics. *Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus*, 41–52.
Kirman, A. P., & Sondermann, D. (1972). Arrow's theorem, many agents, and invisible dictators. *Journal of Economic Theory*, *5*(2), 267–277.
Kripke, S. A. (2011). A puzzle about time and thought. In *Philosophical troubles*. Oxford: Oxford University Press.
Mongin, P. (1994). Harsanyi's aggregation theorem: Multi-profile version and unsettled questions. *Social Choice and Welfare*, *11*(4), 331–354.
Rachels, S. (1998). Counterexamples to the transitivity of 'better than'. *Australasian Journal of Philosophy*, *76*(1), 71–83.
Rawls, J. (1999). *A theory of justice* (revised ed.). Cambridge: Belknapp Press of Harvard University Press.
Schick, F. (1971). Beyond utilitarianism. *The Journal of Philosophy, 68*(20), 657–666.
Sen, A. (1986). Social choice theory. In K. D. Arrow & M. D. Intriligator (Eds.), *Handbook of mathematical economics* (Vol. 3, pp. 1073–1181). Amsterdam: Elsevier.
Sen, A. (1970). *Collective choice and social welfare*. San Francisco: Holden-Day.
Temkin, L. (1987). Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, *16*(2), 138–187.
Temkin, L. (2014). *Rethinking the good: Moral ideals and the nature of practical reasoning*. Oxford: Oxford University Press.
Voorhoeve, A. (2014). Review of Matthew D. Adler: Well-being and fair distribution: Beyond cost-benefit analysis. *Social Choice and Welfare*, *42*(1), 245–54.
Weymark, J. A. (1984). Arrow's theorem with social quasi-orderings. *Public Choice*, *42*(3), 235–246.