Economics and Philosophy, **32** (2016) 283–321 © Cambridge University Press
doi:10.1017/S0266267115000486 First published online 11 January 2016
journals.cambridge.org/eap

AGGREGATING MORAL PREFERENCES

MATTHEW D. ADLER*

Abstract: Preference-aggregation problems arise in various contexts. One such context, little explored by social choice theorists, is metaethical. 'Ideal-advisor' accounts, which have played a major role in metaethics, propose that moral facts are constituted by the idealized preferences of a community of advisors. Such accounts give rise to a preference-aggregation problem: namely, aggregating the advisors' moral preferences. Do we have reason to believe that the advisors, albeit idealized, can still diverge in their rankings of a given set of alternatives? If so, what are the moral facts (in particular, the comparative moral goodness of the alternatives) when the advisors do diverge? These questions are investigated here using the tools of Arrovian social choice theory.

Keywords: metaethics, ideal advisor, preference aggregation, social choice, moral preferences

1. INTRODUCTION

A preference-aggregation problem arises when a ranking of alternatives (for short, the 'collective ranking') is constrained to be sensitive to the preferences of some community of individuals. Such a problem can be formally modelled as follows. Each person in the community has her own ranking of the alternatives, which takes the form of a binary relation required to be well-behaved in some sense (e.g. transitive and reflexive). The collective ranking is also required to be a well-behaved binary relation, and its responsiveness to community preferences is captured via an axiom such as weak or strong Pareto.[1] Further axioms relating the collective ranking to the individuals' preferences may also be imposed.

* Duke Law School, 210 Science Drive, Durham, NC 27708, USA. Email: adler@law.duke.edu. URL: https://law.duke.edu/fac/adler/

[1] The weak Pareto axiom requires that one alternative be ranked strictly better than a second by the collective ranking if everyone in the community strictly prefers the first alternative to the second; while the strong Pareto axiom requires this if everyone in the community weakly prefers the first alternative to the second (on weak preference, see note 5), and at

Social choice theory, in turn, is the body of knowledge that gives us insights into preference aggregation as thus modelled – specifically by telling us which combinations of axioms regarding the individual and collective rankings cannot be jointly satisfied (as in Arrow's famous impossibility theorem), and reciprocally by describing the methodologies for arriving at a collective ranking that are consistent with, or entailed by, one or another compossible group of axioms (Sen 1986; Arrow *et al*. 2002, 2010).

Social choice theory, understood as a rich library of such impossibility and possibility results, is a powerful tool for thinking about preference-aggregation problems. But social choice theory is not itself the source of these problems. They come from normative ethics, political or social theory or practices, or metaethics. That is to say, the identification of the relevant alternatives, and of the relevant community of preference holders; the argument that the collective ranking should satisfy the weak or strong Pareto principle with respect to that community; and the specification of additional conditions as appropriate, proceeds from some substantive view regarding the good or the right (normative ethics); an understanding about how political communities or social institutions generally do or should govern themselves (political or social theory or practices); or a higher-order thesis about the kinds of facts, and kinds of knowledge, that ground first-order normative claims (metaethics).

Two kinds of preference-aggregation problems have been intensively studied: *interest aggregation*[2] and *voting rules*.

(1) *Interest aggregation*. This problem is the focus of scholarship on so-called 'social welfare functions' (Adler 2012 and sources cited therein). The social-welfare-function scholarship can be understood (I suggest) as resting on two fundamental normative premises, one concerning moral goodness, the second concerning well-being. The first is what Broome (1991) terms the Principle of Personal Good, namely: If everyone in the community of welfare subjects (those whose well-being has moral weight) is at least as well off with alternative x as compared with alternative y,[3] and some are strictly better off, then x is morally better than y. The second, call it Preferentism about Well-Being, says that someone is at least as well off with an alternative x as compared with y iff[4] she weakly prefers[5] or

least one of them strictly prefers the first alternative. See section 3 for a formal statement of the weak Pareto axiom.

[2] This terminology is used by Sen. See e.g. Sen (1977, 1986: 1129).

[3] It is a further question for interest aggregation what sort of item the 'alternatives' are – in particular, whether they are outcomes or actions (Adler 2012: ch. 7).

[4] 'Iff' means 'if and only if'.

[5] 'Weak preference' is a disjunction of attitudes: An individual 'weakly prefers' one alternative to a second if she either strictly prefers (affirmatively favours) the first, or is indifferent.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

would prefer x to y and this preference is appropriately laundered.[6] Note how the combination of the Principle of Personal Good and Preferentism about Well-Being yields a preference-aggregation problem: the moral ranking of alternatives must satisfy the strong Pareto principle with respect to the laundered preferences of the community of welfare subjects.

(2) *Voting rules*. The problem of voting rule arises whenever some societal institution is designed so that the choice over some set of alternatives is delegated to a community of voters, each with some ranking of those alternatives. (For surveys of the social choice literature on voting rules, see Brams and Fishburn 2002; Pattanaik 2002.) While the interest-aggregation problem arises within the context of a specific, welfarist, moral view – a view that embraces the Principle of Personal Good – the basis for the voting-rule problem is much more heterogeneous. Virtually every moral theory endorses or at least allows for institutions wherein certain decisions are made by voting among a multimember group. Moreover, theorists might well take the existence of such institutions as 'given' as a matter of societal practices – for example, as a bedrock legal feature of democratic polities – without worrying about yet deeper, moral, foundations. The design of voting rules becomes a problem of preference aggregation once we require that alternatives universally favoured by the voters be collectively preferred. Note that Preferentism about Well-Being is no necessary ingredient of *this* preference-aggregation problem. Plausibly, we might think that a well-designed voting rule should satisfy weak or strong Pareto in terms of voters' preferences regardless of whether voters are motivated by their own well-being, moral considerations or other factors.

This article describes a third, less familiar kind of preference-aggregation problem: the problem of *metaethical aggregation*. The context for this problem is quite different from that of interest aggregation, and different again from that of voting rules. So, too, are the preferences being aggregated. Still, the problem of metaethical aggregation – like these two others – is ripe for exploration using the apparatus of social choice theory. Or so I wish to suggest in this article.

The problem of metaethical aggregation arises from a particular view regarding the nature of normative or value facts. 'Ideal advisor'

[6] Laundering is required because it is implausible that satisfying someone's actual preferences suffices to benefit her. A plausible version of Preferentism about Well-Being says: someone is at least as well off with x as y iff she would weakly prefer the first alternative under conditions of rationality, good information and self-interest (Adler 2012: ch. 3). The moral preferences relevant to the ideal-advisor account (see below, section 3) also need to be laundered ('idealized'), but in a different way. Moral preferences must meet an impartiality screen, while (I suggest) someone's preferences ground her well-being only if they are self-interested. However, the details of the laundering component of Preferentism about Well-Being are not relevant here.

(IA) accounts of normative or value facts analyse them in terms of the preferences of a community of advisors (with 'community' used here to include single-membered groups). Such accounts have a prominent place in the literature on metaethics. For example, Roderick Firth (1952) offers a seminal IA account of moral facts. Peter Railton (1995, 2003) provides two, distinct, IA analyses: one for non-moral good, and one for facts about moral goodness. Michael Smith (1994, 1997, 2004, 2005) offers a very general account: on his view, facts about *normative reasons* are facts about the preferences of a community of advisors. Richard Brandt (1979) analyses facts about *rationality* in terms of advisors' preferences. Both Smith and Brandt then handle moral facts as one application of their more general analyses.

IA accounts give rise to a preference-aggregation problem. Why? The thrust of such an account is that normative or value facts are *constituted* by the preferences of some community. Thus, if *everyone* in the community has the relevant sort of preference for x over y, it must be the case that x possesses the normative/value property while y does not (or possesses the normative/value property to a greater degree than y). In short, it seems very plausible that those who tender such accounts would endorse the weak Pareto principle, specified in terms of the relevant preferences.

To be sure, if the community of advisors is a singleton, the preference-aggregation problem just described will be trivial: the normative or value facts will reduce to the preferences of the one person who comprises the community. With a non-singleton community, however, the problem becomes more interesting.

Specifically, the focus of this article will be on IA accounts of *moral* facts. What is the rule that relates the *moral* preferences of some community of advisors to the *moral* facts? I will refer to this problem as that of 'moral aggregation' or 'aggregating moral preferences'. Although the focus *is* moral facts (in particular, facts regarding *comparative moral goodness*), many aspects of the analysis generalize to other types of normative or value facts.

Section 2 is a very brief survey of IA accounts. The bulk of the article then uses a specific cluster of concepts and axioms to analyse the problem of moral aggregation – a cluster that I call the 'quasi-Arrovian setup'. The quasi-Arrovian setup, presented in section 3, has the following features. Each advisor's moral preferences take the form of a *quasiordering* of the set of alternatives: an ordinal ranking which is transitive and reflexive if not necessarily complete. There is then a moral aggregation function M which associates any given profile of ordinal rankings for each advisor, with a moral goodness ranking $\succcurlyeq^M$ of the alternatives. This moral goodness ranking $\succcurlyeq^M$ is itself a quasiordering. And M satisfies axioms of weak Pareto, intermediate Pareto, Anonymity, Independence, and Across-Situation Ranking Consistency (ASRC) – which constrain the

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

moral goodness ranking to conform in some way with the given profile, or to covary in some way with the moral goodness rankings arising from other possible profiles.

Having presented the quasi-Arrovian setup, I proceed to consider the implications for the moral-aggregation problem of different 'domain' axioms: axioms that specify what *are* the possible profiles of moral preferences that the community of advisors might have. One such axiom (call it 'Ordering Convergence') stipulates that all advisors, once appropriately idealized, necessarily have the very same ranking of the set of alternatives. Some IA theorists have made statements about the convergence of moral preferences which might be read as endorsing Ordering Convergence. With this premise in place, the problem of aggregating moral preferences is readily solved. But should we indeed accept Ordering Convergence? This question is addressed in Section 4.

Section 5 continues to work within the quasi-Arrovian setup, but now relaxes Ordering Convergence. It considers the implication for moral aggregation of permitting profiles in which advisors have non-identical rankings of the set of alternatives. At the other extreme from Ordering Convergence is 'Universal Domain': any logically possible profile is included within Φ, the set of such profiles upon which the moral aggregation function M is supposed to operate. The supposition of Universal Domain together with weak Pareto, intermediate Pareto, Anonymity and Independence dramatically narrows the possibilities for moral aggregation. There is only one moral aggregation function M that satisfies this combination of axioms[7] – a function I will refer to as the 'Unanimity Rule'.

Section 5 also briefly considers domain assumptions intermediate between the two extremes of Ordering Convergence and Universal Domain. In this intermediate territory, a wider range of functional forms for M are consistent with the axioms just mentioned – for example, majority vote. However, adding ASRC plus an additional mild axiom forces us back to the Unanimity Rule.

Section 6 addresses the possibility of moving beyond the quasi-Arrovian set-up: by relaxing Independence or, more radically, both Independence and ASRC. As we shall see, ASRC requires that the moral goodness ranking of a set of alternatives be wholly determined by advisors' ordinal preferences. Independence goes further and says that the comparative moral goodness of two alternatives depends upon the advisors' pairwise ordinal preferences with respect to the two. Why embrace these axioms? What would moral aggregation look like if they were dropped? Section 6 undertakes a preliminary discussion of these topics.

[7] This assumes that the cardinality of the set of alternatives being ranked is greater than two.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

The problem of metaethical aggregation is much too large to be comprehensively addressed in a single article. The ambition, here, is simply to bring the problem into clearer focus. Curiously, the problem has not been discussed in a sustained way by metaethicists. Within social choice theory, a small body of scholars – writing about so-called 'extensive social choice' – have addressed the topic of aggregating moral preferences. (Ooghe and Lauwers 2005 and sources cited therein). More of this work needs to be done. Metaethical aggregation is a problem that can be illuminated using the tools that Arrow and his successors have given us.

2. WHY IDEAL-ADVISOR ACCOUNTS?

I will use the term 'cognitivism' as a shorthand for the family of metaethical views that combine a propositional semantics for normative statements, with a non-error-theoretic view of normative facts (Smith 2005; Schroeder 2010; Miller 2013). According to 'cognitivists', (1) a normative statement has the (sole) semantic content of *asserting* that some normative fact obtains – that some object, e.g. an action, outcome or person, has a normative property; and (2) indeed some facts of this sort do obtain, so that normative statements can be true. Cognitivism is to be contrasted with views according to which normative statements have partly or wholly[8] non-assertoric content (expressing the speaker's emotions, commitments or plans); and with sceptical views that see normative statements as asserting the existence of facts that never obtain, so that normative statements are always false.

Ideal-advisor ('IA') views are a subfamily within cognitivist accounts. IA views analyse normative facts in terms of the preferences of some (single- or multimembered) community of advisors, preferences that are idealised in some way by imposing conditions of good information, rationality, etc. on the advisors. By 'normative facts', I mean facts that involve normative or value properties, such as moral goodness or rightness, well-being, being a reason for, etc.

Such views have figured prominently in contemporary metaethics. Here are perhaps the most famous examples.

(1) *Firth.* Firth proposes that the statement 'x is morally right' be understood to mean 'any ideal observer would approve x', where an ideal observer is an otherwise normal person who is omniscient with respect to non-moral facts; has unlimited powers of imagination; and is disinterested, dispassionate and consistent (Firth 1952).

[8] Traditional expressivists see normative statements as wholly non-assertoric, while hybrid expressivists may see them as partly non-assertoric: *both* asserting a fact, *and* expressing an emotion, commitment, plan, etc. (Ridge 2006, 2014; Schroeder 2009).

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

(2) *Railton*. Railton in a series of writings has advanced two distinct IA views, one for well-being, the second for moral good. (a) A given person's well-being ('non-moral good') is determined by what *she* (if idealized) would want herself to do or want in her actual, non-ideal circumstances.

> Give to an actual individual *A* unqualified cognitive and imaginative powers, and full factual and nomological information about his physical and psychological constitution, capacities, circumstances, history, and so on. *A* will have become *A*+, who has complete and vivid knowledge of himself and his environment, and whose instrumental rationality is in no way defective. We now ask *A*+ to tell us not what *he* currently wants, but what he would want his nonidealized self *A* to want – or, more generally, to seek – were he to find himself in the actual condition and circumstances of *A*. (Railton 2003: 11)

(b) Railton suggests that '*X* is morally wrong' be understood to mean:

> 'We the people (i.e., people in general, including the speaker) would disapprove of allowing *X* as part of our basic scheme of social cooperation were we to consider the question with full information from a standpoint that considers the well-being of all affected without partiality.' (Railton 1995: 69)

(3) *Smith*. Smith offers a general account of 'normative reasons' in terms of the preferences of fully rational agents.

> I have been arguing that the truth of a normative reason claim requires a convergence in the desires of fully rational agents. However note that the convergence required is not at the level of desires about how each such agent is to organize her own life in her own world. … The convergence required is rather at the level of their hypothetical desires about what is to be done in the various circumstances in which they might find themselves. (Smith 1994: 173; see also Smith 1997, 2004: ch. 10, 2005)

Moral reasons, for Smith, are in turn normative reasons with the additional special features (whatever exactly they may be) that are picked out by our platitudes about the domain of the moral (for example, that moral considerations involve important interests of persons other than the actor – thus that I have moral and normative reason to refrain from killing, but a non-moral normative reason to relax after a hard day at work) (Smith 1994: 183–4).

(4) *Brandt*. Brandt, like Smith, offers a very general IA account. For Smith, the analysandum is the property of being a 'normative reason'; for Brandt, it is the property of being 'rational'. Some individual's action is 'rational', according to Brandt, if the action is optimal in light of her underlying desires *and* those desires themselves are 'rational' in the sense of surviving 'cognitive psychotherapy': a 'process of confronting desires with relevant information, by repeatedly representing [that information],

in an ideally vivid way' (Brandt 1979: 113). Brandt then uses his conception of 'rational', together with the concept of a 'moral code' (roughly, norms regulating feelings of guilt), to define 'morally wrong'. He sets forth two possibilities, here:

> I suggest ... that we assign 'is morally wrong' the descriptive meaning 'would be prohibited by any moral code which all fully rational persons would tend to support, in preference to all others or to none at all, for the society of the agent, if they expected to spend a life-time in that society …'
>
> This suggested descriptive meaning is only one possibility. There are others, and perhaps better ones. For instance, if it turned out that not all fully rational persons would tend to support the same moral code in preference to all others, it would be better if 'morally wrong' were given a slightly different descriptive meaning, perhaps 'would be permitted by any moral code which I [or you], if I [you] were fully rational, would tend to support, in preference to all others or to none at all …' (Brandt 1979: 194; cf. Brandt 1955)

Metaethics is a field of controversy, with every account hotly disputed; but IA accounts, clearly, are *prima facie* appealing and have been embraced by some very sophisticated metaethicists. Why? Like all cognitivist accounts – and by contrast with views that see normative statements as doing something other than asserting facts – IA approaches are not troubled by the 'Frege-Geach' problem (Schroeder 2010; Miller 2013). Like other non-error-theoretic approaches, they do not see ordinary normative discourse as radically misguided – as making statements that are never true. By contrast with non-naturalist cognitivism – positing the existence of normative facts that are qualitatively distinct from the facts that figure within physical, psychological and social science – IA views are 'naturalist' in reducing normativity to a certain kind of psychological fact, namely facts about what an observer or observers in a certain informational and emotional state would desire.

Finally, facts about ideal advisors conform to two truisms about normative facts (or at least are well suited to conform to these truisms, given an appropriate specification of the community of advisors and the idealizing conditions). On the one hand, individuals are not infallible about normative facts (if such exist). On the other hand, normative facts, somehow, are tied to motivation: If I genuinely judge that x is morally good (for example), then I am motivated to pursue x. It is initially difficult to see how *any* fact could have both these features, but indeed facts about the preferences of ideal advisors could. Since a given person's preferences need not be ideal, she is *not* infallible about what *ideal* advisors would want. However, the belief that someone with preferences like my own, thinking calmly and with good information, would want me to x, will tend to motivate *me* towards x, notwithstanding that I lack good information,

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

and even if my current deliberational state is less than fully ideal – just as my belief that someone with excellent information and cognitive capacities would believe *p* will tend to cause me (imperfectly informed as I may be) to believe *p*.

IA accounts (to be plausible) should not be understood as giving the *meaning* of the concept or term 'moral' or other normative or value concepts or terms. The statement '*x* is morally good' does not have the same semantic content as '*x* would be approved by a community of ideal advisors'. Rather (to be plausible) such accounts should be understood as identifying the natural properties or property clusters referred to by such terms. For naturalist cognitivists, discovering such properties is quasi-scientific – a matter *both* of understanding what we mean by our normative concepts, the 'truisms' or 'platitudes' constitutive of them, *and* of establishing which physical, psychological or social properties are suitable to be the naturalistic basis for the normative concepts (given how such properties function in light of physical, psychological or social laws). Such discovery is at least roughly analogous to the discovery that the physical substance referred to by the term 'water' is the molecule H_2O – a discovery that combines conceptual analysis (the truism that 'water', as understood by anyone who grasps the concept, is a liquid, if fresh will quench thirst, is found in lakes, rivers and oceans, etc.) and empirical investigation (Jackson 1998).[9]

Let me now suggest a differentiation between two types of IA accounts: 'singleton' accounts, where the community of advisors whose preferences give rise to normative/value facts is guaranteed to have a single member; and 'multimember' accounts, where this is not the case (i.e. the community can have multiple members).

A cross-cutting distinction is between relativist and non-relativist (absolutist) IA accounts (Firth 1952; Smith 1994, 2004: ch. 10). Relativists with respect to some type of normative/value property posit that statements regarding the property can hold true relative to one speaker or group, but false relative to another. This is a familiar idea from the literature on moral relativism (Harman and Thomson 1996). Extrapolating from this literature, one can imagine a type of IA account whereby the truth of the assertion of a particular normative/value fact is relativized to one or another community of advisors. By contrast, an absolutist IA account posits a single community of advisors: A particular normative/value fact either obtains or not, in virtue of the preferences of this single community, and an assertion of this normative/value fact is either true absolutely or not.

[9] Although I find it most plausible that the nexus between moral goodness and advisors' approvals is *a posteriori*, the IA theorist might disagree – claiming that the connection is *a priori*. This issue is orthogonal to the analysis that follows.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

Note now that both absolutist and relativist IA views can assume either the singleton or multimember form. An absolutist IA account might analyse a particular kind of normative/value fact in terms of all persons' idealized preferences (multimember); in terms of God's preferences (singleton); or, perhaps, in terms of the preferences of a particular person (singleton), if the fact concerns the actions or well-being of that particular person. For example, the fact that a particular action of Jane's would be morally wrong could be reduced to Jane's preferences. A relativist IA view might posit that assertions of normative/value fact are true relative to communities which can have multiple members (this counts as a multimember IA account on the definition offered above). For example, moral facts might be relativized to one or another community consisting of individuals in the same society. Alternatively, a relativist view could stipulate that the communities relative to which normative/value statements are true must be singleton.

That IA accounts can assume either the singleton or multimember form is evident even from my brief descriptions of Railton's, Brandt's, Smith's and Firth's views. Railton offers a singleton account of well-being, but not of moral rightness. Whether x is good for Sue's well-being depends upon what Sue (if idealized) would want; whether x is morally right depends upon what some plural 'we' would want.[10] Brandt, quite visibly, oscillates between a singleton and multimember account of moral rightness. Brandt's singleton proposal is: x is morally right relative to a given individual iff she would rationally want a moral code that approves x. Brandt's multimember proposal is: x is absolutely (non-relatively) morally right iff all of us would rationally want a moral code that approves x.

Smith offers a thunderous argument for an absolutist and multimember conception of normative reasons. Facts about normative reasons are made true by the convergent, fully rational desires of all persons.

> On the non-relative conception of normative reasons … it is desirable that p in C just in case we would all desire that p in C if we were fully rational. …
>
> On the relative conception, … matters are quite different …. 'It is desirable that p in C' as assessed from A's perspective is true if and only if A would desire that p in C if A were fully rational, 'It is desirable that p in C' as assessed from B's perspective is true if and only if B would desire that p in C if B were fully rational, and so on and so forth …

[10] Railton suggests that this view is closer to absolutism than relativism. '[A]lthough it is a first-person view … it is a first-person plural, reaching out to 'people in general' rather than any specific group of persons' (Railton 1995: 70–71).

> [I]f normative reasons were indeed relative, then mere reflection on that fact would suffice to undermine their normative significance. For on the relative conception it turns out that, for example, the desirability$_{me}$ of some consideration, p, is entirely dependent on the fact that *my* actual desires are such that, if *I* were to engage in a process of systematically justifying *my* desires, weeding out those that aren't justified and acquiring those that are, a desire that p would be one of the desires that *I* would end up having. But what my actual desires are to begin with is, on this relative conception of reasons, an entirely *arbitrary* matter .... I might have had any old set of desires to begin with, even a set that delivered up the desire that not p after a process of systematic justification. (Smith 1994: 166–7, 172)

In short: 'Acts are right or wrong depending on whether, notwithstanding any contingent and rationally optional culturally induced differences in our actual desires, we would all desire or be averse to the performance of such acts if we had a set of desires that was maximally informed, coherent, and unified' (Smith 2004: 205). Firth's view is also explicitly 'absolutist' (Firth 1952: 320) and multimember: it construes ethical statements as 'assertions about the dispositions of all [hypothetical] beings of a certain kind', namely all ideal observers (1952: 320).

It is well beyond the scope of this article to address the pros and cons of singleton versus multimember IA accounts with respect to moral goodness or rightness, well-being, or other normative/value properties. Rather, the remainder of the article will train its attention on the non-trivial problem of preference aggregation that a multimember IA account faces. The account either needs to show that the community or communities of advisors will be such that members will have identical preferences; or, alternatively, to explain what the normative facts will be when community members have non-identical preferences.

Let us now analyse this problem in detail – focusing specifically, as promised, on moral properties. How do moral facts arise from a multimember community of advisors?

3. AGGREGATING MORAL PREFERENCES: A QUASI-ARROVIAN FRAMEWORK

What follows is *one* cluster of concepts and axioms that might be used to model the problem of aggregating moral preferences. This set-up is similar, in formal structure, to that introduced by Arrow (1951, 1963) in *Social Choice and Individual Values* – but not identical, and so I call it 'quasi-Arrovian'.[11]

[11] An important difference is that Arrow requires both the individual inputs and the collective output to be *complete* rankings. By contrast, I allow for incompleteness: both inputs and output are quasiorderings, possibly incomplete. Note that the Unanimity Rule (section 5 below) is not guaranteed to yield a complete moral goodness ranking – indeed

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

(1) We are given some set of items $\mathbf{S} = \{x, y, z \ldots\}$. The members of any such set (of an appropriate kind) have moral goodness or rightness properties. I focus on moral betterness (comparative goodness), and leave aside absolute goodness and moral rightness.[12]

S, then, is any set of things each pair of which are appropriate for assessment as comparatively morally better or worse[13] – for example, different token actions that a particular agent might undertake at a particular moment in time, or different possible worlds, or different institutions or practices that might exist in a society. The totality of facts about the comparative goodness of the items in **S** can be summarized (I assume) by a moral quasiordering $\succcurlyeq^{M\text{-}\mathbf{S}}$ of **S**. I will call **S** a 'situation', and will refer to the items in **S** as 'alternatives'.

Let #**S** denote the cardinality of **S**. Unless otherwise stipulated, #**S** can take any value: **S** can have any finite number of elements, or be countably or uncountably infinite. One part of the analysis below does assume that #**S** > 2, and that will be stated explicitly.

(2) I assume a finite multimember set **N** of advisors. $\mathbf{N} = \{1, \ldots k, \ldots N\}$, with k used below to denote a particular advisor, and $N > 1$. Which persons and, thus, advisors will exist is, of course, not independent of what we now do – which in turn may depend upon the goodness of actions and outcomes, the very thing that the community of advisors' views is meant to be determining – but to simplify I ignore the possibility of a 'variable population' of advisors. Moreover, I assume that each advisor is an actual person, who exists in the actual world. The actual facts of comparative moral goodness concerning a given situation are, plausibly, fixed by the preferences of actual, not counterfactual advisors. Below, we will consider the kind of counterfactual moral goodness that arises from the non-actual preferences of the actual advisors; but the doubly counterfactual moral goodness arising from the non-actual preferences of non-actual advisors is beyond the scope of the model.

will produce noncomparability between two alternatives if even one advisor strictly prefers one and another strictly prefers the other. This is why my characterization of this rule is consistent with Arrow's impossibility result.

A second difference is that my setup adds an across-situation consistency requirement (ASRC).

[12] There may be significant differences between IA models of absolute goodness or moral rightness as opposed to moral betterness – for example, regarding whether the inputs and output are required to be binary (good/not-good; right/not-right) and with respect to Independence (see below, section 6). These issues cannot be pursued here.

[13] It might be argued that certain pairs of items fail to meet a threshold test of assessability with respect to comparative moral goodness – and that such failure is different from the incomparability of assessable items. For example, it might be seen as nonsensical to ask whether the number 17 is better, worse, equally good, or noncomparable to a particular act. I add the caveat 'appropriate for assessment' to allow for such a threshold test.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

The membership of **N** depends on the particular IA view being modelled. It is appealing to think that moral facts are true absolutely, and true in virtue of what all persons (under idealized conditions) would prefer. This is indeed the position of Smith, Firth and Railton. On this view, **N** consists of all existing persons. Alternatively, **N** might be a proper subset of existing persons – for example, all existing *human* persons, or (on a relativist view) all persons within a given society. Except at one juncture, noted below, these differences in the specification of **N** will not matter to my analysis.

(3) I assume that advisors' preferences are the grounding for moral goodness just in case they meet the following conditions: the preferences must be 'rational', 'well-informed' and 'impartial'. (As a shorthand, I will refer to these conditions, collectively, as the 'idealizing' conditions, and preferences meeting them as 'idealized'.) The conditions are deliberately stated at a high level of generality – so as both to capture the features of advisors' preferences that are clearly demanded by our platitudes about morality, and yet to leave open for further debate and investigation the precise psychological content of 'rational', 'well-informed' and 'impartial'.

Critics of an IA account of moral facts might argue that these idealizing conditions are themselves, to some extent, normative – and that this undercuts the naturalist ambition of the approach, namely to reduce moral facts to ordinary, non-normative facts. But the proponent of the account has an answer. She can concede that 'impartial', 'well-informed' and/or 'rational' *are* normative – concepts that we use in deliberating about what ought to be done – but say that these concepts, *like the concept of moral goodness itself*, refer to non-normative properties. Identifying the psychological attributes corresponding to 'impartial', 'well-informed' and 'rational' involves the very same mixture of conceptual analysis and empirical investigation by which we establish the natural basis for moral goodness. The proponent can say: there is a fact of the matter regarding what psychological state someone is in when he is 'impartial', 'well-informed' and 'rational' – given the nature of various psychological states, as per the laws of psychology, plus the platitudes about 'impartiality', etc., that anyone using *these* normative concepts grasps. In turn, the idealized advisors are individuals in that psychological state; and moral goodness depends upon what such advisors prefer.

(4) I assume that the idealized preferences of each given advisor, k, over **S** take the form of a quasiordering $\succcurlyeq^{k\text{-}\mathbf{S}}$. Thus I allow for an advisor's moral preferences to be incomplete, but not for intransitivities. This is, in effect, a further specification of the term 'rational'. I assert that rationality requires, at least, formal coherence in the sense of having a transitive, although perhaps incomplete, ranking. On the one hand, powerful arguments have been mounted to show that rational moral agents are permitted to rank some alternatives as noncomparable:

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

x neither morally better than y, nor morally worse, nor equally good (Adler 2012: 43–4). The rational permissibility of intransitive preferences is on much shakier ground (Voorhoeve 2013).

The relations of strict preference, weak preference, indifference and noncomparability corresponding to $\succcurlyeq^{k\text{-}\mathbf{S}}$ are denoted, respectively, as $P^{k\text{-}\mathbf{S}}$, $R^{k\text{-}\mathbf{S}}$, $I^{k\text{-}\mathbf{S}}$ and $NC^{k\text{-}\mathbf{S}}$.

(5) Let $\mathbf{P^S}$ denote a profile of moral preferences ($\succcurlyeq^{1\text{-}\mathbf{S}}$, …, $\succcurlyeq^{N\text{-}\mathbf{S}}$). A given $\mathbf{P^S}$ is 'admissible' if it is *possible* (in the relevant sense) for advisors to meet the idealizing conditions and have this profile of preferences. Let $\Phi^{\mathbf{S}}$ denote the set of admissible profiles. A moral aggregation function or rule $M^{\mathbf{S}}$ maps each admissible $\mathbf{P^S}$ onto a moral goodness quasiordering $\succcurlyeq^{M\text{-}\mathbf{S}}$ of $\mathbf{S}$, with $P^{M\text{-}\mathbf{S}}$, $R^{M\text{-}\mathbf{S}}$, $I^{M\text{-}\mathbf{S}}$ and $NC^{M\text{-}\mathbf{S}}$ now denoting the relations of strict and weak preference, indifference and noncomparability corresponding to $\succcurlyeq^{M\text{-}\mathbf{S}}$. Because $\succcurlyeq^{M\text{-}\mathbf{S}}$ is a *moral goodness* quasiordering, we can equally well say that these denote, respectively, the relations of moral betterness, being at least as good as,[14] equal goodness, and moral noncomparability corresponding to $\succcurlyeq^{M\text{-}\mathbf{S}}$.

To avoid clutter, I will sometimes drop the '$\mathbf{S}$' superscript on $\succcurlyeq^{k\text{-}\mathbf{S}}$, $\succcurlyeq^{M\text{-}\mathbf{S}}$, $\mathbf{P^S}$, $\Phi^{\mathbf{S}}$, $M^{\mathbf{S}}$ and the R, I, P, and NC relations associated with $\succcurlyeq^{k\text{-}\mathbf{S}}$ and $\succcurlyeq^{M\text{-}\mathbf{S}}$. However, where I do so, the reader should keep in mind that this superscript is implicit. Moral preferences and profiles thereof are rankings of a particular set $\mathbf{S}$ of items, and the associated set of possible profiles is also indexed to $\mathbf{S}$.

I will use $\succcurlyeq^{M}{}_{\mathbf{P}}$ to denote $M^{\mathbf{S}}(\mathbf{P^S})$ – the moral goodness ranking onto which profile $\mathbf{P^S}$ (for short, $\mathbf{P}$) is mapped by function $M^{\mathbf{S}}$ (for short, M). And I will use $\succcurlyeq^{k}{}_{\mathbf{P}}$ to denote the kth element of a given profile $\mathbf{P}$. Again so as to simplify notation, I drop the $\mathbf{P}$ subscripts on $\succcurlyeq^{k}{}_{\mathbf{P}}$ and $\succcurlyeq^{M}{}_{\mathbf{P}}$ where it is clear from context which profile is intended.

(6) What *does* 'possible' mean, here?

The content of the set of possible preference profiles (here, $\Phi^{\mathbf{S}}$, the set of profiles for situation $\mathbf{S}$) is a topic of great importance to social choice theory, typically referred to as the question of 'domain' definition. (Note that $\Phi^{\mathbf{S}}$ is the *domain* for the function $M^{\mathbf{S}}$, hence the term.) Aggregation functions that are ruled out when combining certain axioms with a universal domain of preferences may no longer be precluded if the domain is restricted. Standard surveys of the field shine a spotlight on the question of domain definition (Sen 1986; Campbell and Kelly 2002; Gaertner 2002). However, there appears to have been little systematic discussion by social choice theorists of the more philosophical topic: what is the appropriate *sense* of the term 'possible' used to specify the domain of preferences?

[14] Item x is morally 'at least as good as' y iff it is either better than y or equally good as y.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

Philosophers of modality distinguish between different *kinds* of possibility: for example, logical, conceptual, epistemic and metaphysical (Divers 2002: ch. 1; Gendler and Hawthorne 2002; Huemer 2007; Egan and Weatherson 2011; Kment 2012). These distinctions are themselves contested, and it is well beyond the scope of this article to address them in detail. It will suffice to say this: a state of affairs is metaphysically possible just in case there is some possible world in which the state of affairs obtains. Metaphysical possibility is fixed by the modal facts – specifically, which worlds are genuinely members of the total collection of possible worlds – just as ordinary facts are fixed by what is true about the particular possible world that is the actual world. Metaphysical possibility is to be contrasted with what we believe to be true in some possible world, what our concepts seem to us to allow or preclude, and what logic allows or precludes. Logical impossibility is fixed by the logical connectives, and implies but is not implied by metaphysical impossibility: it is logically possible that a three-sided figure has four angles, but metaphysically impossible. Goldbach's conjecture, that all even numbers can be expressed as the sum of two primes, has never been proven or disproven. It seems conceivable that either the conjecture or its negation is true, and it is reasonable to believe either. But mathematical truths are true in all possible worlds, and so either the conjecture is metaphysically necessary and the negation impossible, or vice versa.

It is *metaphysical* possibility, I suggest, that determines the content of $\Phi^{\mathbf{S}}$ (for short, Φ). A given profile of moral preferences $(\succcurlyeq^1, \ldots, \succcurlyeq^N)$ is a member of Φ just in case there is truly some possible world in which advisor 1 meets the idealizing conditions and has the preferences over **S** described by $\succcurlyeq^1$ *and* advisor 2 meets the idealizing conditions and has the preferences over **S** described by $\succcurlyeq^2$ *and* … advisor N meets the idealizing conditions and has the preferences over **S** described by $\succcurlyeq^N$. Why conceptualize Φ as the set of metaphysically possible profiles with respect to **S**, rather than the set of logically, conceptually or epistemically possible profiles? Recall the ambition of the IA account of normative properties: to show how these properties, such as moral goodness, reduce to natural facts, specifically facts about the (in some way) idealized preferences of advisors. The moral aggregation function M is a kind of law (or, even better, *bridge principle*) of nature, mapping certain naturally occurring features of the world (advisors' preferences) onto the facts about moral goodness that these features determine. A putative M, to be a genuine bridge principle, must be capable[15] of mapping any **P** that might naturally occur, onto a goodness ranking $\succcurlyeq^M$. But whether a given **P** could naturally occur is a question of *metaphysical possibility*: a question about

[15] 'Capable' within the limits of whichever axiomatic requirements there are (weak Pareto, intermediate Pareto, etc.) that govern the aggregation of advisors' preferences.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

which arrangements of advisors' preferences are to be found somewhere in the total collection of possible worlds, given the nature of preferences and the nature of the idealizing conditions for the advisors.

To be more concrete: the advisors' idealized preferences have certain properties. On the IA account of moral facts, these are the properties picked out by the totality of our truisms about normativity and morality and about motivation, preferences and other features of human psychology. These are the properties I am referring to with the terms 'rational', 'well informed' and 'impartial'. *Which* properties these are is a matter for empirical investigation; that may not be clear to us *ab initio*. And it may turn out to be metaphysically impossible for a particular profile **P*** of rankings of some **S** to have all of these properties. In no possible world does a group of 'well informed', etc., advisors have *that* collection of rankings. If so, **P*** should not be in Φ. It is not the function of the natural bridge principle *M* to operate on arrangements of preferences that could never actually occur, just as it is not the function of an ordinary law of nature (a causal regularity) to determine what an impossible state of affairs or event would causally produce.

To be sure, IA theorists are not omniscient, and will therefore have different beliefs about what is metaphysically possible. The formal apparatus here is a partly specified model of the determination of moral facts. It will be useful only to those theorists who already have certain beliefs about such determination[16]; and, among those theorists, will be further specified depending on what further beliefs they hold. In particular, as we will see in section 4, differing beliefs among IA theorists about what is 'rational' will yield quite different beliefs about the metaphysical possibility of a community of advisors meeting the idealizing conditions yet diverging in their rankings of a set of alternatives. A theorist's beliefs inevitably shape her modelling of the phenomenon of interest (here, the determination of moral facts). But it hardly follows that beliefs, thoughts or concepts are the target of the model. A profile $(\succcurlyeq^1, \ldots, \succcurlyeq^N)$ belongs in Φ, according to a given IA theorist, if she believes the profile to be metaphysically possible. This is certainly *not* the same as saying that Φ contains all profiles which are epistemically possible, or which it is reasonable to believe are metaphysically possible!

(7) In short: *M* maps each metaphysically possible (henceforth 'possible') profile **P** of moral preferences over the set **S** of items, onto a goodness quasiordering of those items. This ranking, denoted $M(\mathbf{P})$ or

[16] For example, that moral facts are fixed by the preferences of advisors that are (inter alia) rational in the sense of being transitive – as reflected in the requirement of the quasi-Arrovian setup that advisors' rankings be quasiorderings.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

$\succcurlyeq^{M}_{\mathbf{P}}$, specifies what the facts about moral goodness *would be*, were the advisors to have the moral preferences set forth in profile **P**.

But where are the *actual* facts of moral goodness in this picture? The answer is (relatively) straightforward. Let **P*** be the *actual* moral preferences of the advisors. More precisely, since we are imposing idealizing conditions on preferences that advisors might not actually meet, $\mathbf{P^*} = (\succcurlyeq^{1^*}, \ldots, \succcurlyeq^{N^*})$, where $\succcurlyeq^{k^*}$ are the preferences of advisor k in the nearest possible world – nearest to the actual world – in which all the advisors meet the idealizing conditions. For short, I will refer to **P*** as the advisors' 'actual idealized' preferences. Actual moral goodness is given by the ranking, $M(\mathbf{P^*}) = \succcurlyeq^{M}_{\mathbf{P^*}}$, onto which this profile is mapped by the moral aggregation function M.[17]

One question that can be raised at this point is why M should be expected to operate on metaphysically possible profiles of idealized preferences that are not the advisors' actual idealized preferences. Why not see the moral aggregation function for a given situation as having a singleton domain, namely what advisors prefer in the nearest possible world in which they meet the idealizing conditions?[18]

Suffice it to say that natural laws support counterfactuals (the chemical law that salt dissolves in water tells us what would happen if this packet of salt were dropped in that glass), and also that moral discourse intelligibly asks questions regarding counterfactual moral requirements or values, namely what would be morally right or morally better if certain facts were otherwise. Both points suggest that Φ should include counter-actual as well as actual idealized profiles, at least to some extent. M, if genuinely the natural law by which moral facts arise from advisors' preferences, should not only tell us that the actual comparative moral goodness of the items in **S** is given by $M(\mathbf{P^*})$ – with **P*** the advisors' actual idealized preferences – but also that the moral goodness of the items in **S** would be given by $M(\mathbf{P})$, were **P** instead of **P*** to be the

[17] If moral goodness is truly a kind of fact, then it can be modally qualified, like physical, psychological and social facts. We can differentiate actual moral goodness from possible moral goodness. And what (for the IA) theorist could be the difference between actual and possible moral goodness, other than the difference between actual and possible advisor preferences?

A robust view of rationality would have it that idealized advisors never differ in their preferences. See below, section 4. A yet more robust variation would have it that advisors in all possible worlds have the very same convergent preferences. In other words, Φ contains but a single profile **P**. Such a view is highly contestable, but in any event does not undermine what I have just said about actual goodness. On such a view, actual moral goodness *is* still determined by the actual profile of advisors' preferences. However, such profile no longer varies across worlds.

[18] This is just to ask whether moral aggregation should be modelled using a 'single-profile' or 'multi-profile' approach. The choice between the two approaches has been much debated by social choice theorists (Roberts 1980; Fleurbaey and Mongin 2005).

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

advisors' preferences in the nearest possible world where the preferences meet the idealizing conditions. Moreover, by including **P** as well as **P*** in its domain, the natural law M serves to *explain* why actual moral goodness is given by $M(\mathbf{P^*})$ – namely, that actual idealized preferences are **P*** *rather than* **P**. Conversely, if M does not operate on counterfactual but metaphysically possible profile **P**, then it is indeterminate what the moral facts would be, were the advisors' idealized preferences in the nearest possible world to be **P**. But what grounds does the IA theorist have to believe in such indeterminacy? [19]

(8) What, now, are the constraints on M?

First, M must satisfy the axioms of *Pareto indifference* and *weak Pareto* with respect to all admissible profiles of advisors' moral preferences. If all advisors are morally indifferent between two items, the two are equally morally good. If all advisors strictly prefer one item to a second, the first is morally better. That is: (a) If $x\ I^k_{\mathbf{P}}\ y$ for all k, then $x\ I^M_{\mathbf{P}}\ y$. (b) If $x\ P^k_{\mathbf{P}}\ y$ for all k, then $x\ P^M_{\mathbf{P}}\ y$.

A key feature of the IA account is that the preferences of the advisors in the stipulated community, under idealizing conditions, are *constitutive* of the moral facts. It is not possible for the advisors, under those conditions, to be collectively mistaken about the facts. This constitutive role of the advisors' preferences is captured by the weak Pareto and Pareto indifference axioms. On other metaethical views, such axioms would be problematic: there may be *no* community of persons relative to which the moral goodness ranking needs to satisfy weak Pareto and Pareto indifference. For example, if moral facts are constituted by God's ranking (with God a superhuman being who is not a person), the unanimous preferences of any community of idealized persons might diverge from moral goodness. Similarly, if goodness is a natural property with a particular explanatory role, then (perhaps) any community of persons could be collectively mistaken about how that property behaves – just as we (perhaps) might all be mistaken about the laws of physics.

I do not build in *strong Pareto* as a foundational axiom of the IA approach. Strong Pareto says: If $x\ R^k_{\mathbf{P}}\ y$ for all k, and $x\ P^j_{\mathbf{P}}\ y$ for some j, then

[19] Admittedly, natural laws might be expected to operate only on counterfactuals that are not 'wildly' unrealistic. Since the formal apparatus builds in the premise that the advisors are actually existing persons, this qualification seems less applicable. Included in Φ are all possible profiles of idealized preferences regarding **S** that this particular group of actual persons might have – both their preferences in the nearest possible world in which they all meet the idealizing conditions, and their idealized preferences in more distant worlds. Yet more remote preferences belonging to a different community of advisors are not in Φ; and so 'wild' counterfactuals arising by supposing a change in this community are already excluded. If the reader is not satisfied by this structural exclusion, she might add an additional qualification that $(\succcurlyeq^1, \ldots, \succcurlyeq^N)$ is in Φ only if the world in which the advisors have those preferences is not 'too far' from the actual world.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

$x\ P^M_{\mathbf{P}}\ y$. While the moral weak Pareto and Pareto indifference axioms see the unanimous strict preference or indifference of the advisors as sufficient for, respectively, moral betterness or equal goodness, strong Pareto (to the extent it goes beyond weak Pareto) identifies a condition of *disagreement* among the advisors as nonetheless sufficient for moral betterness. It is not clear that the IA theorist should be committed to this from the start.

To be sure, it may emerge that a moral aggregation function entailed by *other* axioms ends up satisfying the strong Pareto axiom. Indeed, as we shall see, this is true of the Unanimity Rule. In other words, it may turn out to be the case that the IA theorist is logically required to endorse the strong Pareto axiom as the result of *other* principles that she finds immediately appealing. But because strong Pareto does not directly express a key commitment of the IA approach – in particular, that advisors' preferences are constitutive of moral facts – I do not incorporate strong Pareto into the quasi-Arrovian setup itself.

By contrast, strong Pareto has direct appeal in the context of interest aggregation (Adler 2012: ch. 5). And perhaps it also does for voting rules. The absence of a foundational commitment to strong Pareto is one important difference between the metaethical variant of preference aggregation and these other variants (cf. Rabinowicz 2015).

A fourth kind of Pareto principle should now be mentioned. So as to avoid terminological confusion, let's call this '*intermediate Pareto*': If $x\ R^k_{\mathbf{P}}\ y$ for all k, then $x\ R^M_{\mathbf{P}}\ y$.

Like weak Pareto and Pareto indifference, intermediate Pareto is intuitively attractive. Imagine that all advisors weakly prefer x to y. This means, first, that no advisor strictly prefers y to x. But then the moral aggregation function should surely *not* rank y as better than x. To do so would contradict the universal view of the advisors. Observe, second, that if all advisors weakly prefer x to y then no advisor counts the two as noncomparable. But then, surely, the moral aggregation function should not rank the two as noncomparable – again, this would be to contradict the universal view of the advisors.[20]

Note that intermediate Pareto implies Pareto indifference. I therefore describe the quasi-Arrovian setup as including the axioms of weak Pareto and intermediate Pareto rather than (redundantly) weak Pareto, intermediate Pareto and Pareto indifference.

[20] I am indebted to an editor of *Economics and Philosophy* for this defence of intermediate Pareto. I find it to be quite powerful. Some readers may, however, resist it – seeing moral noncomparability as a plausible default when advisors disagree. Consider the case where some advisors strictly prefer x to y, while all others are indifferent – a kind of disagreement. In this case, some readers may find it plausible that the alternatives *are* morally noncomparable. In note 30 below, I briefly discuss the implications for the analysis of replacing intermediate Pareto with Pareto indifference.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

(9) Second, M satisfies *Anonymity*. Each advisor has equal weight in determining moral facts – assuming her preferences meet the idealization screen. Formally, if the preferences in profile $\mathbf{P^*}$ are a permutation of the preferences in profile $\mathbf{P}$, $M(\mathbf{P}) = M(\mathbf{P^*})$. Again, this flows from the view that the community's preferences are constitutive of moral facts. If these were, rather, merely evidence of moral facts grounded in some other way, some community members might have special expertise with respect to these facts, and Anonymity would be unwarranted.

(10) Third, M^S must satisfy a requirement of *Across-Situation Ranking Consistency* (ASRC). Note that the $\mathbf{S}$ superscript has been made explicit; it will be significant in what follows.

ASRC expresses several ideas. First, it is advisors' ordinal preferences that determine the moral goodness ranking of alternatives. To be sure, that idea also informs the structure of the moral aggregation function: the function for a given situation operates upon a profile of advisor ordinal rankings (quasiorderings) of that situation. However, there is nothing in the ordinality of the inputs to $M^{\mathbf{S}}$ which requires $M^{\mathbf{S}}$, the moral aggregation function for situation $\mathbf{S}$, to be similar in any further sense to $M^{\mathbf{T}}$, the moral aggregation function for a different situation $\mathbf{T}$. The formal setup, thus far, allows the choice of moral aggregation function for a given situation to hinge upon the specific features of the alternatives being compared. For example, if $\mathbf{S}$ and $\mathbf{T}$ each have three elements, with the elements of $\mathbf{S}$ actions of a particular sort, and the elements of $\mathbf{T}$ actions of a different sort, the aggregation functions for $\mathbf{S}$ and $\mathbf{T}$ might be quite dissimilar.

The IA theorist would (I think) resist such differentiation – which is in tension with the core claim of IA theory, that advisors' preferences are constitutive, and wholly so, of moral goodness. It is *just* the pattern of advisor preferences, not further facts about the alternatives, that fixes the goodness ranking; nor are these further facts 'smuggled in' by having them determine the specific form of the function from advisors' preferences to the goodness ranking. This is the second idea expressed by ASRC – that if advisors have 'the same' preferences in two situations, the moral goodness ranking should be 'the same'. We now go back to the first idea, ordinality: advisors' preferences are 'the same' in two situations if their ordinal preferences are 'the same'.

In short, ASRC says that if each advisor has 'the same' ordinal ranking of the items in $\mathbf{S}$ as in $\mathbf{T}$, then the moral goodness ranking should be 'the same'. Formally: Let $\#\mathbf{S} = \#\mathbf{T}$, and let $\pi(\cdot)$ be a bijection from situation $\mathbf{S}$ to situation $\mathbf{T}$. Let $\mathbf{P^S}$ be a profile of preferences in $\Phi^{\mathbf{S}}$ and $\mathbf{Q^T}$ a profile of preferences in $\Phi^{\mathbf{T}}$. Let M^S be the moral aggregation function for $\mathbf{S}$ and M^T for $\mathbf{T}$, such that $M^S(\mathbf{P^S})$ is a quasiordering of $\mathbf{S}$, and $M^T(\mathbf{Q^T})$ a quasiordering of $\mathbf{T}$. Let $R^{M}{}_{\mathbf{P}}$ denote the morally-at-least-as-good relation corresponding to $M^S(\mathbf{P^S})$, and similarly $R^{M}{}_{\mathbf{Q}}$ that relation corresponding to $M^T(\mathbf{Q^T})$.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

If for all x, y in **S** and for every advisor k, $x\ R^k_{\mathbf{P}}\ y$ iff $\pi(x)\ R^k_{\mathbf{Q}}\ \pi(y)$ and $y\ R^k_{\mathbf{P}}\ x$ iff $\pi(y)\ R^k_{\mathbf{Q}}\ \pi(x)$, *then*: for all x, y in **S**, $x\ R^M_{\mathbf{P}}\ y$ iff $\pi(x)\ R^M_{\mathbf{Q}}\ \pi(y)$ and $y\ R^M_{\mathbf{P}}\ x$ iff $\pi(y)\ R^M_{\mathbf{Q}}\ \pi(x)$.

(11) The final element of the quasi-Arrovian setup is *Independence*. It says, roughly, that the comparative moral goodness of two items depends just on advisors' pairwise rankings of those items. Independence *could* be expressed in an inter-situational version – namely that if advisors all have the very same rankings of the *x*-*y* pair as the *z*-*w* pair, with *x*-*y* and *z*-*w* belonging to the same situation or two different situations, the moral goodness ranking of the two must be the same. However, since across-situational consistency is already brought into play by ASRC, and since it will be illuminating to highlight the axiomatic work done by this inter-situational requirement as contrasted with Independence in the less ambitious, intra-situational sense, I formalize Independence as an intra-situational axiom. For any given pair of items in some situation, if each advisor's ranking of the two items under profile **P** is the same as her ranking under profile **Q**, the moral aggregation function must map the two profiles onto the same comparative moral goodness ranking of the two items.

Formally: Let x and y be any two alternatives from **S**, **P** and **Q** any two profiles of preferences in $\Phi^{\mathbf{S}}$. *If* for all k, $x\ R^k_{\mathbf{P}}\ y$ iff $x\ R^k_{\mathbf{Q}}\ y$ and $y\ R^k_{\mathbf{P}}\ x$ iff $y\ R^k_{\mathbf{Q}}\ x$, *then*: $x\ R^M_{\mathbf{P}}\ y$ iff $x\ R^M_{\mathbf{Q}}\ y$ and $y\ R^M_{\mathbf{P}}\ x$ iff $y\ R^M_{\mathbf{Q}}\ x$.

Independence is, of course, a key element of the original Arrow framework. It is also the most controversial. Why take it on board? Let me hold this question in abeyance until section 6. We proceed by, first, adopting the quasi-Arrovian setup, including Independence and seeing its implications for *M*; and, then, in section 6 considering the arguments for and against this axiom.[21]

4. THE QUASI-ARROVIAN SETUP PLUS ORDERING CONVERGENCE

'Ordering Convergence' is a (highly restrictive) stipulation regarding the admissibility of preference profiles. It says that a profile is a member of Φ only if each component preference has the very same ranking of the

[21] Last, how does this setup relate to existing scholarship on 'extensive social choice'? (Ooghe and Lauwers 2005, and sources cited therein). In that literature, a community of 'planners' – these might be interpreted as my 'advisors' – ranks outcomes, each using her own social welfare function. That is, each advisor has moral preferences that, in their substance, endorse Preferentism about Well-Being and the Principle of Personal Good. She arrives at her moral ranking, $\succcurlyeq^k$, by using an advisor-specific social welfare function W^k to solve the *interest-aggregation* problem that those with these substantive moral commitments face.

In other words, extensive social choice combines metaethical aggregation (as per an IA approach to metaethics) with interest aggregation (with respect to the substantive content of moral values). The framework here is more general, since it allows but does not require that advisors have moral preferences that are welfarist in their content.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

set **S** of items. Let a statement of identity between moral preferences, $\succcurlyeq^k = \succcurlyeq^l$, mean that: for all x, y in **S**, $x\ R^k\ y$ iff $x\ R^l\ y$. *Ordering Convergence*, thus, says: **P** is an element of Φ only if $\succcurlyeq^1_{\mathbf{P}} = \succcurlyeq^2_{\mathbf{P}} = \cdots = \succcurlyeq^N_{\mathbf{P}}$.

If Ordering Convergence is adopted, the problem of moral aggregation becomes trivial. By immediate upshot of Pareto indifference and weak Pareto, $M(\mathbf{P})$ must be some extension of the common advisor ranking.[22] If advisors have complete moral preferences, $M(\mathbf{P})$ must simply be the common ranking.[23]

Various IA theorists, most prominently Michael Smith,[24] have argued for the convergence of advisors' preferences. Do his arguments indeed make a persuasive case for Ordering Convergence?

Ordering Convergence should not be conflated with a particular aggregation rule M or family of such rules. In the next part of this article, I will discuss an aggregation rule, the Unanimity Rule, which is such that the weak convergence of advisors' preferences in ranking any two items is a necessary condition for those items to be morally comparable (one better than the other or the two equally good, rather than noncomparable). But convergence, here, is a feature of the rule – not of Φ. As we shall see, the Unanimity Rule operates perfectly well on preference profiles that include non-identical rankings – indeed, even if Φ contains every logically possible profile of advisor preferences.[25] The Unanimity Rule is a methodology for aggregating advisors' preferences, perhaps non-identical – a methodology that has the feature of propagating non-comparability *from* preference divergence – while Ordering Convergence is the view that a profile with divergent preferences is just metaphysically impossible from the outset.

Smith writes: 'the truth of a normative reason claim requires a convergence in the desires of fully rational agents' (Smith 1994: 173). This statement, standing alone, does not show that Smith endorses Ordering Convergence. Rather, it says that *if* there is to be genuine normative reason in favour of one option rather than a second, then fully rational agents must desire the first.

[22] That is, for all x, y: if $x\ I^1\ y$ (and thus $x\ I^2\ y$, $x\ I^3\ y$, …), then $x\ I^M\ y$; and if $x\ P^1\ y$ (and thus $x\ P^2\ y$, $x\ P^3\ y$, …), then $x\ P^M\ y$.

[23] $M(\mathbf{P}) = \succcurlyeq^1_{\mathbf{P}}$ $(= \succcurlyeq^2_{\mathbf{P}} = \cdots = \succcurlyeq^N_{\mathbf{P}})$. Absent completeness in the advisors' preferences, Ordering Convergence in the quasi-Arrovian setup does not imply this, but rather the slighter weaker result in footnote 22. Someone who embraces Ordering Convergence might rebel at this weaker result, since it allows that one item could be morally at least as good as a second even if all advisors count the two as noncomparable. This could be ruled out by adding an axiom that: if $x\ NC^k\ y$ for all k, then $x\ NC^M\ y$. Ordering Convergence in the quasi-Arrovian setup plus this axiom entails $M(\mathbf{P}) = \succcurlyeq^1_{\mathbf{P}}$ $(= \succcurlyeq^2_{\mathbf{P}} = \cdots = \succcurlyeq^N_{\mathbf{P}})$ even with incomplete advisor preferences.

[24] See also Firth (1952); Jackson (1998: 137); but see Brandt (1955).

[25] This is 'Universal Domain', defined in section 5.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

At other points, however, Smith tries to show that all idealized advisors must have the very same preferences. To argue thus *is*, clearly, to put forth the case for Ordering Convergence. Smith writes:

> [T]o be fully rational an agent must not be suffering from the effects of any physical or emotional disturbance, she must have no false beliefs, she must have all relevant true beliefs, and she must have a systematically justifiable set of desires, that is, a set of desires that is maximally coherent and unified. Furthermore, I argue that it is part of what we mean when we say that a set of desires is systematically justifiable that the desires that are elements in that set are desires that other people too would have if they had a systematically justifiable set of desires. Fully rational agents converge in the desires that they have, and converge by definition, because it is part of what we mean by the rational justification of our desires that people who have such desires have a justification for them that other people also could see to be a justification. (Smith 1997: 89)

More specifically, we might understand Smith as leveraging the *rationality* component of idealization into an argument for Ordering Convergence.

Is the argument persuasive? Smith asserts that 'rational' desires/preferences have the following feature (call it SJ): 'people who have such desires have a justification for them that other people also could see to be a justification'. Assume, now, that any two advisors k and l have the same full information, including knowledge about what each other's preferences are; moreover, each has SJ preferences; finally, the two have common knowledge of all this. Then if $\succcurlyeq^k \neq \succcurlyeq^l$, k will reason: 'I can't see my current preferences as justifiable to l, since l, with the very same information as me, and reasoning just as well, has different preferences'.

Note that this argument works only if 'having a justification for my desires' means 'having a non-agent-centred justification' – a justification that excludes indexicals. Consider a case in which k prefers Renaissance painting to modern painting, and l has the opposite preference. If k can provide a systematic justification for the preference by saying, 'Renaissance painting better fits with *my* deepest aesthetic values, the ones that have been inculcated in *me* since childhood', and l by saying 'modern painting fits better with *my* deepest aesthetic values, the ones inculcated in *me* since childhood', the two can have full and common knowledge of everything relevant, including the requirement of systematic justification, and still end up with different preferences.

So let us revise SJ to NACSJ (non-agent-centred, systematically justifiable). If 'rationality' requires NACSJ preferences, and if the idealizing conditions for moral facts require the advisors to have preferences that are NACSJ and sufficiently informed, we have a good argument for Ordering Convergence. On these suppositions, it is

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

metaphysically impossible for rational and well informed advisors to have divergent moral preferences.

A difficulty here is that nothing in the concept of 'rationality' clearly entails the NACSJ feature. A long tradition in thinking about rationality, going back to Hume, and embodied in modern economics and decision theory, thinks of 'rationality' in purely formal terms (Hooker and Streumer 2004; Weirich 2004). Such a view identifies conditions regarding the internal coherence of preferences, such as transitivity, reflexivity, consistency across choice situations, or the axioms of expected utility theory; and then says that satisfying such conditions is not only necessary, but *sufficient*, for someone's preferences to be rational. The view is 'purely formal' because these coherence conditions permit preferences that do not conform to any commonly shared standard of desirability. I can 'rationally' prefer something 'just because I want it'. Indeed, the purely formal view says that 'rational' preferences may seem substantively crazy – such as a preference to eat a saucer of mud, or Hume's example of preferring the destruction of the world to scratching my finger.

Nor does introducing the further idealizing element of 'impartiality' (as a condition on *moral* preferences) transform this formal conception of rationality into one that incorporates NACSJ. There's no apparent incoherence in Sarah explaining her impartial preferences to herself in Sarah-centred terms: 'Abstracting from my own position in society, I morally prefer much higher income-tax rates that would move us toward a much more equal distribution of well-being, even at substantial cost to overall well-being; and the reason I have this preference is because welfare egalitarianism fits *my* deepest values, the ones that my parents and I have always held.'

Smith might respond that the properties of preferences picked out by the concept of 'rationality' may not be obvious to those using that concept; and that, indeed, those properties include the NACSJ property, even though many who use that concept mistakenly think otherwise. (This would be like a mistaken belief that water is not H_2O.) To be sure, Smith himself is not omniscient; he cannot know for sure which properties our normative language refers to. But he can say that he *reasonably believes* 'rational' preferences to be NACSJ; and that if preferences do indeed have this feature, Ordering Convergence follows.

Several other arguments for Ordering Convergence should be mentioned. First, the *relativistic* IA theorist (who views moral statements as true or false relative to one or another community, and sees the preferences of idealized community members as the empirical basis for the truth-value of such statements) might restrict the set of eligible communities so that each satisfies Ordering Convergence. She might argue that a community's preferences can be the touchstone for a relativized moral statement only if the community genuinely shares a

moral code – and indeed does so in the strong sense required by Ordering Convergence. Such an argument, of course, is not available to the non-relativistic (absolutist) IA theorist.

Second, the non-relativistic IA theorist might allow that fully informed and rational *persons* can have divergent moral preferences, but include only *humans* in the community of advisors, and seek to make the empirical showing that idealized *humans* (given their genetic makeup) converge in what they morally prefer.

Finally, consider the disjunctive claim: *either* Ordering Convergence is true, *or* there are no moral facts. At one point, Smith makes comments suggestive of this claim.

> The truth of [Smith's account] depends on more than mere conceptual analysis [that is, the non-relativistic IA analysis of normative reasons], it depends, as well, on the substantive fact that *there is* a set of desires that we would all converge upon if we had a set of desires that was maximally informed, coherent and unified. Even if the conceptual analysis is impeccable, absent the power of rational argument – that is, absent the power of information, together with considerations of coherence and unity – to elicit common desires in us, [Smith's account] entails that there are no moral facts at all. (Smith 2004: 205)

The disjunctive claim, plus the truism of ordinary moral discourse that there *are* moral facts, would imply Ordering Convergence.

However, absent independent reason to accept Ordering Convergence, there is no reason to believe the disjunctive claim. Why believe it? (1) The disjunctive claim might, conceivably, be a feature of moral goodness that we discover by investigating the structure of moral aggregation. But there is no such discovery to be had. There exist moral aggregation functions – in particular, the Unanimity Rule – that yield facts about comparative moral goodness even without Ordering Convergence. (2) The disjunctive claim might express a truism of ordinary moral thought that any metaethical view should try to respect. But does it? Who (except for a few IA theorists) finds it pretheoretically obvious that convergence of moral preferences is the only alternative to moral nihilism? Indeed, many of us find the negation of the disjunctive claim to be pretheoretically obvious – observing moral disagreement all around us, and yet finding ourselves asserting moral truths in the teeth of such disagreement.

To sum up: the non-relativistic IA theorist who includes all persons (not just humans) in the community of advisors has *one* reasonable basis for endorsing Ordering Convergence – for believing that only profiles with identical moral preferences are metaphysically possible. That basis is the reasonable belief that the property of preferences picked out by the concept of *rationality* forces convergence.

However, since it is also reasonable to deny that 'rational' persons need to see their choices as systematically justifiable, or reasonable to accept agent-centred rationales as one such justification, it is also reasonable to reject Ordering Convergence. What doing so means for moral aggregation is the question we now consider.

5. THE QUASI-ARROVIAN SETUP WITH DIVERGENT PREFERENCES

For a given situation **S**, let $\mathbf{UD^S}$ be the set of all N-fold concatenations of quasiorderings of **S**. Universal Domain holds true in a given situation **S** iff $\Phi^{\mathbf{S}} = \mathbf{UD^S}$. $\Phi^{\mathbf{S}}$, the domain for the moral aggregation function, includes every combination of formally well-behaved rankings of the situation.

Whether the IA theorist believes that Universal Domain holds true in a given situation depends, of course, on her beliefs about the properties picked out by the terms 'rational', 'well-informed' and 'impartial'. It also depends on the nature of the situation. For example, let $\mathbf{S} = \{x, y, z\}$ be such that the three options are identical except that x involves the torturing of no babies, y the torturing of 10 and z the torturing of 100,000. Then (absent a radical disjunction between the idealized advisors and what we observe among non-ideal moral thinkers) it is hard to believe that any advisor would prefer z to y to x, let alone that all the elements of $\mathbf{UD^S}$ are metaphysically possible. By contrast, consider $\mathbf{T} = \{x^*, y^*, z^*\}$, with the three options otherwise identical except that x^* causes non-fatal physical harm to a group of persons, y^* a substantial loss of liberty and z^* pain and suffering. Then (since we do not observe non-ideal moral thinkers agreeing on a strict ranking of any two of these harms) it is easier to believe that all patterns of preference among x^*, y^* and z^* are possible for the ideal advisors.

I first discuss the implications of Universal Domain. If Universal Domain holds true in a given situation **S** (and $\#\mathbf{S} > 2$), the moral aggregation function for that situation must be the Unanimity Rule.

Conversely, if $\Phi^{\mathbf{S}}$ has a narrower domain, other moral aggregation functions – for example, majority vote – may be consistent with the intra-situational axioms of the quasi-Arrovian Setup (regardless of $\#\mathbf{S}$): weak Pareto, intermediate Pareto, Anonymity and Independence. However, by adding Across-Situation Ranking Consistency (ASRC), and an assumption that Universal Domain holds true in a sufficient number of situations, we force the moral aggregation function in *every* situation (with $\#\mathbf{S} > 2$) to be the Unanimity Rule.

The surprising result of this section is that the connection between moral goodness and convergent moral preferences emerges not from a demanding view of rationality (such as Michael Smith's), but from a very different direction. If rationality (and the other idealizing conditions) are *loose* enough that a very wide diversity of advisor preferences are

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

metaphysically possible in a sufficient range of situations, the combination of the intra- and inter-situational axioms of the quasi-Arrovian setup makes convergence a necessary condition for moral comparability.

5.1. Universal Domain Implies the Unanimity Rule

In the original Arrow setup, both the individual inputs and the collective output are required to be *complete* orderings. Arrow shows that there is no aggregation function satisfying this requirement plus Universal Domain, Independence, weak Pareto and Non-Dictatorship, where there are three or more alternatives being ranked and at least two persons in the population.

In an important contribution, Weymark (1984) shows that this impossibility dissipates if the collective output is allowed to be a quasiordering. Weymark, specifically, departs from the Arrow framework by (1) allowing the collective output to be incomplete, albeit retaining the requirement that individuals' inputs be complete; (2) substituting strong Pareto for weak Pareto; and (3) strengthening Non-Dictatorship to Anonymity. Weymark then proves that the only aggregation function possible in his modified framework is the following: alternative x at least as good as alternative y iff everyone weakly prefers x to y.

I will term this the *Unanimity Rule*.[26] Transposed into the context of moral aggregation, it says that one alternative is morally at least as good as a second iff all the advisors weakly prefer the first to the second. That is: $x\ R^M_{\mathbf{P}}\ y$ iff $x\ R^k_{\mathbf{P}}\ y$ for all k.

In the Appendix, I demonstrate that the only moral aggregation function consistent with the quasi-Arrovian setup (dropping the inter-situational axiom of ASRC, not needed for this result[27]) plus Universal Domain is the Unanimity Rule, if more than two alternatives are being compared with respect to moral goodness.

Theorem. *If* #S $>$ 2, *and Universal Domain holds true in* **S**, *a moral aggregation function* $M^{\mathbf{S}}$ *satisfies weak Pareto, intermediate Pareto, Independence and Anonymity iff* $M^{\mathbf{S}}$ *is the Unanimity Rule*.

Proof. See Appendix.

This characterization result differs from that of Weymark (1984) in two ways. First, both the collective output (the moral goodness ranking)

[26] This rule is better known to social choice theorists as the 'strong Pareto' rule, which is indeed the term that Weymark uses. So as to avoid any risk that readers from other fields might confuse this rule with the strong Pareto *axiom*, I have, for purposes of this article, picked a different name for the rule.

[27] Clearly, adding back ASRC does not create an impossibility. If the moral aggregation rule $M^{\mathbf{S}}$ for every situation **S** is the Unanimity Rule, $M^{\mathbf{S}}$ satisfies weak Pareto, intermediate Pareto, Independence, Anonymity *and* ASRC. whether or not Universal Domain is true of **S**.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

and the individuals' inputs (the advisors' preferences) are quasiorderings, possibly incomplete. Second, the strong Pareto axiom is not used in the characterization. Instead, the Pareto axioms that I use are the two 'built into' the quasi-Arrovian setup: weak Pareto and intermediate Pareto.[28]

Nonetheless, it should be emphasized that the Unanimity Rule *satisfies* the strong Pareto axiom. A different formulation of the Unanimity Rule, logically equivalent to that given above, is as follows. (1) Two alternatives are morally equally good iff all advisors are indifferent between them. ($x\ I^M_{\mathbf{P}}\ y$ iff $x\ I^k_{\mathbf{P}}\ y$ for all k.) (2) One alternative is morally better than a second iff at least one advisor strictly prefers the first and all others weakly prefer it. ($x\ P^M_{\mathbf{P}}\ y$ iff $x\ P^i_{\mathbf{P}}\ y$ for some i and $x\ R^k_{\mathbf{P}}\ y$ for all k.) (3) Two alternatives are morally noncomparable otherwise. ($x\ NC^M_{\mathbf{P}}\ y$ iff either (a) $x\ NC^k_{\mathbf{P}}\ y$ for some k or (b) there exist k, l, such that $x\ P^k_{\mathbf{P}}\ y$ and $y\ P^l_{\mathbf{P}}\ x$.)

It is immediately evident from prong (2) of this equivalent formulation of the Unanimity Rule that strong Pareto is satisfied. One lesson of the Theorem, therefore, is that the IA theorist who endorses the elements of the quasi-Arrovian setup is logically committed to endorsing strong Pareto, in a situation with Universal Domain – even though she may not find that strong Pareto has immediate intuitive appeal.

A second, striking feature of the Unanimity Rule is that it satisfies 'Output Convergence' (by contrast with Ordering Convergence). *Output Convergence*: Two alternatives x and y are morally comparable (either one better than the second, or the two equally good) only if there is one of the two alternatives that is weakly preferred to the other by *all* the advisors. Equivalently, if even one advisor counts x and y as noncomparable, or if two advisors have conflicting strict preferences – one strictly preferring x to y, the second strictly preferring y to x – then the alternatives are noncomparable. As can be seen from the initial statement of the Unanimity Rule, or from prong (3) of the equivalent formulation, this Rule does indeed satisfy Output Convergence.

Thus a corollary of the Theorem is the following.

Corollary. *If* $\#S > 2$, *and Universal Domain holds true in* $\mathbf{S}$, *a moral aggregation function* $M^{\mathbf{S}}$ *satisfies weak Pareto, intermediate Pareto, Independence and Anonymity only if* $M^{\mathbf{S}}$ *satisfies Output Convergence*.[29]

[28] A variation of the Theorem for the case where advisors' preferences are required to be complete also holds true. In this variation, Universal Domain is dropped, and instead it is supposed that $\Phi^{\mathbf{S}}$ is all N-fold concatenations of *complete* quasiorderings of $\mathbf{S}$. This variation is not demonstrated separately in the Appendix, but the proof there can be easily adapted to prove it. Lemmas 1 and 2 continue to apply. From the premise that $\mathbf{N}$ is the unique β-oligarchy and that advisors' rankings are complete, it follows immediately that: $x\ R^M_{\mathbf{P}}\ y$ implies that $x\ R^k_{\mathbf{P}}\ y$ for all $k \in \mathbf{N}$.
Adding intermediate Pareto, the characterization of the Unanimity Rule follows.

[29] An in-depth analysis of the $\#\mathbf{S} = 2$ case with the quasi-Arrovian setup and Universal Domain will not be undertaken here. It should be noted that if $\#\mathbf{S} = 2$, there are moral

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

The Theorem and Corollary will not be surprising to social choice theorists, who are familiar with the dramatic implications of the combination of Independence and Universal Domain. But metaethicists may find it puzzling. Think of the results this way. Independence imposes an inter-profile, intra-situational consistency constraint: a constraint regarding how the moral goodness ranking varies with the different possible profiles of advisor preferences for a given situation. The more such possible profiles there are, the stronger the 'bite' of this constraint. At the limit, if all profiles are possible – Universal Domain – the class of eligible moral aggregation functions is sharply circumscribed.

Ordering Convergence rules out non-identical preferences from the get-go. By contrast, the Theorem and Corollary show how convergence of moral preferences in a different sense – Output Convergence – is a *consequence* of the quasi-Arrovian structure of moral aggregation. Ironically, a tolerance for any pattern of disagreement or agreement among the advisors, at the outset, yields the Unanimity Rule, and therewith Output Convergence: moral noncomparability between two alternatives absent a *consensus* among the advisors in identifying one of the two as at least weakly preferable to the other.

Output Convergence does not go so far as to require that all advisors have exactly the same ranking of two items for moral comparability. In particular, it allows that one item can be morally better than a second if some advisors strictly prefer the first, while all others are indifferent between the two. (That is to say, Output Convergence and strong Pareto are logically consistent.) Thus we might say that Output Convergence makes the weak convergence of advisor preferences with respect to two items a precondition for their moral comparability.

Note finally that the Unanimity Rule is *neutral* between alternatives: the pattern of advisor preferences that yields moral betterness, equal goodness or incomparability between two alternatives is the *same* for any two. This is, intuitively, as it should be; and the Theorem shows neutrality to be an implication of the quasi-Arrovian setup plus Universal Domain, rather than needing to be stated as a further axiom.[30]

aggregation functions that satisfy weak Pareto, intermediate Pareto, Independence and Anonymity but do not satisfy Output Convergence. This is true, for example, of the majoritarian function M^* discussed below in section 5.2. With $\#\mathbf{S} = 2$, M^* will clearly yield a quasiordering of $\mathbf{S}$ and satisfy all the axioms just stated except Output Convergence. (Transitivity says: if x is at least as good as y and y is at least as good as z, then x is at least as good as z. This is applicable to the case of $\#\mathbf{S} = 2$ since x, y, and z need not be distinct.)

By contrast, what is observed in section 5.2 is that if $\#\mathbf{S}$ can exceed 2, M^* is sure to yield a quasiordering of $\mathbf{S}$ only if $\Phi^{\mathbf{S}}$ satisfies a domain restriction such as 'extremal restriction' or 'single-peakedness' whereby $\Phi^{\mathbf{S}}$ is a proper subset of $\mathbf{UD}^{\mathbf{S}}$.

[30] How do the results of this section change if Pareto indifference is substituted for intermediate Pareto? In an unpublished companion paper, I prove that if $\#S > 2$, and

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

5.2. The Implications of Relaxing Universal Domain

Let us continue to hold to one side the ASRC axiom. Where Universal Domain is *not* true in a given situation – if $\Phi^{\mathbf{S}}$ is only a proper subset of $\mathbf{UD}^{\mathbf{S}}$ – to what extent are moral aggregation functions *other* than the Unanimity Rule consistent with the rest of the quasi-Arrovian setup?

This is a large question which I can hardly begin to answer here. Rather, I will use the example of majority vote to illustrate the possibility of a non-Unanimity Rule that is consistent with the intra-situational axioms of the quasi-Arrovian setup (those except ASRC), given sufficiently narrow domains. Majority vote is worth singling out because much of the existing social-choice literature that relaxes Universal Domain but retains Independence has focused on this procedure. Moreover, majority vote adds an important perspective to the discussion thus far of convergence.

Define the majoritarian function M^* as follows: for every x and y in $\mathbf{S}$, x is morally at least as good as y iff the number of advisors who strictly prefer x to y is at least as large as the number who strictly prefer y to x.

M^* clearly satisfies weak Pareto, intermediate Pareto, Anonymity and Independence. Moreover – now borrowing from the literature – it can be shown that if $\Phi^{\mathbf{S}}$ is appropriately limited, every profile in $\Phi^{\mathbf{S}}$ will be mapped by M^* onto a quasiordering of $\mathbf{S}$ (indeed, a complete ordering). In particular, this will be true if $\Phi^{\mathbf{S}}$ is such that in every profile $\mathbf{P}$, every advisor has a complete ranking of the alternatives and the advisors' rankings satisfy a condition known as 'extremal restriction'. It will also be true if N is odd and $\Phi^{\mathbf{S}}$ is such that in every profile $\mathbf{P}$, every advisor has a complete ranking of the alternatives and the advisors' rankings are 'single peaked' (Sen 1970, 1986; Gaertner 2002, 2009). The intransitivities of combining M^* with a universal domain of preferences – so famously illustrated by social choice theorists beginning with Condorcet – disappear as $\Phi^{\mathbf{S}}$ shrinks.

The domain restrictions for M^* just mentioned do *not*, of course, entail Ordering Convergence. They admit profiles in which advisors have non-identical rankings of the items in $\mathbf{S}$.

Universal Domain holds true in $\mathbf{S}$, a moral aggregation function $M^{\mathbf{S}}$ satisfies weak Pareto, Pareto indifference, Independence and Anonymity iff $M^{\mathbf{S}}$ is one of a broader family of rules that includes but is not limited to the Unanimity Rule. The rules in this family all satisfy Output Convergence and Neutrality but not necessarily strong Pareto. Consider, in particular, the rule which says: alternative x is better than alternative y iff all advisors strictly prefer x to y; alternatives x and y are equally morally good iff all advisors are indifferent between the two; otherwise x and y are noncomparable. This rule yields a quasiordering of any situation, even with Universal Domain, and satisfies weak Pareto, Pareto indifference, Independence and Anonymity *but not* intermediate Pareto or strong Pareto; and it conforms to Output Convergence and Neutrality.

Moreover, by contrast with the Unanimity Rule, M^* does *not* satisfy Output Convergence. If more than half of the advisors strictly morally prefer x to y, x will be ranked better than y *regardless of what the other advisors prefer*.

However, consider what happens if we add the ASRC axiom, plus the supposition that Universal Domain is true in a sufficient number of situations. Specifically, for any situation **S**, there is at least one situation **T** such that #**S** = #**T** and such that Universal Domain holds true of **T**. *Then the moral aggregation function for every situation with cardinality greater than two must be the Unanimity Rule*. Why? There is a one-to-one correspondence between the elements of **S** and the elements of **T**. Moreover, because Universal Domain is true of **T**, it follows that for any profile $\mathbf{P^S}$, there is an isomorphic profile $\mathbf{P^T}$ – whereby each advisor's ranking of the elements of **S** is the same as her ranking of the corresponding elements of **T**. By the characterization theorem above, if #**T** >2, the mapping $M^{\mathbf{T}}$ from $\mathbf{P^T}$ onto the moral goodness ranking of **T** must be the Unanimity Rule. By ASRC, the mapping $M^{\mathbf{S}}$ from $\mathbf{P^S}$ onto the moral goodness ranking of **S** must be the very same rule.

6. BEYOND THE QUASI-ARROVIAN SETUP

The analysis to this point has worked within the quasi-Arrovian setup. The setup might be revised in various ways, either by changing the requirement that advisors' inputs and the moral output be quasiorderings, or by changing the axioms. I will here briefly discuss the latter type of revision – in particular dropping Independence or Across-Situation Ranking Consistency (ASRC). It is harder to see how the IA theorist could reject the Pareto axioms (which express the constitutive role of the community of idealized advisors in determining moral facts); or the Anonymity axiom (which expresses the equal role within that community of each advisor).

One possible revision to the quasi-Arrovian setup is to drop Independence but not ASRC. The setup, thus amended, retains the idea that advisors' ordinal preferences determine the moral goodness ranking of a set of alternatives. But it does not require that such determination occur on a pairwise basis. The comparative moral goodness of two alternatives need not be fixed by the advisors' ranking of those two, taken alone. Instead, it is the advisor-by-advisor ordinal preferences with respect to the entire set of alternatives that fix the moral goodness ranking of that entire set. The social choice literature describes aggregation functions that satisfy ASRC while violating Independence, the most famous being the Borda rule (Brams and Fishburn 2002; Pattanaik 2002).

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

A second amendment is to drop both Independence and ASRC.[31] The idea, here, would be that moral goodness depends in part on non-ordinal features of advisors' preferences – so that the pairwise ranking of two alternatives cannot be fixed by advisors' ordinal rankings of the two (Independence) and indeed the ranking of a whole set of items cannot be fixed by advisors' ordinal rankings of the whole set (ASRC).

Imagine, specifically, that moral aggregation proceeds using an approach analogous to the social-welfare-function (SWF) framework – which, as mentioned earlier, is a standard tool developed by social choice theorists for the context of interest aggregation. A given alternative is mapped onto a vector of 'utility' numbers. With N individuals in the population whose well-being is of concern, a given alternative x is mapped onto the vector (list) of numbers $(u_1(x), \ldots, u_N(x))$, with $u_i(x)$ the utility number representing the well-being of individual i given x. An SWF is a rule for ranking such vectors (Bossert and Weymark 2004; Adler 2012; Weymark 2016).

By analogy, the IA theorist *might* think that advisors' ordinal rankings of a given situation plus the relevant non-ordinal features of their preferences can be summarized in *moral utility numbers*. On this view, moral aggregation proceeds by translating a profile of advisor preferences over the alternatives into a vector of moral utility numbers, one for each alternative; and then by using some rule to rank these vectors. Because the moral utility numbers are capturing more than advisors' ordinal rankings of the situation, neither Independence nor ASRC can be expected to hold.

Should we indeed drop Independence, or Independence and ASRC, from our model of moral aggregation? Or is the quasi-Arrovian framework correct to include both axioms? How the IA theorist should answer these questions depends upon what it is reasonable to believe concerning the grounding of moral facts. Any IA theorist believes certain metaethical propositions that are basic to the IA approach: for example, that moral statements can be true or false, that there are moral facts, that moral facts are grounded in natural (physical, psychological or social) facts, and more specifically are grounded in the preferences of a group of idealized advisors. The IA theorist believes all this in virtue of considerations mentioned earlier, and is reasonable to do so. Now we can ask: Would it be epistemically more reasonable to believe that preferences constitute moral facts in the specific manner set forth by Independence and ASRC, or instead to deny this?

The metaethical discussion required for a full treatment of this question is well beyond what can be accomplished here. Let me simply make some initial, suggestive remarks.

[31] It is very hard to see why the IA theorist would seriously entertain a third possibility: retaining Independence but dropping ASRC.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

To begin, there is reason to doubt that there are non-ordinal features of the advisors' preferences, relevant to the determination of moral goodness and mirrored in moral utility numbers. *This* challenge to ASRC and Independence seems weak. Moral goodness is (as a matter of ordinary moral discourse) a purely ordinal property: we think of actions, outcomes and other items subject to moral assessment as being better or worse than each other, *not* as possessing degrees of moral goodness. So why wouldn't the advisors themselves, thinking morally, also have preferences that are purely ordinal?

There is an important disanalogy, here, between interest aggregation and moral aggregation. Ordinary discourse about *well-being* supports comparisons of well-being differences. It would be unexceptionable to say something like: 'Ron should spend that extra money on headache therapy, not the big screen TV, because getting rid of his chronic headaches will make a much bigger difference to his life than a larger TV.' Moreover, ordinary discourse about well-being does not share the scepticism about interpersonal comparisons that some economists espouse. No one (except an economist) would rebel at the statement that Stanley, who has a decent job, good mental and physical health, and a loving spouse and family, is better off than Xavier, who is unemployed, poor, depressed and alone.

The well-being utility functions that serve as the input to SWFs contain information about well-being differences and interpersonal comparisons, as well as tracking preferences: $u_i(x) \geq u_i(y)$ iff i has an appropriately laundered weak preference[32] for x over y. $u_i(x) - u_i(y) \geq u_i(z) - u_i(w)$ iff the difference in i's well-being between x and y is at least as large as the difference in i's well-being between z and w. $u_i(x) \geq u_j(y)$ iff i in x is at least as well off as j in y. To be sure, constructing well-being utility functions that reflect all this information is not a trivial matter (Adler 2016); but the supposition that there *are* genuine facts about well-being differences and interpersonal level comparisons, to be measured by well-being utility numbers, is not problematic.[33]

By contrast, it is hard to see what the underlying facts *are* which are supposed to be mirrored in moral utility numbers. What would be the counterpart of well-being differences, or of interpersonal comparisons of well-being levels – differences in moral satisfaction, or comparisons of levels of moral satisfaction? Talk of moral satisfaction is no part of

[32] Recall that welfare-relevant preferences need to be laundered, in the manner stipulated by whichever version of Preferentism about Well-Being has been adopted. See above, section 1.

[33] One standard SWF, the leximin SWF, is not sensitive to information about well-being differences; but this caveat does not undercut the supposition that there *is* difference information to be had, as needed, and mirrored in utility numbers.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

ordinary moral discourse. It *would* be weird to say 'Ron should give his cash to Oxfam, not Greenpeace, because helping to prevent hunger makes a bigger difference to his moral satisfaction than helping to save the planet'; or to say 'Phyllis is a big believer in animal rights, while Francine is a sceptic, and so if government were to regulate treatment of animals that occurs in factory farms Phyllis would be morally better off – more satisfied, morally – than Francine'. Whether a putative metaethical fact is *genuine* – specifically, here, whether advisors' moral preferences have a non-ordinal aspect, in turn determinative of moral goodness – depends, in some more or less complicated way, on both the truisms constitutive of our moral concepts (as evidenced by what is said or not said in moral discourse) and on physical, psychological and social features of the world. On the first score, at least, the case for such non-ordinal information is weak.

Turning now to Independence: this axiom, too, seems warranted by ordinary moral discourse. Here, it should be stressed that our topic has been moral betterness (comparative moral goodness), rather than moral rightness or absolute moral goodness. Whether one or another alternative is *right* depends upon the totality of alternatives available. And of course the *best* alternative in a given situation depends upon the comparative moral goodness of all the pairs. But it seems to be a truism of ordinary moral talk that the comparative goodness of two alternatives depends just upon the characteristics of the two. For example, if government is choosing between a policy x that increases a population's happiness but limits individuals' choices, and a policy y that increases individual liberties at the expense of happiness, it would be odd to think that the truth of the proposition 'x is better than y' might depend upon whether government was *also* considering a policy z targeted at population health.

Still, a plausible indirect argument can be mounted for rejecting Independence. Assume that we reject the supposition that non-ordinal features of advisor preferences help determine moral goodness. We thus accept ASRC. ASRC, plus the other elements of the quasi-Arrovian setup, plus the supposition that Universal Domain holds true in a sufficient number of situations, entails that the moral aggregation function for every situation with cardinality greater than two must be the Unanimity Rule. That was the upshot of section 5. And – it might be argued – the Unanimity Rule is too ready to propagate moral noncomparability from the heterogeneity of moral preferences that we observe around us and might imagine to characterize even idealized advisors. Via a process of reflective equilibrium – used here not in refining our substantive moral views, but in refining our *metaethical* understandings – we might end up rejecting Independence, despite its intuitive appeal, once we see what it yields when combined with other axioms that we find even *more* appealing.

7. CONCLUSION

This article has identified the problem of metaethical aggregation; offered a roadmap of different frameworks for analysing the problem; explored at length one such framework, the 'quasi-Arrovian setup'; and provided a characterization of the Unanimity Rule. As a consequence of this moral aggregation rule, convergence comes in at the 'back end', as Output Convergence (not Ordering Convergence): the weak convergence of advisors' moral preferences with respect to two alternatives is a necessary condition for the moral comparability of the two.

I have focused, throughout, on the problem of moral aggregation as arising from a particular type of metaethical account, the IA account. Actually, other accounts may also give rise to that problem. In particular, expressivism (traditional or hybrid) may do so, since someone may commit herself to a plan whereby she does whatever is approved by a community of advisors (Ridge 2006). I don't think the roadmap and analyses here would be substantially different if transposed to the expressivist's version of the aggregation problem.

The reader familiar with recent work in social choice on 'aggregation' may wonder why I have focused throughout on preference aggregation, saying nothing about *judgement* aggregation (List 2012). In the judgement-aggregation problem, each individual in a community makes logically coherent judgements about a group of propositions. We want a set of collective judgements that is logically coherent as well – and bears appropriate axiomatic connections to the individuals' judgements, for example satisfying a unanimity axiom (the analogue here of weak Pareto) which says that a proposition unanimously believed is in the collective set.

I believe that metaethical aggregation is, first and foremost, an issue of *preference* rather than *judgement* aggregation.[34] If I am wrong about this, we have yet another topic on which the tools of social choice can usefully illuminate questions in metaethics.

ACKNOWLEDGEMENTS

Many thanks for helpful comments and criticism to Vincent Conitzer, Andrew Forcehimes, Laurie Paul, Walter Sinnott-Armstrong and the participants in the Vanderbilt Conference on Rational Choice and

[34] The advisors' judgements about moral goodness are not constitutive of the truth of propositions regarding moral goodness. Rather, on the IA view, facts about the comparative moral goodness of items are constituted by some *M* function applied to the totality of advisors' idealized preferences. Propositions about moral goodness, in turn, are true if they fit these facts. The exact details of the idealization are left open. If idealized advisors have less than perfect information, no advisor may know for sure what the others' preferences are, or what *M* is.

Philosophy. Special thanks are owed to two anonymous referees for exceptionally thorough reports, and to the editors of *Economics and Philosophy* for detailed and very helpful guidance that has substantially improved this article. The usual disclaimer applies.

REFERENCES

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

Adler, M.D. 2016. Extended preferences. In *The Oxford Handbook of Well-Being and Public Policy*, ed. M. D. Adler and M. Fleurbaey. Oxford: Oxford University Press.

Arrow, K. J. 1951. *Social Choice and Individual Values*. New York, NY: Wiley.

Arrow, K. J. 1963. *Social Choice and Individual Values*, 2nd edn. New York, NY: Wiley.

Arrow, K. J., A. K. Sen and K. Suzumura, eds. 2002. *Handbook of Social Choice and Welfare*, vol. 1. Amsterdam: North-Holland.

Arrow, K. J., A. Sen and K. Suzumura, eds. 2010. *Handbook of Social Choice and Welfare*, vol. 2. Amsterdam: North-Holland.

Bossert, W. and J. A. Weymark. 2004. Utility in social choice. In *Handbook of Utility Theory*, ed. S. Barberà, P. J. Hammond and C. Seidl, vol. 2 (*Extensions*), 1099–1177. Boston: Kluwer Academic.

Brams, S. J. and P. C. Fishburn. 2002. Voting procedures. In *Handbook of Social Choice and Welfare*, ed. K. J. Arrow, A. K. Sen and K. Suzumura, vol. 1, 173–236. Amsterdam: North-Holland.

Brandt, R. B. 1955. The definition of an 'ideal observer' theory in ethics. *Philosophy and Phenomenological Research* 15: 407–413.

Brandt, R. B. 1979. *A Theory of the Good and the Right*. Oxford: Clarendon Press.

Broome, J. 1991. *Weighing Goods: Equality, Uncertainty, and Time*. Oxford: Blackwell.

Campbell, D. E. and J. S. Kelly. 2002. Impossibility theorems in the Arrovian framework. In *Handbook of Social Choice and Welfare*, ed. K. J. Arrow, A. K. Sen and K. Suzumura, vol. 1, 35–94. Amsterdam: North-Holland.

Divers, J. 2002. *Possible Worlds*. London: Routledge.

Egan, A. and B. Weatherson, eds. 2011. *Epistemic Modality*. Oxford: Oxford University Press.

Firth, R. 1952. Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research* 12: 317–345.

Fleurbaey, M. and P. Mongin. 2005. The news of the death of welfare economics is greatly exaggerated. *Social Choice and Welfare* 25: 381–418.

Gaertner, W. 2002. Domain restrictions. In *Handbook of Social Choice and Welfare*, ed. K. J. Arrow, A. K. Sen and K. Suzumura, vol. 1, 131–167. Amsterdam: North-Holland.

Gaertner, W. 2009. *A Primer in Social Choice Theory*, Rev. edn. Oxford: Oxford University Press.

Gendler, T. S. and J. Hawthorne, eds. 2002. *Conceivability and Possibility*. Oxford: Oxford University Press.

Harman, G. and J. J. Thomson. 1996. *Moral Relativism and Moral Objectivity*. Oxford: Blackwell.

Hooker, B. and B. Streumer. 2004. Procedural and substantive practical rationality. In *The Oxford Handbook of Rationality*, ed. A. R. Mele and P. Rawling, 57–74. Oxford: Oxford University Press.

Huemer, M. 2007. Epistemic possibility. *Synthese* 156: 119–142.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon Press.

Kment, B. 2012. Varieties of modality. In *The Stanford Encyclopedia of Philosophy* (Winter 2012 edition), ed. E. N. Zalta. <http://plato.stanford.edu/archives/win2012/entries/modality-varieties/>.
List, C. 2012. The theory of judgment aggregation: an introductory review. *Synthese* 187: 179–207.
Miller, A. 2013. *Contemporary Metaethics: An Introduction*, 2nd edn. Cambridge: Polity Press.
Ooghe, E. and L. Lauwers. 2005. Non-dictatorial extensive social choice. *Economic Theory* 25: 721–743.
Pattanaik, P. K. 2002. Positional rules of collective decision-making. In *Handbook of Social Choice and Welfare*, ed. K. J. Arrow, A. K. Sen and K. Suzumura, vol. 1, 361–394. Amsterdam: North-Holland.
Rabinowcz, W. 2015. Aggregation of value judgments differs from preference aggregation. In *Uncovering Facts and Values*, ed. A. Kuzniar and J. Odrowaz-Sypniewska (Poznan Studies in the Philosophy of the Sciences and the Humanities).
Railton, P. 1995. Moral realism: prospects and problems. In *Moral Knowledge? New Readings in Moral Epistemology*, ed. W. Sinnott-Armstrong and M. Timmons, 49–81. New York, NY: Oxford University Press.
Railton, P. 2003. *Facts, Values, and Norms: Essays toward a Morality of Consequence*. Cambridge: Cambridge University Press.
Ridge, M. 2006. Ecumenical expressivism: finessing Frege. *Ethics* 116: 302–336.
Ridge, M. 2014. *Impassioned Belief*. Oxford: Oxford University Press.
Roberts, K. W. S. 1980. Social choice theory: the single-profile and multi-profile approaches. *Review of Economic Studies* 47: 441–450.
Sen, A. K. 1970. *Collective Choice and Social Welfare*. San Francisco, CA: Holden-Day.
Sen, A. K. 1977. Social choice theory: a reexamination. *Econometrica* 45: 53–89.
Sen, A. 1986. Social choice theory. In *Handbook of Mathematical Economics*, ed. K. J. Arrow and M. D. Intriligator, vol. 3, 1073–1181. Amsterdam: North-Holland.
Schroeder, M. 2009. Hybrid expressivism: virtues and vices. *Ethics* 119: 257–309.
Schroeder, M. 2010. *Noncognitivism in Ethics*. New York, NY: Routledge.
Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.
Smith, M. 1997. In defense of *The Moral Problem*: a reply to Brink, Copp, and Sayre-McCord. *Ethics* 108: 84–119.
Smith, M. 2004. *Ethics and the A Priori: Selected Essays on Moral Psychology and Meta-Ethics*. Cambridge: Cambridge University Press.
Smith, M. 2005. Meta-ethics. In *The Oxford Handbook of Contemporary Philosophy*, ed. F. Jackson and M. Smith, 3–30. Oxford: Oxford University Press.
Voorhoeve, A. 2013. Vaulting intuition: Temkin's critique of transitivity. *Economics and Philosophy* 29: 409–423.
Weirich, P. 2004. Economic rationality. In *The Oxford Handbook of Rationality*, ed. A. R. Mele and P. Rawling, 380–398. Oxford: Oxford University Press.
Weymark, J. A. 1984. Arrow's theorem with social quasi-orderings. *Public Choice* 42: 235–246.
Weymark, J. A. 2016. Social welfare functions. In *The Oxford Handbook of Well-Being and Public Policy*, ed. M. D. Adler and M. Fleurbaey. Oxford: Oxford University Press.

Appendix

The *Unanimity Rule* for some situation $\mathbf{S}$ is the moral aggregation function $M^{\mathbf{S}}$ defined as follows: for all $\mathbf{P}^{\mathbf{S}} \in \Phi^{\mathbf{S}}$ and all $x, y \in \mathbf{S}$, $x\ R^{M\text{-}\mathbf{S}}\ y$ iff $x\ R^{k\text{-}\mathbf{S}}\ y$ for all $k \in \mathbf{N}$.

Theorem. *If* #S > 2, *and Universal Domain holds true in* $\mathbf{S}$, *a moral aggregation function* $M^{\mathbf{S}}$ *satisfies weak Pareto, intermediate Pareto, Independence and Anonymity iff* $M^{\mathbf{S}}$ *is the Unanimity Rule*.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

Proof of the Theorem

In what follows, the **S** superscript is omitted as **S** is held fixed. '$a \leftrightarrow b$' is used to mean 'a if and only if b'[35] and '$a \rightarrow b$' is used to mean 'if a, then b'.

It is straightforward to show that the Unanimity Rule satisfies weak Pareto, intermediate Pareto, Independence and Anonymity for any $\Phi \subseteq \mathbf{UD}$ and for any #**S**. (Recall that $\mathbf{UD}^{\mathbf{S}}$ – for short, **UD** – is the set of all N-fold concatenations of quasiorderings of **S**.) What is established here is that if $\Phi = \mathbf{UD}$ and #S > 2, the four axioms are sufficient to characterize the Unanimity Rule.

A set of advisors $\mathbf{G} \subseteq \mathbf{N}$ is *almost decisive* for the ordered pair of alternatives $(x, y) \in \mathbf{S} \times \mathbf{S}$ if for any $\mathbf{P} \in \Phi$, $[x\ P^k_{\mathbf{P}}\ y$ for all $k \in \mathbf{G}$ and $y\ P^i_{\mathbf{P}}\ x$ for all $i \notin \mathbf{G}] \rightarrow x\ P^M_{\mathbf{P}}\ y$. A set of advisors $\mathbf{G} \subseteq \mathbf{N}$ is *decisive* if for any $\mathbf{P} \in \Phi$ and for any $x, y \in \mathbf{S}$, $[x\ P^k_{\mathbf{P}}\ y$ for all $k \in \mathbf{G}] \rightarrow x\ P^M_{\mathbf{P}}\ y$.

Lemma 1. *If a group* **G** *is almost decisive for some ordered pair* (x, y)*, then it is decisive*.

Proof. Sen's 'field expansion' lemma (Sen 1970: 42–5; Weymark 1984) establishes this result for the case in which Φ is the set of all profiles of complete orderings of **S**. The same proof applies when, as here, Φ is the set of all profiles of quasiorderings of **S**. □

A set of advisors $\mathbf{G} \subseteq \mathbf{N}$ is a *β-oligarchy* (Weymark 1984) if (i) **G** is decisive and (ii) for any $k \in \mathbf{G}$ and $x, y \in \mathbf{S}$, $x\ P^k_{\mathbf{P}}\ y \rightarrow$ not $y\ R^M_{\mathbf{P}}\ x$.

Lemma 2. *There is a unique β-oligarchy and it is* **N**.

Proof. Weymark (1984) generally considers a 'collective choice rule' with the domain all profiles of complete orderings of a set of alternatives, and the codomain either all binary relations of the set of alternatives, or some subset of such binary relations. Here, the domain of M is $\Phi = \mathbf{UD}$, all profiles of quasiorderings of **S**, and its codomain is all quasiorderings of **S**. Both Weymark's Lemma 2 (stating that at most one oligarchy exists) and Theorem 1 (stating that there exists a β-oligarchy) hold true for a collective choice rule with the domain and codomain of M. The proof of Lemma 2 applies without modification. To modify Weymark's proof of Theorem 1: when he assumes $y\ P^M_{\mathbf{P}}\ x$ (using our notation), it is instead assumed that $y\ R^M_{\mathbf{P}}\ x$. Moreover, an additional group **D** is considered consisting of the individuals in group **G** who regard x and y as noncomparable. In this new profile, these individuals rank w below x and y.

By Anonymity, no group other than **N** can be a β-oligarchy. This establishes the second part of the result. □

Next, it is shown that if some profile **P** is such that each advisor's ranking of x and y is the same as her ranking of y and z, then if x is morally at least as good as y under **P**, y must be morally at least as good as z under **P**.

Two profiles **P** and **P*** 'coincide on $\{x, y\}$' if, for all $k \in \mathbf{N}$, $x\ R^k_{\mathbf{P}}\ y \leftrightarrow x\ R^k_{\mathbf{P}^*}\ y$ and $y\ R^k_{\mathbf{P}}\ x \leftrightarrow y\ R^k_{\mathbf{P}^*}\ x$.

Lemma 3. *For any* $x, y, z \in \mathbf{S}$*, let* $\mathbf{P} \in \Phi$ *be such that: for all* $k \in \mathbf{N}$*,* $x\ R^k_{\mathbf{P}}\ y \leftrightarrow y\ R^k_{\mathbf{P}}\ z$ *and* $y\ R^k_{\mathbf{P}}\ x \leftrightarrow z\ R^k_{\mathbf{P}}\ y$*. Then* $x\ R^M_{\mathbf{P}}\ y$ *implies that* $y\ R^M_{\mathbf{P}}\ z$*.*

Proof.

Let **P** be as stated in the lemma. Consider two profiles **P***, **P′** such that (a) **P** and **P*** coincide on $\{x, y\}$ and $y\ I^k_{\mathbf{P}^*}\ z$ for all $k \in \mathbf{N}$, and (b) **P** and **P′** coincide on $\{y,$

[35] Abbreviated above as 'a iff b'.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press

$z\}$ and $y\ I^k_{\mathbf{P}'}\ x$ for all $k \in \mathbf{N}$. By Universal Domain, $\mathbf{P^*}, \mathbf{P}' \in \Phi$. Note that (a) and (b) imply that (c) $\mathbf{P^*}$ and $\mathbf{P}'$ coincide on $\{x, z\}$.

Suppose that $x\ R^M_P\ y$. By Independence $x\ R^M_{\mathbf{P^*}}\ y$. By Pareto indifference and the transitivity of $R^M_{\mathbf{P^*}}$, it follows that $x\ R^M_{\mathbf{P^*}}\ z$. By Independence, $x\ R^M_{\mathbf{P}'}\ z$. By Pareto indifference and the transitivity of $R^M_{\mathbf{P}'}$ it follows that $y\ R^M_{\mathbf{P}'}\ z$. By Independence, $y\ R^M_P\ z$. □

Lemma 4. *For all* $x, y \in \mathbf{S}$ *and all* $\mathbf{P} \in \Phi$, $x\ R^M_{\mathbf{P}}\ y$ *implies that* $x\ R^k_{\mathbf{P}}\ y$ *for all* $k \in \mathbf{N}$.

Proof. The result is trivial if $x = y$. We consider the case $x \neq y$.

Because $\mathbf{N}$ is a β-oligarchy, $x\ R^M_{\mathbf{P}}\ y$ implies that there does not exist a $i \in \mathbf{N}$ for which $y\ P^i_{\mathbf{P}}\ x$. Thus for all $k \in \mathbf{N}$, either $x\ R^k_{\mathbf{P}}\ y$ or $x\ NC^k_{\mathbf{P}}\ y$. Let $\mathbf{G} = \{i \in \mathbf{N}: x\ R^i_{\mathbf{P}}\ y\}$. Suppose that $\mathbf{G} \neq \mathbf{N}$; we show that this yields a contradiction. Because $\#\mathbf{S} > 2$, there is some $z \in \mathbf{S}$ distinct from both x and y. Consider a profile $\mathbf{P}' \in \Phi$ that coincides with $\mathbf{P}$ on $\{x, y\}$ such that (a) $x\ I^i_{\mathbf{P}'}\ y \rightarrow y\ I^i_{\mathbf{P}'}\ z$ and $x\ P^i_{\mathbf{P}'}\ y \rightarrow y\ P^i_{\mathbf{P}'}\ z$, for all $i \in \mathbf{G}$; (b) $y\ NC^k_{\mathbf{P}'}\ z$ for all $k \notin \mathbf{G}$; and (c) $z\ P^k_{\mathbf{P}'}\ x$ for all $k \notin \mathbf{G}$. By Universal Domain, $\mathbf{P}' \in \Phi$.

Independence together with the premise $x\ R^M_{\mathbf{P}}\ y$ imply that $x\ R^M_{\mathbf{P}'}\ y$. By Lemma 3, $y\ R^M_{\mathbf{P}'}\ z$. By the transitivity of $R^M_{\mathbf{P}'}$, it follows that $x\ R^M_{\mathbf{P}'}\ z$. But $x\ R^M_{\mathbf{P}'}\ z$ and $z\ P^k_{\mathbf{P}'}\ x$ for all $k \notin \mathbf{G}$ contradicts Lemma 2. □

By Lemma 4 and intermediate Pareto, M is the Unanimity Rule. This completes the Proof.

BIOGRAPHICAL INFORMATION

Matthew D. Adler is Richard A. Horvitz Professor of Law and Professor of Economics, Philosophy, and Public Policy at Duke University. He is the author of *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis* (Oxford University Press, 2012), and, with Marc Fleurbaey, the editor of the forthcoming *Oxford Handbook of Well-Being and Public Policy*. His scholarship in recent years has focused on the 'social welfare function' framework for evaluating governmental policies – both the philosophical foundations of this framework in normative ethics and the theory of well-being, and its practical implementation in areas such as risk regulation and climate change.

https://doi.org/10.1017/S0266267115000486 Published online by Cambridge University Press