AGGREGATING EXTENDED PREFERENCES

HILARY GREAVES AND HARVEY LEDERMAN

Abstract. An important objection to preference-satisfaction theories of well-being claims that they cannot make sense of interpersonal comparisons of well-being. A tradition dating back to Harsanyi (1953) attempts to solve this problem by appeal to so-called *extended* preferences, that is, roughly, preferences over situations *whose description includes one's ordinary (non-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.

1. Introduction

Queen Victoria was better off than an average Roman slave. 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 in a middle-class family is better off than one who does not. The first of these obvious truths compares the *levels* of well-being of different individuals; the second compares 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.

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. Preferences as ordinarily conceived, such as Agnes's preference for eating meat as opposed to eating fish, concern only what would happen in the relevant situation as viewed from the agent's own perspective. Of course, Agnes may have preferences for what happens to people other than herself: for example, Agnes may prefer Brandon's eating fish to Brandon's eating meat, because if Brandon eats fish he will not complain of indigestion later. But Agnes's preferences of this kind do not concern *being Brandon* and eating meat as opposed to *being Brandon* and eating fish: they only concern the various possibilities for Brandon's diet as seen from Agnes's point of view. So Agnes's preferences and Brandon's preferences seem to be preferences over different sorts of objects: Agnes's perferences concern how things would be from Agnes's point of view, while Brandon's concern how things would be from Brandon's point of view. But neither person's preferences tell us how to compare a situation viewed from Brandon's perspective with one viewed from Agnes's perspective. And since each individual's well-being is supposed to be

Date: November 8, 2015. Please do not cite without permission!
The authors contributed equally to this paper.

determined by his or her preferences, if preferences do not tell us how to compare the invidividuals' well-being it is unclear how they could be compared at all.

A popular line of response to this problem invokes *extended preferences*. According to this response, people do not just have preferences over alternatives viewed from their own perspectives; they also have preferences over situations viewed from others' perspectives. For example, Agnes may prefer being Agnes while Agnes eats meat to being Brandon while Brandon eats meat (because in that scenario she, as Brandon, would experience such bad indigestion). Since each individual is claimed to have extended preferences, all individuals have preferences over the same kinds of objects: situations viewed from every individuals' perspectives. Seen in this light, interpersonal well-being comparisons only seemed to pose a problem because we thought all preferences were ordinary preferences; we did not take account of people's extended preferences as well.

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. Assuming that one does not wish to retreat to expressivism or subjectivism (and we will assume, following proponents of the extended preferences program, that we do not), it remains to be said how individuals' extended preference rankings will combine to determine facts about well-being. 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 a growing consensus has recognized that individuals may not have the same extended preference relation. 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 celebrated impossibility theorem, and explore ways of avoiding analogues of Arrow's result in this context.[1]

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 the basic setup of Arrow's theorem. Section 5 describes a variant on Arrow's theorem which is based on weaker assumptions, which are 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:

[1]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) points out the incomparability problem discussed in section 3.

denying that the aggregation rule satisfies the formal condition of Independence of Irrelevant Alternatives.

2. Setup

We can state the problem of aggregation formally in a simple, abstract setting. There is a finite population of individuals N, where each $i \in N$ has an extended preference relation $\succeq_i$ over a finite set of extended alternatives X. (For simplicity, we assume throughout the paper that both N and X are finite. As far as we're aware nothing depends essentially on this simplification.) In a moment we'll say something about what these 'extended alternatives' might be, but any proponent of extended preferences will believe extended preferences are a binary relation on some set or other. For the whole of the paper, we will think of individuals' preferences as binary relations, and binary relations as sets of ordered pairs; a binary relation over X is thus a subset of $X \times X$, an element of $\mathcal{P}(X \times X)$. A *profile* of extended preferences is a specification of each individual's extended preferences, that is, a function from N into $\mathcal{P}(X \times X)$, or equivalently, as we will sometimes write, a vector of preference relations $\langle \succeq_i \rangle_{i \in N}$. The aim of the extended preferences program is to define an *aggregation rule* $f : \mathcal{P}(X \times X)^N \to \mathcal{P}(X \times X)$, which takes in a profile of extended preferences $(\langle \succeq_i \rangle_{i \in N})$ and outputs the objective well-being ordering, an ordering over extended alternatives $\succeq \in \mathcal{P}(X \times X)$. As usual, we use $\succ_i$ and $\succ$ to represent the 'strict' portions of the relations derived from $\succeq_i$ and $\succeq$: $x \succ_i y$ just in case $x \succeq_i y$ and $\neg y \succeq_i x$ (and similarly for $\succ$).

That, then, is the formal setting for the problem of aggregation. But the constraints which will make the problem a problem will be motivated by features of the project we are engaged in: of producing interpersonal well-being comparisons from extended preferences. Motivating these constraints will sometimes require that we think more concretely about what exactly the extended alternatives are on which extended preferences are defined. As far as we're aware, nothing important hangs on the details of the framework we're about to introduce, but it will occasionally be useful to have this more concrete framwork to refer to. Let a *choice-situation* be a structure $\langle W, N, E \rangle$, where W is a set of logically possible worlds and N is a finite set of individuals. We identify the set of extended alternatives X with a set of centred worlds $X = W \times N$. These centred worlds specify not just what the world is like, but which individual is the 'centre' of the world; this is the formal implementation of our loose talk of 'a situation seen from an agent's point of view' in the introduction. The property of being i in world w is associated with the centred world $\langle w, i \rangle$; if $i \neq j$, then this centred world differs from the property of being j in w, which is associated with $\langle w, j \rangle$. The final component of the structure, $E : N \to (\mathcal{P}(X \times X)^W)$, assigns each individual a preference relation over these extended alternatives. In the most general case, E might depend on which element of w is the actual one. But a simpler model will be sufficient for the purposes of this paper: we will assume that individuals' preferences do not vary across the worlds we are considering (and thus E may be thought of as constant on W; $E : N \to \mathcal{P}(X \times X)$). This concrete model is an instance of the more abstract setting introduced above, if we let each $\succeq_i$ be given by the value of $E(i)$.

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; these relations may,

for example, fail to be transitive. But it is worth mentioning that typically preferences are also taken to be defined over *lotteries* of extended alternatives. If various well-known constraints are imposed on preferences over lotteries, the preferences in question can be represented by utility functions. On some views of these representation theorems, the utility functions then determine *intra*personal comparisons of units of well-being. If the overall well-being ordering (the output of the aggregation rule) were also to satisfy the needed axioms, the overall ordering would give rise to *inter*personal unit comparisons as well.

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.[2] But some arguments later in the paper will rely on the possibility of moving to the usual decision theoretic setting, and that is why we have mentioned it 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. In this section we will introduce the problem, by considering a particular aggregation rule, the Strong Pareto Rule, which has recently been advocated by Matthew Adler (2012, 53) for aggregating extended preferences. The Strong Pareto Rule is defined as follows:

Strong Pareto: For all $x, y \in X$, $x \succeq y$ if and only if for all $i \in N$, $x \succeq_i y$.

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.

We will begin with a simple problem with this particular rule. This simple problem will lead into a much more general statement of the problem in subsequent sections. The problem is that given diversity in individuals' preferences, the rule leads to massive incomparability. This is certainly the case if, as Adler himself

[2]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 each agent's preferences are represented by a vNM utility function, and the output ordering is also represented by a vNM utility function then if the output ordering satisfies an Ex Ante Strong Pareto condition it will be representable by a weighted sum of the individual vNM utility functions. This Strong Pareto condition is extremely plausible in the context of the EP program; as is the claim that well-being should be representable by a vNM utility function. So Harsanyi's theorem shows that any acceptable EP aggregation rule must have a particular functional form: it must be representable by a function from profiles of utility functions to 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 EP theorist's aggregation rule must belong. The 'multi-profile results' of which we are aware (e.g. Mongin (1994)), and which might otherwise threaten the EP program, rely on a condition similar to independence of irrelevant alternatives; accordingly they do not offer a new challenge to the EP theorist over and above the one we will develop in sections 4–8.

sometimes effectively suggests,[3] the 'constituency' – the members of N – exhibit all rationally permissible extended-preference relations.[4] It is natural to suppose that in this setting the only constraint on rational preferences are 'purely structural' ones, so that 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. 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 x and y will be incomparable in the output ordering generated by the Strong Pareto rule: for since it is not the case that for every $i \in N$ $x \succeq_i y$ and also not the case that for every $i \in N$ $y \succeq_i x$, it follows that it is not the case that $x \succeq y$ and also not the case that $y \succeq x$.

An obvious and natural response is for the preference-satisfaction theorist to backtrack, and say that the 'constituency', N, exhibits not all rationally permissible extended-preference orderings, but only the extended preferences that are the rational version of preferences which are actually possessed by some individual. Call this the 'actualist' EP theory, as opposed to the 'possibilist' EP theory considered in the previous paragraph. The simple argument of the preceding paragraph will not affect the actualist EP theory: 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 case. According to the Strong Pareto Rule, every person in some sense 'has a veto' regarding every pair of extended alternatives: that is, the rule's output will refrain from ranking x as being even *weakly* better than y whenever there is *any* person who strictly prefers y to x. It only takes one person to 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, then it is overwhelmingly plausible that for almost any pair of extended alternatives, 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 are incomparable in terms of well-being.

Such massive incomparability is radically implausible: it amounts to denying the data with which we started. To say that virtually all situations are incomparable to virtually all others is no better than to say that interpersonal well-being comparisons are 'meaningless'. For the extended preferences program to have this implication is for it to end in failure.

---

[3] Adler's (sometime) suggestion is 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, 226-7). This is presumably extensionally equivalent to including all rationally permissible extended preferences.

[4] Depending on how rich the set of rational preference relations is, there may be cardinality problems with the population exhibiting all rational preference relations in a single profile, but the argument in the main text turns only on a mild closure condition of the relevant profile, not on its truly instantiating all such preference relations.

What went wrong with the Strong Pareto Rule? We can state the problem with the Strong Pareto rule in a somewhat more abstract form. Individual $i \in N$ *has a veto* over alternatives $x, y \in X$ under some aggregation rule f just in case if $x \succ_i y$ in a given profile then in the output ordering under f it is not the case that $y \succ x$. 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 $x, y \in Z$. The Strong Pareto rule is spineless on the set of all extended alternatives: as a result, even slight variability in the preferences of the constituency gives rise to massive incomparability.

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. But, as we will show in the remainder of the paper, 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 one can show that any aggregation rule which satisfies much weaker conditions will be spineless, even if it does not have a dictator. This result already looks troubling for the prgoram, since we saw that spinelessness led to an implausible degree interpersonal incomparability. Moreover, the weaker assumptions used in this result are much more compelling than the assumptions Arrow's original theorem in the setting of extended preferences. The theorem thus presents a serious challenge to the EP program (Section 5).

4. Arrow's Theorem

Arrow's result can be formulated as follows. Let N be a fixed and finite set of individuals. Let X be an arbitrary set of cardinality at least 3. Recall that a binary relation $R \subseteq X \times X$ is an *ordering* if it is reflexive, transitive and complete. Let $\mathcal{R}$ be the set of orderings of X. An *aggregation rule (AR)* is a function $f : D \to \mathcal{P}(X \times X)$, where $D \subseteq \mathcal{P}(X \times X)^N$. An *ordering aggregation rule (OAR)* is an aggregation rule where the inputs are assumed to be orderings; that is, its domain $D \subseteq \mathcal{R}^N$. We use $\mathbf{R}$ to range over elements of $\mathcal{P}(X \times X)^N$; given such a 'preference profile' $\mathbf{R}$, and a subset $Y \subseteq X$, we write $\mathbf{R}|_Y$ to denote the restriction of $\mathbf{R}$ to Y, that is: $\langle \{\langle x, y\rangle \in R_i : x, y \in Y\}\rangle_{i \in N}$. Notice that we use $\mathbf{R}$ for relations in general; the notation $\mathbf{R}$ reflects the fact that we often require that the domain consist of at least all profiles where each individual's preferences are orderings.

We consider the following conditions on an AR f:

UD (Unrestricted Domain): $D = \mathcal{R}^N$.
R (Reflexivity): $\forall \mathbf{R} \in D$, $f(\mathbf{R})$ is reflexive.
T (Transitivity): $\forall \mathbf{R} \in D$, $f(\mathbf{R})$ is transitive.
C (Completeness): $\forall \mathbf{R} \in D$, $f(\mathbf{R})$ is complete.
WP (Weak Pareto): $\forall x, y \in X, (\forall i \in N(xR_iy)) \to xf(\mathbf{R})y$.
IIA (Independence of Irrelevant Alternatives): $(\forall x, y \in X)(\forall \mathbf{R}, \mathbf{R}' \in D)((\mathbf{R}|_{\{x,y\}} = \mathbf{R}|'_{\{x,y\}}) \to (xf(\mathbf{R})y \leftrightarrow xf(\mathbf{R}')y))$.
ND (Non-dictatorship): $\neg(\exists i \in N)(\forall \mathbf{R} \in D)(\forall x, y \in X)(xf(\mathbf{R})y \leftrightarrow xR_iy)$.

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

Theorem 1 (Arrow, 1963)**.** *There is no OAR satisfying UD, C, T, 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 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. It is also extremely difficult to see how any aggregation rule that violated the Weak Pareto condition could count as grounding betterness-for-the-individual facts in extended preferences. The spirit of the EP 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, surely the betterness facts should match individuals' unanimous judgments when such unanimity exists.

Most of the remaining conditions, however, are inappropriate in the EP context:

First, the framework itself is inappropriate: it requires an output ordering to be determined only on the basis of individuals' input preference *relations* or even *orderings*, but in fact in principle we are considering individuals' preferences over *lotteries over* extended alternatives, and thus we have 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 highly questionable: an aggregation rule for the purposes of the EP program need aggregate only *rational* extended preferences, but it may well be that even some orderings of X (that is, elements of $\mathcal{R}$) are such that it is rationally impermissible to hold the associated extended-preference ordering.

Third, the Completeness requirement 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 result. But they do not allow for an escape from a closely related one. 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. We will take these three replies in turn, as we lead up to the statement of the spinelessness theorem.

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, *the limited kind of utility information* that is available in the EP setting does not suffice to make available any essentially new aggregation rules. In other words, the first objection just described – that an aggregation rule on preference relations misses out on important information which is preserved in the utility function – cannot help to avoid an impossibility theorem, at least if IIA is in place.

The basic point is that in this setting, we have only the utility information that is recoverable from individuals' (extended-) preferences over extended lotteries. 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 $u : X \to \mathbb{R}$. Let U be the set of all such utility functions; an n-tuple of utility functions is therefore an element of U^N. A *utility aggregation rule (UAR)* is a function $\overline{f} : \overline{D} \to \mathcal{P}(X \times X)$, where $\overline{D} \subseteq U^N$. We write $\mathbf{u}$ for a typical element of U^N, u_i for the ith component of $\mathbf{u}$, and $\mathbf{u}(x)$ for the vector of real numbers $\langle u_1(x), ...u_N(x)\rangle$. 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 $\pi_{CNC} : U^N \to U^N$ 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 $\overline{f}(\mathbf{u}) = \overline{f}(\pi_{CNC}\mathbf{u})$.

Restrictions analogous to those stated above can be imposed on UARs, as well as on OARs, 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, becomes:

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

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

Definition 2. Say that a UAR $\overline{f}$ *reduces to the OAR* f iff $\forall x, y \in X, \forall \mathbf{u} \in U^N (x\overline{f}(\mathbf{u})y \leftrightarrow xf(\mathbf{R})y)$, where $\mathbf{R}$ is the preference profile that is ordinally represented by the utility profile $\mathbf{u}$ (that is, for all $i \in N$ and all $x, y \in X$,xR_iy if and only if $u_i(x) \geq u_i(y)$).

Lemma 3 (Sen, 1970)**.** *Let $\overline{f}$ be a UAR satisfying IIA* and CNC. Then, there exists an OAR f such that $\overline{f}$ reduces to f.*

Proof. This is part of the proof of Sen's Theorem 8*2.[5] □

[5]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 (of 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 applies *a fortiori* to these richer objects. Thus we lose no generality in considering only UARs in our sense.

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

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 UD 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 – the property of spinelessness already leads to an unacceptable degree of incomparability. If the conditions of the theorem are true, the result is damning for the extended preferences program.

Replacing UD with Sufficient Diversity. In place of UD, our result will only require *some* diversity among possible extended preferences. But the diversity will be comparatively minimal. 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 relations on $\{x, y, u, v\}$, there exists a profile $\mathbf{R} \in D$ whose restriction to $\{x, y, u, v\}$ is $\mathbf{r}$.

Our claim is that, the restriction to rational preferences notwithstanding, it will always be possible to find a set Z with respect to which SD is indeed true, and that is simultaneously such that if the aggregation rule were spineless on Z, that would result in a problematic degree of incomparability.

Quasi-transitivity. We will not need to assume the full transitivity condition. Instead, given a profile $\mathbf{R}$, let $f^P(\mathbf{R})$ be the asymmetric portion of $f(\mathbf{R})$. Then for this result we need only[6]

> **QT (Quasi-transitivity):** For all $\mathbf{R} \in D$, $f^P(\mathbf{R})$ is transitive.

[6]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 (of the weak betterness relation), 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.

Anonymity. Permutations of N and of X act in a natural way on the space of binary relations on X and N-tuples thereof; we thereby have a natural notion of what it means for an AR to be *invariant* under such permutations. We do not require that the aggregation rule be invariant under permutations of individuals *simpliciter*, since it is plausible that the aggregation rule might (for example, and relying on our concrete model of extended preferences) assign significance to the special connection between a given individual and extended alternatives in which she herself is the centre.[7] The following weaker requirement, however, *is* a reasonable expression of the principle that the aggregation rule should ultimately not privilege any individual over any other:

> **A (Anonymity):** For any permutation π of N, there exists a permutation ρ of X such that f is invariant under (π, ρ).

Appendix A makes precise the notion of 'invariance' used in this principle; the theorem below relies on that precise version of the Anonymity condition.

Spinelessness. 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 that led, in the context of the Strong Pareto Rule, to excessive incomparability. Let us define these notions formally.

Definition 4. Let $G \subseteq N$. Individual $i \in G$ *has a veto for* the pair $x, y \in X$ iff for all profiles $\mathbf{R} \in D$, if $x \succ_i y$ then not $yf^P(\mathbf{R})x$.

Definition 5 (Spinelessness)**.** An OAR f is *spineless with respect to* $Z \subseteq X$ iff every individual has a veto for every pair of alternatives in Z.

The Spinelessness theorem. Finally, we introduce a new name for the set of rational preference. Let $L_{rat} \subseteq \mathcal{P}(X \times X)$ be the set of rational preference relations, which so far have no formal conditions on them.

Whatever relations count as rational, we have

Theorem 6. *Let f be an OAR with domain $D = L_{rat}^N$. Let $Z \subseteq X$ be any set of extended alternatives such that $(L_{rat}|_Z)^N$ satisfies SD. Suppose that f satisfies R, WP, QT, IIA and A. Then f is Spineless with respect to Z.*

The proof is in Appendix A.[8]

Perhaps the most important aspect of this theorem is what is *not* stated here. UD has been replaced with SD, but we also no longer require *any analogue of completeness*.

5.3. **Conceptual Upshot of the Result.** The Strong Pareto rule leads to massive incommensurability because it is spineless. A natural response to this problem is to seek an alternative aggregation rule. This put us in the region of Arrovian impossibility results, but it was at first sight unclear that these results carried over to our setting. For one thing, in our setting we have utility information, not just preference orderings. But Sen's lemma shows that in the presence of

[7] See Greaves & Lederman (n.d., Section 6) for discussion of this 'special connection'.

[8] 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).

IIA, utility information cannot afford any new aggregation rules which were not already available using orderings alone. A more promising response to Arrow's theorem was to observe that both UD and completeness are too demanding in the EP context. Theorem 6 shows that these points, while correct, also fail to open the door to any acceptable aggregation rule. An impossibility result can be derived even if we weaken UD substantially and drop completeness altogether. No matter what aggregation rule we choose, if it satisfies the assumptions of Theorem 6 it is guaranteed also to be spineless, and hence to give rise to an unacceptable degree of incomparability.

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

6. Actualism, Domain Restrictions and Social Constructivism

We first consider the possibility that the problem could be met by restricting the domain, going beyond the denial of UD to deny also the weaker assumption we called SD. The most natural strategy for doing this is related to one we have seen before: to restrict attention to the preferences of actual, living individuals. What the Spinelessness Theorem shows is that any aggregation rule which is well-behaved on a sufficiently diverse domain of profiles will be spineless. But it remains possible that there could be a non-spineless aggregation rule that is well-behaved on some domain which is not sufficiently diverse.

We argued earlier that the rational versions of actual living individuals' preferences are diverse in the sense that there are *some* alternatives on which individuals exhibit diverse preferences. But we did not argue (and we do not believe) that the actual preferences of individuals are sufficiently diverse in the technical sense: plausibly it is not the case that for *every* quadruple of alternatives, there is some living individual whose extended preferences match one of the logical possibilities for an ordering on those alternatives.

In order for this restriction to actually-existing individuals' preferences to escape the theorem, however, *either* it must be a contingent matter how preferences determine well-being, *or* it must be a contingent matter that the formal conditions on the output ordering are properties of well-being comparisons – (for example) that the relation of 'better off than' is transitive. For if the mechanism by which preferences determine well-being were a necessary matter, and if it is a necessary matter that comparisons of well-being exhibit these logical properties, then since the domain of all possible preferences plausibly *is* sufficiently diverse, the domain restriction of 'going actual' won't help. Given a plausible modal recombination principle for preferences (if it is possible that an individual have rational preference ordering R, and possible that an individual have preference ordering R', then it is possible that some (possibly different) pair of individuals have R and R' respectively), it would follow that the domain of the rule which (necessarily) determines well-being on the basis of preferences is sufficiently diverse after all, so that the rule itself is spineless.

The problem is that neither of these options seems to us at all attractive. The relevant logical properties of 'better off than' seem to us *logical* properties of the relation of better off than: if they hold at all, then they hold of necessity. Moreover, since the relationship between preferences and well-being is supposed to form part of an analysis of well-being, it seems implausible that it would turn out to be a

contingent matter that this analysis was true. There are of course well-known examples of the contingent *a priori*, but it is unclear how this example could fall under that category. If it is not *a priori*, we have the surprising result that there should be some observations we could make which would help to make to determine the properties of the aggregation rule which determines well-being on the basis of preferences.

That is a first objection to this reply to the theorem: it requires assuming that two aspects of well-being which are as clear as any examples of necessary truths are not necessary but are in fact contingent. Our second objection derives from a quite different aspect of the reply. The ordinary preference-satisfaction theory is constructivist, in the (minimal) sense that it constructs the facts about what is good for a particular individual from *that individual's* attitudes. This is, of course, the source of the theory's chief attractions: for instance, it allows the theory straightforwardly to ground evaluative facts in a naturalistically acceptable basis, and avoids the problem of 'alienation' that arguably plagues objective-list theories (Railton, 1986, 9). It is also the source of some of the main objections to preference-satisfaction theory. For instance, it is typically thought that preference-satisfaction theories only concern ideally rational and fully-informed, self-interested preferences. Even with these qualifications added, the preference satisfaction theorist is committed to there being a true reading of such counterfactuals as: 'if Pat were ideally coherent and really preferred grass-counting, then grass-counting would be better for Pat'.[9]

We presume that the preference satisfaction theorist will have made her peace with the true readings of these counterfactuals. The point we wish to press now is that any actualist version of the EP program threatens a significant expansion of this existing oddness. Recall that actualist versions of the EP program aggregate only the (idealised) preferences of *actually* existing individuals, while possibilist versions aggregate the preferences of *all possible* rational preferences. The expansion in oddness occurs if the aggregation rule – like the Strong Pareto rule – makes the betterness-for-the-individual relation between pairs of individuals depend on what everyone actually prefers. In our concrete setting, this would be to say, for example, that the comparison between Makena-centred and Laurence-centred extended alternatives depends nontrivially on *everyone's* extended preferences, not only on those of Makena and Laurence. In such a 'social constructivist' approach, the betterness-for-the-individual comparisons between Makena-centred and Laurence-centred extended alternatives will also depend in part on which other individuals

[9]We don't think it's a viable option to claim that even given the truth of preference-satisfaction theory there would be no true reading of such conditionals. One option that might seem initially promising is to claim that the consequent of the conditional is mandatorily read by reference to the actual standards for well-being, so that since grass counting is in fact not better for Pat (because he doesn't prefer it), the consequent of the conditional is false. One might take hope from a related reading of conditionals involving the metre-stick; even on the supposition that the metre stick necessarily determines the length of one metre, there is a false reading of the sentence 'If the metre stick were three centimetres longer, then it would still be a metre.' But the problem is that (again on the hypothesis that the metre-stick necessarily sets the length of one metre) there is *also* a reading on which the same sentence is true. On the 'mandatory actual standards' theory of the counterfactuals involving odd preferences, however, there would be no true reading of the sentence, so these counterfactuals would be very different from any other English examples we're aware of. Positing some new linguistic phenomenon with no independent motivation seems to us more bizarre than the result it was supposed to avoid: that the relevant counterfactuals have a true reading.

exist (or on which other individuals are included in the constituency for the purposes of applying the aggregation rule). That is, such a theory will then be committed to there being true readings of some counterfactuals of the following two forms:

(X1): As things stand, Makena eating meat is better off than Laurence eating fish, but if Quinn had preferred to be Laurence eating fish than to be Makena eating meat, then Makena eating meat would not have been better off than Laurence eating fish.

(X2): As things stand, Makena eating meat is better off than Laurence eating fish, but if Quinn had never been born, then Makena eating meat would not have been better off than Laurence eating fish.

In our view this constitutes a new, significant objection to an actualist version of the EP program.[10]

We conclude that 'going actualist' is highly unattractive. On the one hand, it avoids SD at the cost of denying either that it is a necessary matter how preferences determine well-being, or that apparently 'logical' properties of comparisons of well-being are themselves necessary. Even if we could resign ourselves to swallowing this difficult pill, there is the further problem of counterfactuals which are bizarre even by the bizarre standards of the preference-satisfaction theorist.

7. Aggregation Rules that Violate Quasi-Transitivity

We have argued above for conditions SD, WP, A and Non-Spinelessness. Given Theorem 6, 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 take these remaining possibilities in 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 $x_1,...,x_n$ $(n \geq 2)$ such that $x_1 \succ x_2 \succ ... \succ x_n \succ x_1$ – 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 & 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: they introduce the following additional 'Neutrality' condition:

[10] As mentioned in the previous note, the issue is that there seems to be *no* reading on which these conditionals are true, whereas the preference-satisfaction theorist is committed to there being *some* reading on which they are.

N (Neutrality): f is invariant under permutations of alternatives. That is, for any permutation ρ of X, (i) the domain D of f is invariant under ρ, and (ii) $\rho f(\mathbf{R}) = f(\rho\mathbf{R})$.

At first sight, Neutrality seems 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 it may be positively inappropriate, in the present context, to impose Neutrality. It is arguably reasonable for an aggregation rule to treat the relationship between Makena's extended preferences on the one hand, and extended alternatives that are centred on Makena, as special.[11] Thus, there may be an aggregation rule that satisfies Acyclicity, violates Neutrality, satisfies the axioms of Theorem 6 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 this approach is *possible,* would-be 'betterness relations' that satisfy Acyclicity but not Quasi-Transitivity are, in our view, strange indeed. So we regard this way of replying to the theorem as comparatively unattractive. The next subsection turns to the final response, which we think is the most promising of the three.

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 EP context.[12] It *is* formally very natural, and there are also no clear arguments against it, but in the absence of a positive argument for it, we think denying IIA is the natural route to take. We therefore recommend this avenue of investigation to the extended-preference theorist.

But how exactly should we do this? There are many known rules which violate IIA, and it is not our aim to decide here on a particular one. Instead, we will narrow down the options by pointing out some rules that strike us as *un*promising in the context of the EP 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 an ordering aggregation rule that violates IIA is the *Borda rule*. The rule can be described as follows: suppose that the number $|X|$ of alternatives is finite, and assign a nonnegative integer $n(i,\alpha)$ to each pair consisting of an individual i and an alternative α, such that for each individual i, the most-preferred alternative is assigned the highest integer $n(i,\alpha) = |X|$, while the least-preferred alternative is assigned $n(i,\alpha) = 1$. The overall Borda score for a given alternative is given by summing these rankings across all individuals:

[11] See again Greaves & Lederman (n.d., Section 6) for discussion of why this relationship might be thought to be 'special'.

[12] In contexts of 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 EP 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.

$B(\alpha) = \sum_{i \in N} n(i, \alpha)$. The Borda rule then ranks one alternative above another just in case the first has a higher Borda score $B(\alpha)$ than the second.

The Borda rule, however, is merely an ordinalist analogue of a utilitarianism that is based on a very different way of producing interpersonal well-being comparisons in a preference-satisfaction setting. Unlike extended preference theory, this alternative approach, which we call *structuralism,* seeks to define interpersonal comparisons on the basis of the structure that is already present in a profile of preference orderings, without expanding the objects of ordinary preferences.

The most straightforward structuralist proposal is a close analog of the Borda rule. Suppose that the number of ordinary alternatives 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 might then define interpersonal well-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 $\frac{n(i,x)-n(i,y)}{n(j,v)-n(j,w)}$.

The structuralist approach relies on the possibility of calibrating individuals' utility functions, that is, of choosing one utility function for each individual as a distinguished representative of that individual's positive affine family of utility functions. The most common such selection rule, the 'zero-one rule', is available in any situation in which every individual's utility is bounded above and below: that is, if for each individual there is either a most-preferred and a least-preferred alternative or, failing that, a lowest upper bound and a greatest lower bound to that individual's utility (for any given utility function in that individual's positive-affine family). In that case there exists, for each individual, a unique vNM utility function such that the greatest lower bound is zero, and the least upper bound is one; the zero-one rule selects, for each individual, the utility function that has this property. (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 the mean and the variance.

Much more could be said here. But this brief discussion of structuralism is enough to present our objection to the Borda rule. The objection is this: it is hard to see why one would be willing to use any such approach for the purpose of aggregating *extended* preferences, if one was not already willing to use such structuralist considerations directly to construct interpersonal comparisons of *ordinary* preferences. (If, for example, one rejected the zero-one method of interpersonal comparisons on grounds of concern that some individuals' ordinary preferences might globally be stronger than others and that this should be taken into account, presumably one should think the same of extended preferences, and thus reject the Borda method of aggregating extended preferences.) But, of course, if one *is* happy with a structuralist standard of interpersonal comparison for ordinary preferences, then the motivation for the EP program dissolves. The motivations for the EP program, after all, were to save ordinary preference theory from the problem of interpersonal comparisons. But if structuralism does just as good a job, then this motivation vanishes. The general lesson of this example is that the extended-preference theorist must take care, in seeking an escape route from the Arrow-like

theorem of section 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 unpromising, but for quite a different reason. In an interesting recent series of contributions to the literature on social choice theory, Fleurbaey (2007) and Fleurbaey & Maniquet (2008a,b) investigate aggregation rules according to which the output ranking of alternatives 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 EP 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 social choice context 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 social-choice context will even be definable in the EP 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 (i) unlike the Borda rules, can be appealed to without undermining the motivation for the EP program and (ii) unlike the rules investigated by Fleurbaey and Maniquet, can be defined in the EP context.

We regard the following avenue as worthy of further investigation. The *Kendall tau distance* between two binary relations $\succeq, \succeq'$ is the number of ordered pairs $\langle x, y\rangle$ such that $x \succ y$ but $y \succ' x$. Relative to an input profile $\mathbf{R} \in D$, the *Kemeny score* of a candidate output relation $\succeq$ is the sum of the Kendall tau distances between $\succeq$ and the input relations $\succeq_i$ of each individual $i \in N$. The *Kemeny-Young rule* selects, for any input profile, that output relation[13] 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 any significant degree of Spinelessness.

The extended-preference theorist will not want to use the Kemeny-Young rule itself, if only for the reason that that rule, like the others discussed in this section, is an 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 an ordering aggregation rule would be to throw away relevant information; further, since we ultimately want interpersonal unit- as well as level-comparisons, the EP theorist should seek an aggregation rule whose *output*, too, is a utility function rather than merely an ordering. (We noted in section 5.1, following Sen, that every utility aggregation rule *that satisfies an independence condition* reduces to an ordering 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

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

that the Kemeny-Young rule, as it stands, applies only when the set of alternatives is finite, which is not true of the EP 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 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 a growing consensus has recognized that extended preferences are not shared by all individuals. In this setting, where extended preferences may 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 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 order; there may well be living individuals whose well-being levels are incomparable.

But even if we relax assumptions which are inappropriate in this setting, we can still prove an apparently challenging result. As we have shown, any aggregation rule satisfying comparatively modest assumptions will be guaranteed to be spineless. We considered three responses to this problem on behalf of the extended preference theorist. The first – attempting to deny sufficient diversity – seemed to us hopeless; it led to contingency in the ways in which preferences determined well-being, and in any event to a form of 'social' constructivism that seemed unacceptable. The second – denying the quasi-transitivity of 'better-off-than' – was perhaps not hopeless but is nonetheless unattractive. It seems eminently plausible – even if it is not uncontroversial – that well-being comparisons are not just acyclic, they are also quasi-transitive. The third response – denying IIA – seems to us more plausible, although we do not know of a concrete solution along these lines.

We ourselves think that the extended preferences program cannot offer a solution to the troubles which beset the preference-satisfaction theory of well-being (see Greaves & Lederman (n.d.)). But we hope that this paper will help those who are more sanguine about its prospects to isolate aggregation rules which will be useful for their purposes. Many seem to be attracted to the preference satisfaction theory of well-being without taking seriously some of its most glaring defects. One of these defects is the problem of interpersonal comparisons of well-being. We cannot make progress on this problem simply by mulling over vague dicta about the prospects for a solution. We can make progress only by looking at the details of the theories on offer. This paper presents some first small steps in that direction. We hope others will be able to make greater strides.

Appendix A. Proof of Theorem 6

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. Let $G \subseteq N$ be an arbitrary set of individuals. Let $x, y \in X$ be any alternatives. Let$f$ be an arbitrary ordering aggregation rule, and let $D \subseteq \mathcal{P}(X \times X)$ be the domain of f. Then

- G is *semidecisive w.r.t.* (x, y) iff $\forall \mathbf{R} \in D, (\forall i \in G x P_i y \wedge \forall i \notin G y P_i x) \rightarrow xPy$.
- G *is decisive w.r.t.* (x, y) iff $\forall \mathbf{R} \in D, (\forall i \in G x P_i y) \rightarrow xPy$.
- G is *decisive* iff G is decisive w.r.t. every pair of alternatives.
- i *has a veto w.r.t.* (x, y) iff $\forall \mathbf{R} \in D, y P_i x \rightarrow \neg xPy$.
- i *has a veto* iff i has a veto w.r.t. every pair of alternatives.
- G is an *oligarchy* relative to f iff (i) G is decisive, and (ii) every member of G has a veto.

Let π, ρ be permutations of N, X respectively. Any such permutations act in a natural way on the space $\mathcal{R}^N$ of N-tuples of orderings of X:

- For any $\mathbf{R} \in \mathcal{R}^N$, define $\pi\mathbf{R}$ by the condition that $\forall i \in N, (\pi\mathbf{R})_i = R_{\pi(i)}$.
- For any $\mathbf{R} \in \mathcal{R}^N$, define $\rho\mathbf{R}$ by the condition that $\forall i \in N, x \left(\rho\mathbf{R}\right)_i y \leftrightarrow (\rho x) R_i (\rho y)$.

Say that the domain D of f is invariant under π iff $D = \{\pi(\mathbf{R}) : \mathbf{R} \in D\}$; similarly it is invariant under ρ iff $D = \{\rho(\mathbf{R}) : \mathbf{R} \in D\}$. If the domain D of f is invariant under ρ, then we can define an aggregation rule ρf in a natural way:$\forall \mathbf{R} \in D, (\rho f)(\mathbf{R}) = \rho^{-1} f(\rho\mathbf{R})$. Similarly, if D is invariant under π, then there is a natural way to define a 'permuted' aggregation rule of πf by requiring that: $\forall \mathbf{R} \in D, \pi f(\mathbf{R}) = f(\pi\mathbf{R})$.

We can now give the formal statement of the last condition required for our impossibility result:

A (Anonymity, formal statement): For any permutation π of N, there exists a permutation ρ of X such that (i) D is invariant under ρ, and (ii) $\pi\rho f = f$.

Our claim (recall) is

Theorem 6. *Let f be an OAR with domain $D = L_{rat}^N$. Let $X \subseteq X_1$ be any set of extended alternatives such that $(L_{rat}|_X)^N$ satisfies SD. Suppose that f satisfies R, WP, QT, IIA and A*. Then f is Spineless with respect to X.*

The proof is as follows.

Lemma 7. *(Field Expansion Lemma) Let $D \subseteq \mathcal{R}^N$ satisfy SD. Let f be an OAR with domain D satisfying QT, WP and IIA. 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). □

Lemma 8. *Let f be an OAR whose domain satisfies SD. Then there is at most one oligarchy relative to f.*

Proof. Let G, G' be oligarchies relative to f. Suppose, for a contradiction, that $G \neq G'$; WLOG, suppose that $G' \backslash G \neq \emptyset$. Consider any profile such that

$\forall i \in G, x P_i y$;

$\forall i \in G' \backslash G, yP_i x$.

Since G is decisive, we must have xPy. But since every member of G' has a veto, we must have $\neg xPy$. Contradiction. □

Given Lemmas 7 and 8, we can establish the following:

Lemma 9. *(Weymark's oligarchy theorem) Let $D \subseteq \mathcal{R}^N$ satisfy SD. Let f be an OAR with domain D satisfying R, QT, WP and IIA. Then there exists a unique oligarchy relative to f.*

Proof. Weymark (1984), Theorem 1. □

Given that $(L_{rat}|_X)^N$ satisfies SD, applying Lemma 9 to the OAR $f|_{(L_{rat}|_X)^N}$, establishes that there exists a unique oligarchy with respect to $f|_{(L_{rat}|_X)^N}$.

We next show that $G = N$. First, note that since the domain is a product space, it is closed under permutations of individuals. Suppose next for contradiction that $G \subset N$. Now let π be any permutation of N that maps some members of G to members of $N \setminus G$ (since $G \subset N$, such a permutation exists). It is straightforward to check that if G is an oligarchy relative to f, then, for any permutation ρ of X such that the domain is invariant under ρ, πG is an oligarchy relative to $\pi\rho f$. Anonymity, however, requires that there exist a ρ such that $f = \pi\rho f$. Since we have chosen π such that $\pi G \neq G$, this contradicts Lemma 8.

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

References

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

Adler, Matthew. 2012. *Well-being and fair distribution: Beyond cost-benefit analysis*. Oxford University Press. 4, 5

Adler, Matthew. forthcoming. Extended preferences. *In:* Adler, Matthew, & Fleurbaey, Marc (eds), *Oxford Handbook of Well-Being and Public Policy*. 2

Arrow, K. 1963. *Social choice and individual values*. 2nd edn. New York: Wiley. 7, 18

Austen-Smith, David, & Banks, Jeffrey. 2000. *Positive political theory I*. University of Michigan Press. 13

Brown, Donald. 1973. *Acyclic choice*. Cowles Foundation Discussion Papers 360. Cowles Foundation for Research in Economics, Yale University. 13

Fleurbaey, M., & Maniquet, F. 2008a. Fair social orderings. *Economic theory*, **34**(1), 25–45. 16

Fleurbaey, Marc. 2007. Social choice and just institutions: New perspectives. *Economics and philosophy*, **23**(1), 15–43. 16

Fleurbaey, Marc, & Maniquet, Francois. 2008b. Utilitarianism versus fairness in welfare economics. *Pages 263–280 of:* Fleurbaey, Mark, Salles, Maurice, & Weymark, John (eds), *Justice, political liberalism, and utilitarianism*. Cambridge University Press. 16

Greaves, Hilary, & Lederman, Harvey. *Extended preferences*. Unpublished MS. 10, 14, 17

Harsanyi, John. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, **63**(4), 309–321. 4

Harsanyi, John C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, **61**(5), 434. 1

Isbell, JR. 1959. Absolute games. *Page 357 of:* Tucker, Albert William, & Luce, Robert Duncan (eds), *Contributions to the theory of games*, vol. 4. Princeton University Press. 15

Mongin, Philippe. 1994. Harsanyi's aggregation theorem: multi-profile version and unsettled questions. *Social Choice and Welfare*, **11**(4), 331–354. 4

Rachels, Stuart. 1998. Counterexamples to the transitivity of better than. *Australasian Journal of Philosophy*, **76**(1), 71–83. 13

Railton, Peter. 1986. Facts and values. *Philosophical Topics*, **14**(2), 5–31. 12

Schick, Frederic. 1971. Beyond utilitarianism. *The Journal of Philosophy*, 657–666. 15

Sen, A. 1986. Social choice theory. *Pages 1073–1181 of:* Arrow, K. D., & Intriligator, M. D. (eds), *Handbook of mathematical economics*, vol. 3. Elsevier. 18

Sen, Amartya. 1970. *Collective choice and social welfare*. Holden-Day. 8, 10

Temkin, Larry. 1987. Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, **16**(2), 138–187. 13

Temkin, Larry. 2014. *Rethinking the good: moral ideals and the nature of practical reasoning*. Oxford University Press. 13

Voorhoeve, Alex. 2014. Review of Matthew D. Adler: Well-being and fair distribution. Beyond cost-benefit analysis. *Social Choice and Welfare*, **42**(1), 245–54. 2

Weymark, John A. 1984. Arrow's theorem with social quasi-orderings. *Public Choice*, **42**(3), 235–246. 10, 18, 19