The Reasons Aggregation Theorem

RALPH WEDGWOOD

0. AGGREGATING REASONS

In many cases, when an agent faces a choice between alternative courses of action, there are *reasons* both *for* and *against* each of these alternatives.

For example, suppose that you face a choice between two Italian restaurants – Luna Caprese and Sole Siciliano. Suppose that considerations of food quality tell in favour of Luna Caprese and against Sole Siciliano, while considerations of price and economy tell in favour of Sole Siciliano and against Luna Caprese. This is a situation in which there is *one* reason – grounded in considerations of food quality – in favour of Luna Caprese and against Sole Siciliano, and also *another* reason – grounded in considerations of price and economy – in favour of Sole Siciliano and against Luna Caprese.

	Food quality	Economy
Luna Caprese	Reason *For*	Reason *Against*
Sole Siciliano	Reason *Against*	Reason *For*

Nonetheless, it might be clear that *overall*, or *all things considered* (ATC), there is *more reason* for one alternative than for another. For example, perhaps Luna Caprese is only *slightly* better than Sole Siciliano in terms of food quality, while Sole Siciliano is *much* better than Luna Caprese in terms of economy. Then overall, or ATC, there is more reason to choose Sole Siciliano than Luna Caprese.

Somehow, then, there must be a way of *combining* or *aggregating* the different reasons for and against each of the relevant alternatives, to yield an overall or ATC verdict on *how much reason* ATC there is for the agent to take each of these alternatives.

This overall or ATC verdict on these alternatives is a conclusion about *how much reason* ATC there is for the agent to take each of these alternatives. There is a way of understanding 'should' and 'ought' so that if, out of *all* available alternatives, one has most reason ATC to do *A*, then *A* is what one ATC *should* or *ought to* do. In this way, the correct way of aggregating the reasons for and against the available alternatives determines what one (in this sense) ATC should or ought to do.

How can these different reasons be aggregated? On many ways of thinking about reasons, this is a deep and inscrutable question – perhaps even a question where no systematic answers are to be found. In this paper, I shall sketch an alternative view of reasons, on which this question has a quite straightforward answer.

In this paper, I shall sketch a number of premises about reasons, and about how they can be aggregated. Some considerations will be given in support of each of these premises, but I will not be able to offer anything like a full defence of these premises here. The main goal of this paper is to show that, once all of these premises are in place, we can derive an analogue of a certain famous theorem from social choice theory. According to this theorem, the aggregation of reasons is necessarily *additive*. After explaining the theorem, I shall briefly consider some objections to this kind of additive aggregation. As I shall argue, these objections can all be addressed. In this way, I hope to marshal some considerations in favour of this essentially additive conception of the aggregation of reasons.

1. REASONS FOR ACTION ARE GROUNDED IN VALUES

Many philosophers hold that there is an intimate connection between *reasons for action* and *values* – where every "value" is a *way* of being *good*. As Joseph Raz (1999: 23) puts it: "reasons [for actions] are facts in virtue of which those actions are good in some respect and to some degree".

For example, the fact that I promised my mother to put up the Christmas decorations is a reason for me to put up the Christmas decorations because it is a fact in virtue of which the action of my putting up the Christmas decorations is *good*. Specifically, it is a fact in virtue of which the action of my putting up the Christmas decorations is an act of *fidelity* or *promise-keeping* – which is one respect in which an action can be good to a non-trivial degree.

If this is true, it will surely also be true that a reason *against* an action is a fact in virtue of which that action is *bad* in some respect and to some degree. The fact that your pushing a certain button would be an act of torturing another person is a reason *against* your pushing the button, because it is a fact in virtue of which your pushing the button is a *bad* thing to do. Specifically, it is a fact in virtue of which your pushing the button is an act of *victimizing* and *exploiting* the other person, or *infringing that person's rights* – which are respects in which actions can be bad to a non-trivial degree.

As I shall put it here, every reason for an action is "*grounded*" in a certain way or respect in which the action is good. Each of these "ways" or "respects" in which actions can be good is what I shall call a "value". I shall assume that there are many different values of this kind – that is, there are many different ways in which actions can be good or bad. For example, there is the value of promoting self-interest, and there are also numerous moral values as well – such as fidelity, fairness, beneficence, non-maleficence, and so on. These values can *conflict* with each other – as when one action *A* is better than an alternative action *B* in terms of self-interest, even though *B* is better than *A* in terms of one of the moral values, such as fidelity or fairness or beneficence.

This, then, is the first premise of my derivation of how the aggregation of reasons is possible – the premise that all reasons for action are grounded in values in this way.

This premise is compatible with many different views of how reasons are grounded in values. For example, it is compatible with the consequentialist view that all reasons for actions are grounded in the values that are *promoted* by those actions, and also with the nonconsequentialist view that some reasons for action are grounded in the values that are instantiated by the *actions themselves*.[1]

This premise is also compatible with a *desire-based* view of reasons (similar to that of Schroeder 2007). This is because one way in which something can be good is by being *good for* achieving an end *E* – where *E* is an end that the agent desires. Finlay (2014) argues that *every* normative use of 'good' expresses an end-relational way of being good of this kind. Even if this view is mistaken (as I believe), we may assume that the values – the ways of being good – include such end-relational ways of being good. (This assumption clearly guarantees the existence of a huge number of values – since it implies that for every end *E*, one of the values that exists is the value of being good for achieving *E*.)

Different theories of reasons will give different accounts of *which* of the many different values that exist are the *reason-providing values*, for the situation of an agent at a particular time. For example, the desire-based view of reasons will make the following two claims: first, whenever the agent desires an end *E* at a particular time, the value of being good for achieving *E* is a reason-providing value for that agent at that time; and secondly, there are no other reason-providing values.

I am in fact myself sympathetic to a much more *objectivist* view of the values that ground reasons for action. According to this more objectivist view, all the reason-providing values are more "objective" values that can be exemplified by actions – such as the value of self-interest, and the various moral values (such as fidelity, fairness, and beneficence) that I have canvased above. However, for the purposes of this paper, I can remain neutral about which values are the reason-providing values for an agent at a time. The first premise of the derivation that I shall discuss this paper is simply that all reasons for action are grounded in values of some kind.

2. REDUCING GOODNESS TO BETTERNESS

The second premise that I shall discuss here is a version of John Broome's principle that "goodness is reducible to betterness". Broome (1999: 163) formulates this principle as follows:

> If you knew everything about betterness – of every pair of alternatives, you knew whether one was better than the other, and which – then you would know everything there is to know about goodness. There is nothing more to goodness than betterness.

Specifically, as Broome (1999: 166) claims, "'*X* is good' reduces to '*X* is better than the standard'" – where this standard of comparison is simply determined by the context in which the term 'good' is used. In the case of *actions*, the standard of comparison seems typically to be the contextually salient alternative actions. In short, to say that an action is good is simply to say that it is better than the salient alternative(s).

To be precise, I should note that my version of this premise is slightly different from Broome's. First, as I have explained in the previous section, I am assuming here that there are

[1] For a sketch of this form of non-consequentialism, see Wedgwood (2009).

many different values – many different ways in which things can be good, or many different kinds of goodness. So, the premise that I need for my argument is that *each* of these kinds of goodness is reducible to the *corresponding* kind of betterness.

Moreover, I shall also allow that one may need to know a bit more about a given kind of betterness than the facts that Broome explicitly mentions, in order to know "everything that there is to know" about the corresponding kind of goodness. Specifically, one may also need to know, not only whether one of the two alternatives in question is better, and if so which, but also *how much* better (in terms of the relevant value) the better of the two alternatives is. For some purposes, one may also need to know the *explanation* of all these facts: that is, one may need to know *why A* is better than *B*, to the extent that it is, or why *C* is not better than *D*, and so on.

Apart from these qualifications, however, the second premise of the derivation that will be discussed here is fundamentally similar to Broome's principle that "goodness is reducible to betterness". Combining Raz's view of reasons with Broome's view of goodness yields the following two principles:

- A reason *for* an action is a fact in virtue of which that action is in a relevant way *better* than the relevant alternative(s).
- A reason *against* an action is a fact in virtue of which that action is in a relevant way *worse* than the relevant alternative(s).

These principles imply that, if an action is neither better nor worse in any relevant way than any relevant alternative, then there are no reasons either for or against that course of action. This seems to be a plausible implication of these principles.

What are "the relevant alternatives" here? We might be tempted to think that they are always *all* the actions that are available to the agent. But this tempting thought cannot be exactly right. Suppose that three alternatives are available to the agent – *A*, *B*, and *C* – where, with respect to one of the relevant values *V*, *A* is better than *B*, and *B* is better than *C*. Clearly, any fact in virtue of which these three acts are ranked in this way is a reason *for A* and a reason *against C*. However, if this tempting thought were correct, this fact would be a reason neither for nor against *B*: if the relevant alternatives to *B* include both *A* and *C*, then *B* is neither better than the relevant alternatives nor worse than the relevant alternatives – it is better than one relevant alternative and worse than another. Yet it seems that this fact must be a reason either for or against *B* – since it clearly bears on the question of whether or not to do *B*.

For this reason, we must suppose that the conversational context in which there is discussion about which actions are "good" and which are "bad", or about what is a reason "for" or "against" a given option, determines the "relevant alternatives" in a specific way. In particular, unless none of the available actions is either better or worse than any other in the relevant way, then the context must determine a sufficiently restricted set of "relevant alternatives" to each available action so that every action is in the relevant way either better or worse than all its relevant alternatives. Given our principles, this ensures that there will always be a reason either for or against every available action (at least unless none of them is either better or worse than any other in the relevant way).

As we shall see, however, it will not matter for our purposes exactly how the conversational context does this. It may be convenient in ordinary conversation to draw a line between good and bad, or between reasons for and reasons against, but if Broome's principle is correct, the most fundamental ethical theory need only focus on the *comparison* between the available actions in terms of the degree to which they exemplify each of the relevant values. It is this essentially comparative version of the value-based conception of reasons for action that is the second premise of the argument of this paper.[2]

3. AGGREGATING VALUES

On this values-based view, then, the comparisons of the available actions in terms of the specific *reason-providing values* $\{V_1, V_2\ldots\}$ must be capable of being somehow combined or aggregated to produce an overall ATC comparison in terms of *how much reason* there is for these actions. Indeed, fundamentally, it is not the reasons themselves that are combined or aggregated in this way; it is the specific *reason-providing values* that are combined or aggregated to determine how much reason ATC there is for all the available actions.

So far, I have just characterized values as "ways of being good". Is there a more illuminating general characterization of values that we can give? I propose that, in general, a value-concept is a concept that (a) *compares* or *ranks alternatives*, in a way that (b) can be expressed by terms like 'better' and 'worse', and (c) essentially plays a *reasoning-guiding role*.

By this criterion, the notion of *how much reason* ATC there is for the available alternatives is also a value-concept:

- **a.** This concept clearly compares or ranks alternative actions, in terms of *how much reason* ATC there is for those alternatives.
- **b.** The thought that there is ATC most reason for us to take course of action *A* can naturally be expressed by saying that *A* is ATC the "best thing for us to do".[3]
- **c.** This notion also seems to play a role in guiding our practical reasoning, by guiding us towards *choosing* any course of action that we judge ourselves to have most reason ATC to take.

Thus, the notion of how much reason ATC there is for the available alternatives is itself a value-concept. Our question concerns how the specific reason-providing values can be aggregated to yield the overall value of how much reason ATC there is for these alternatives.

This is an importantly different conception of aggregation from other conceptions that have been discussed in the philosophical literature. The utilitarian tradition – including the "ideal utilitarianism" of G. E. Moore (1903) – tends to think of aggregation as concerning the

[2] Another account that this view of reasons can be made compatible with is that of Snedegar (2018) – although unfortunately I do not have space to explain how to make these views compatible with each other here.

[3] Some readers may question whether this second point is compatible with *supererogatory* actions – that is, with cases in which there is an action that is in some way *better* than all alternatives, but not such that one "ought" ATC to do it. This short answer to this question is that, in these cases, the supererogatory action is *morally* best or *most admirable*, but not in this sense "the (unique) best thing to do ATC"; on the contrary, in these cases, both the supererogatory action and some morally permissible but less admirable action are such that no alternative option is an "ATC better thing to do". For a defence and exploration of this view of supererogation, see Wedgwood (2013), and also Section 6 below.

structure of a *single* kind of goodness. More specifically, this tradition interprets aggregation as concerned with the question of how the degree of goodness exemplified by a "whole" is determined or affected by the degrees of goodness that are exemplified by the "parts" of that whole. (For example, on many versions of classical utilitarianism, the degree to which the happiness of a group of people is good is interpreted as a function – such as the *sum* – of the degrees to which the happiness of each member of that group is good.)

The conception of aggregation that I am developing here is crucially different. On my conception, for there to be any question of the relevant kind of aggregation, there must be at least *three different* values that are all exemplified by the members of the *same* set of alternatives. Specifically, there must be at least two specific reason-providing values $\{V_1, V_2, \ldots\}$, each of which is exemplified to some degree by each of the relevant alternatives $\{A_1, A_2, \ldots\}$. The question is how the degrees to which these alternatives exemplify these specific reason-providing values $\{V_1, V_2, \ldots\}$ determine how these alternatives compare in terms of a third *overall* value – the value of how much reason ATC there is for these alternatives – which somehow combines and aggregates the specific reason-providing values.

This idea of aggregating these specific reason-providing values into some overall or ATC value is clearly analogous to *social choice theory*. The goal of social choice theory is to explore how the preferences or utility functions of the individual members of society can be aggregated to produce a "social" preference or utility function. As I shall argue in this paper, this is analogous to the aggregation of the specific reason-providing values into the "overall" value of how much reason ATC there is for the alternatives in question.[4]

4. ADVANTAGES OF THIS APPROACH

Before developing this analogy between the aggregation of reasons and social choice theory, however, I shall canvas some considerations that support the picture of reasons that I have sketched so far. Specifically, as I shall explain in this section, this picture allows us to solve some puzzles that rival views have found extremely challenging.

On the account that was sketched above, a reason in favour of an action is a fact that *explains* why the action is better than the relevant alternatives. Now, typically, an explanation of a fact presupposes some *background* of *normal circumstances*; this background is fixed by the context in which the explanatory claim is made.

This helps to solve a puzzle due to Shyam Nair (2016, 59): "both heat and rain, taken individually, function as reasons to not run; still, the combination of heat and rain together might function as a weaker reason to not run (say, because the heat is less onerous when there is rain)." How can the conjunction of two stronger reasons result in a weaker reason in this way?

On my picture of reasons, this phenomenon is easily explained. Suppose that the background of normal circumstances *B* includes: 'Normally: if it rains, it's not hot, and if it's hot, it

[4] For a pioneering discussion of the analogy between the aggregation of values and social choice theory, see Hurley (1989). For surveys of the most important results of social choice theory, see Sen (2017) and Mongin and Pivato (2016).

doesn't rain (though occasionally it rains even when it's hot).' We may think of this background as a probability distribution – specifically, a probability distribution that assigns low but non-zero conditional probability to the proposition that it's hot, given the assumption that it's raining, and also assigns low but non-zero conditional probability to the proposition that it's raining, given the assumption that it's hot.

The relevant reason-providing value seems to be indicated by Nair's allusion to how "onerous" it is to run. This value provides a *strong* reason not to run if, according to the relevant probability distribution, running can be expected to be *much* more onerous than not running. It provides a *weak* reason not to run if, according to the relevant probability distribution, running can be expected to be only *slightly* more onerous than running. In a typical context, the relevant probability function will be the result of conditionalizing this background *B* on the fact that is described as a "reason" in that context.

Thus, according to the probability distribution that results from conditionalizing the background *B* on 'It's hot', *running* is *significantly* worse than *not running* in terms of "expected onerousness". Similarly, according to the probability distribution that results from conditionalizing *B* on 'It's raining', *running* is again significantly worse than *not running* in terms of expected onerousness. However, according to the probability distribution that results from conditionalizing *B* on 'It's hot and it's raining', *running* has only *slightly* greater expected onerousness than *not running*. This explains why 'It's hot and it's raining' is a weaker reason against running than either 'It's hot' or 'It's raining'.

My picture also helps us to solve a puzzle about *counting* reasons (Fogal 2016, 88). Suppose that Ronnie loves to dance and there will be dancing at the party. Then, depending on what information is included in the background *B*, different statements will be true about what counts as Ronnie's reason to go to the party:

- If the background *B* includes 'There'll be dancing', then 'Ronnie loves to dance' is a reason (provided by the value of Ronnie's pleasure) for him to go to the party.
- If the background *B* includes 'Ronnie loves to dance', then 'There'll be dancing' is a reason (provided by the value of Ronnie's pleasure) for him to go.to the party.
- If the background *B* includes neither fact, then 'Ronnie loves to dance and there'll be dancing' is a reason (provided by the value of Ronnie's pleasure) to go to the party.

However, all these three reasons are provided by the *same* value, relative to *different* backgrounds. Consequently, it would be a mistake to try to aggregate these reasons. What we fundamentally need to aggregate are the different reason-providing values – not the facts that we call the "reasons", which explain how the relevant alternatives compare in terms of these reason-providing values. Statements that presuppose *different* backgrounds create a mere illusion of different reasons. In this case, these statements are really just different ways of talking about the *same* practically significant fact – the fact about how the options of *going to the party* and *not going to the party* compare with respect to the reason-providing value of Ronnie's pleasure.

One piece of evidence in favour of my picture of reasons, as it seems to me, is the neat way in which it can solve these puzzles. Obviously, this is not a conclusive proof of this picture; but in my judgment it does give us a reason to take this picture of reasons seriously. In what follows, I shall explore the implications of this picture.

5. THE ANALOGY WITH SOCIAL CHOICE THEORY

To develop the analogy with social choice theory, a number of further premises need to be in place. These further premises fall into two groups.

First, each of the relevant values – the specific reason-providing values, and the value of how much reason ATC there is for the relevant alternatives – must be capable of being measured or represented by means of a *utility function* – and all these utility functions must be defined in terms of the *same* probability distribution.

Secondly, a pair of *Pareto* principles need to hold about the relationship between (a) the specific reason-providing values and (b) the overall value of how much reason ATC there is for the relevant alternatives. Intuitively, such Pareto principles seem highly plausible. For example, consider a case in which the only two reason-providing values are (i) beneficence and (ii) self-interest. Suppose that *A* is at least as good as *B* in terms of both beneficence *and* self-interest, and better than *B* in terms of self-interest. Then it seems plausible that there is more reason ATC for *A* than for *B*.

Once these two groups of premises are in place, an important theorem follows. Before getting to that theorem, however, I shall articulate these premises more precisely, and I shall canvas some considerations that seem to me to make these premises plausible and worth taking seriously.

i. EXPECTED VALUE MEASUREMENT

First, then, we need the assumption that each of the relevant values is capable of a kind of *cardinal measurement*. Specifically, to fix ideas, I shall assume that every relevant value can be measured on an *interval* scale – that is, a scale (like the Celsius and Fahrenheit scales for temperature) on which the zero point and the unit are arbitrary, but the ratios between *differences in value* are not arbitrary. In other words, it can be literally true that the difference in value between *A* and *B* is *twice* the difference between *B* and *C*, and so on.

Next, we need the assumption that the measurement of each of these values takes an "expectational" form. According to this assumption, each of the relevant alternatives has a precise degree of this value at every *possible world* that is compatible with that alternative. However, this value is not only exemplified by these alternatives at these possible worlds; it is also exemplified by certain related *uncertain prospects*, and the value of these uncertain prospects depends on the relevant *probabilities*.

What exactly are these "uncertain prospects"? Suppose, for example, that the alternatives are *going Left* and *going Right*. Then one uncertain prospect is the act of *going Left* in all the relevant possible worlds (where one does not know for certain which of these worlds one is in); another uncertain prospect is the "mixed strategy" of *going Left* if a certain coin toss lands heads and *going Right* if it does not land heads. In general, each of these uncertain prospects corresponds to a set of possible worlds – where, for each world in the set, one performs one of the alternative actions that has a precise degree of value at that world.

If the measurement of the value takes an "expectational" form, then the value of the "mixed strategy" that I have just described is the *weighted sum* of the precise value of going Left at each world where the coin lands heads and the precise value of going Right at each world where the coin does not land heads – weighting each of these worlds by the conditional probability of the world, given the assumption that one takes this "mixed strategy". In general, the value of each uncertain prospect is the weighted sum of the precise values of the relevant alternatives at the relevant worlds, weighting each world by the relevant probability of the world, conditional on the prospect in question.

Why should it be that the measurement of these values must take this expectational form? Many different answers to this question could be suggested, but here is the answer that seems most promising to me. Perhaps it is just an essential part of each value's job description that it needs to be able to assess, not just alternatives relative to worlds, but also uncertain prospects of the kind that I have been discussing. Moreover, in assessing these uncertain prospects, the value needs to be able to get together with *any* probability function that is defined over the relevant worlds, to yield an assessment of these uncertain prospects from the standpoint of that probability function.

In our everyday use of evaluative terms like 'good' and 'bad', 'better' and 'worse', we are often willing to apply these terms to uncertain prospects, from the standpoint of a certain epistemic perspective. For example, even before knowing the jury's verdict, a defence attorney might say to her client, 'It's good that the jury has reached their verdict so quickly'. What makes this statement true is that the uncertain prospect of *the jury's reaching their verdict quickly* is rationally assessed as more valuable than the salient alternative, from the standpoint of the attorney's epistemic perspective. I suggest that we can model these sorts of epistemic perspectives by means of probability functions.

In this way, each value provides comparative assessments of the relevant prospects, from the standpoint of each probability function – assessments that could be expressed by saying such things as 'Prospect *A* is better than prospect *B*', and the like. Arguably, these comparative assessments have to meet the conditions that are imposed by the so-called "axioms" of expected utility theory.[5] (For example, these assessments must be *transitive*, and so on.) If these assessments must indeed meet all these conditions, then it follows that the measurement of the value must take this expectational form. In other words, the measurement of the value takes the form of a utility function.

On this approach, then, the value of uncertain prospects is relativized to the standpoint of particular epistemic perspectives – where we are assuming here that each of these epistemic perspectives can be modelled by a probability function. For example, consider the epistemic perspective of the agent who is deciding what to do. The agent can make assessments from this perspective of how the relevant alternatives compare *both* in terms of the specific reason-

[5] More precisely, I have in mind the conditions that are imposed by the axioms of the representation theorem that Bolker (1967) proved for the decision theory of Jeffrey (1965 / 1983); for a lucid account of these axioms, see Joyce (1999: Chap. 4). However, my approach differs from Bolker's in that the probability function *P* is given independently of the comparative value assessments – just as in the older proofs that were due to von Neumann and Morgenstern (1944 / 1953). If the probability function *P* is given independently, Bolker's axioms guarantee the existence of a unique utility function (given an arbitrary choice of a unit and zero point). The general idea of using the axioms of decision theory to provide a measure of goodness is due to Broome (1992, Chap. 6). However, my approach differs from Broome's in two respects: (a) I do not assume that any special probability distribution is privileged as the one that is uniquely involved in the nature of the value – on my approach, comparative assessments in terms of the value can be made from the standpoint of *any* probability distribution; (b) I prefer to rely on Bolker's axioms rather than on those of Savage (1954).

providing values *and* in terms of the overall value of how much reason ATC there is these alternatives; this perspective can be modelled by a probability function *P*(•). To fix ideas, let us assume that the values that we are concerned with here – at least as they apply to uncertain prospects of the kind that we have discussed – are all relativized to this epistemic perspective.

This gives us some reason to accept the following premise that we need for our argument: all the values that we are concerned with here can be measured by means of functions that have the form of *utility functions* – where all these utility functions are defined in terms of the same probability function *P*(•).

ii. THE PARETO PRINCIPLES

The last group of premises that I need for my argument is that the relation between the specific reason-providing values and the overall ATC value of these alternatives obeys two *Pareto principles*.

The first Pareto principle is a *Pareto Indifference* principle: if all the specific reason-providing values favour two alternatives *A* and *B* exactly *equally*, then there is exactly as much reason ATC for *A* as for *B*; *A* is *exactly as good* as *B* in terms of how much reason ATC there is for them. The second Pareto principle is a *Pareto Preference* principle: if every reason-providing value favours *A* at least as strongly as *B*, and at least one such value favours *A more strongly* than *B*, then there is *more* reason ATC for *A* than for *B*; *A* is *better than B* in terms of how much reason ATC there is for them.

What exactly do I mean by talking about how "strongly" a reason-providing value "favours" an alternative? As I explained in the previous subsection, all these values are relativized to the standpoint of a certain probability function *P*. So, we might think that for a reason-providing value to favour *A* and *B* equally (as assessed from the standpoint of *P*) is just for *A* and *B* to have exactly *equal* expected value, according to this probability function *P*.

If this is what we mean by how "strongly" this value favours an alternative, then the Pareto principle is what is known as an "*ex ante* Pareto principle", which encodes a strong kind of neutrality about *risk*.

For example, consider self-interest – the value of promoting one's own well-being. Suppose that one alternative *A* involves getting 100 units of well-being for certain, while a second alternative *B* involves taking a gamble between 200 and 0 units of well-being at equal odds. In this case, the expected value of *A* and *B*, in terms of how well they promote one's well-being, is exactly equal. If well-being is the only reason-providing value in this case, and the measure of how "strongly" a value favours an alternative is just the expected value of the alternative, then the Pareto Indifference principle implies that there is exactly as much reason ATC for *A* as for *B*.

It seems to me that we should avoid being committed to this strong kind of neutrality. Agents can reasonably be risk-averse, and these agents would rightly believe that they have more reason ATC for the more cautious option *A* than for the risker alternative *B*.[6] So, if we are to

[6] I believe that a *rational* agent's choices and preferences will align with her rational expectations of how much reason ATC she has for each option. In this sense, a rational agent may be reasonably risk-averse about all values *other than* the value of how much reason there is ATC for each option. If this is right, then, for every

defend these Pareto principles, we need to adopt a different measure of how "strongly" a specific reason-providing value favours an alternative. In general, the strength with which a value favours an alternative need not be identical to the expected value of the alternative.

However, the strength with which a value favours an alternative relative to a world *w* is surely an *increasing function* of the alternative's value at *w*. For example, if *A* is *better for the agent* at *w* than at *v*, then the value of goodness-for-the-agent favours *A* more strongly relative to *w* than to *v*; and the greater this difference in goodness-for-the-agent, the greater the difference between how strongly this value favours *A* relative to *w* and relative to *v*.

Still, the strength with which a value *V* favours an alternative *A* relative to a world *w* may reflect, not just the value of *A* in terms of *V* at *w*, but also some reasonable attitude towards risk. If the agent is reasonably risk-averse, then the strength with which *V* favours *A* may be a *concave* function of *A*'s value in terms of *V*. As we might put it, it may be that *goodness-in-terms-of-V* makes a *declining marginal contribution* to the strength of the *V*-grounded reason in favour of the alternatives that are better in terms of *V*.

We can illustrate this point by returning to our example of the choice between getting 100 units of well-being for certain and taking a gamble between 0 and 200 units at equal odds. The difference in well-being between 0 and 100 units is the same as the difference between 100 and 200 units. But if well-being makes a declining marginal contribution to the strength of reasons, then the difference between how "strongly" the value of well-being "favours" getting 100 units and how strongly it favours getting 0 units is *greater* than the difference between how strongly it favours getting 200 units and how strongly it favours getting 100 units. On this view, we are not compelled to say that the value of well-being favours the two courses of action equally strongly – even if the expected level of well-being is the same.

With this revised understanding of how strongly a reason-providing value favours an alternative, the following Pareto principles seem plausible (in general, how strongly a value *V* favours an alternative *A*, when estimated from the standpoint of a probability function *P*, is just the expected strength with which *V* favours *A* according to *P*):

a. If (when estimated from the standpoint of *P*) *every* reason-providing value *V* favours *A exactly as strongly* as *B*, then (when estimated from the standpoint of *P*) *A* is *exactly as good* as *B* in terms of how much reason ATC there is for them.

b. If (when estimated from the standpoint of *P*) *every* reason-providing value *V* favours *A at least as strongly* as *B*, and *some* such value *V′* favours *A more strongly* than *B*, then (when estimated from the standpoint of *P*) *A* is *better than B* in terms of how much reason ATC there is for them.

Evidently, much more could be said about these Pareto principles. In what follows, however, we shall simply assume that they are correct, in order to explore their implications.

value except for this one, we cannot derive a correct measure of the value by constructing a utility function from the agent's rational preferences between gambles involving the value: that derivation would wrongly say that the expected well-being of getting 100 units for certain is *greater* than the expected well-being of taking a 50/50 gamble between 0 and 200 units. This is why I sketched a different method for measuring the value in the previous subsection (note 5 above).

iii. THE THEOREM

With all these premises in place, we can now derive an analogue of John Harsanyi's (1955) "Social Aggregation Theorem".[7]

> The measure of how much reason ATC there is for each relevant alternative *A* is a *weighted sum* of the measures of how strongly *A* is favoured by the relevant reason-providing values V_1, V_2, ….

Formally, we can without loss of generality understand this theorem by focusing only on *reasons against* alternatives.

If a specific reason-providing value *V* does not favour *any* alternative *B* more strongly than *A*, then *A* is *optimal* in terms of *V*. In that case, there is *no* *V*-provided reason against *A*; the measure of how much *V*-provided reason there is *against A* is then 0. For every other alternative *B*, the measure *m*(*B*) of how much *V*-provided reason there is against *B* is given by much *more strongly V* favours the optimal alternatives (like *A*) than *B*.

For every reason-providing value V_i, and every such measure m_i of how much V_i-provided reason there is against these alternatives, there is some reasonable way of assigning a weight α_i – a positive real number – to m_i. The measure of how much reason ATC there is against an alternative *A* is given by this weighted sum:

$$\sum_i \alpha_i \, m_i (A)$$

The *greater* this weighted sum is, the *more reason* there is ATC *against* the alternative *A*. According to the interpretation of 'ought' that was canvased in Section 0 above, an alternative is one that you *ought not* ATC to take if and only if there is ATC *more reason against it* than against some other available alternative. In this way, it is this essentially additive weighing of the reason-providing values that determines what you ought ATC to do.

For instance, let us return to the example that we considered at the outset, of the choice between two Italian restaurants, Luna Caprese and Sole Siciliano. Suppose that there are two reason-providing values – economy and food quality.

- The value of food quality does not provide any reason against Luna Caprese at all, and it provides a *weak* reason against Sole Siciliano.
- The value of economy does not provide any reason against Sole Siciliano, but it provides a *strong* reason against Luna Caprese.

Adding the reasons against each alternative together, the combined reasons against Luna Caprese are clearly weightier than the combined reasons against Sole Siciliano. There is therefore more reason ATC against going to Luna Caprese than against going to Sole Siciliano. Going to Sole Siciliano is what you ATC ought to do.

On this conception of the aggregation of reasons, the metaphor of "weighing reasons" is not merely rough and suggestive. On the contrary, there is an *exact* parallel between the weighing of reasons and the weighing of physical quantities. On this conception, each of the reason-providing values grounds a certain quantity for each alternative – a certain amount of *reason*

[7] For the details of the derivation, see Broome (1992, Chap. 10).

against that alternative; this quantity is zero if the alternative is optimal in terms of the value, and greater than zero if the alternative is suboptimal. These quantities can then all be added up, to produce a measurement of how much reason ATC there is against the alternative. Finally, we compare how much reason ATC there is against the alternative with how much reason ATC there is against every other alternative. This is exactly like weighing collections of physical objects, with the goal of selecting the lightest or least heavy of these collections.

6. WHERE DO THE "WEIGHTS" COME FROM?

According to our premises, each of the reason-providing values can be measured on a unique interval scale, which is formally like a utility function. Together with some reasonable attitude towards risk, this provides a measure of how "strongly" this value "favours" the relevant alternatives. However, according to these premises, these values are all measured on *different* scales. There is no common scale on which all these values can be measured.

Thus, these weights do not come from the specific reason-providing values. Instead, they must come from a different source. It seems to me that the most plausible account is that these weights come from the standpoint of *the agent's practical reason*. In effect, they represent the way in which the agent can reasonably convert these independent measures of the specific reason-providing values into a way of measuring these values' contributions to how much reason ATC there is for the relevant alternatives. In other words, these weights express a judgment of the *relative importance* of these specific values – as measured by the measures in question – for determining what the agent ought ATC to do.

Various different views of these weights are possible here. One possible view is that there is a unique correct weighting of these measures of the specific reason-providing values, fixed by objective features of the agent's situation – such as the facts about which options are available and the values that are at stake. On this view, the agent has no discretion in weighting these measures of the values: a unique correct weighting is dictated by these objective features of the agent's situation.

On an alternative view, these objective features of the situation only fix a range of equally legitimate or permissible weightings, and there are several alternative ways in which the agent may legitimately exercise her practical reason in weighting these measures of the values in one way rather than another. On this view, then, some judgments of these values' relative importance are simply incorrect; but once these incorrect judgments are excluded, there may be a range of judgments of these values' relative importance that are all equally legitimate. It may be one mark of *saintly* agents – that is, agents who routinely perform actions that others reasonably regard as supererogatory – that they assign unusually great relative importance to moral values, compared to the value of self-interest. By contrast, less saintly agents may assign comparatively greater relative importance to self-interest. Both the more saintly and the less saintly agents may be equally reasonable – neither kind of agent is necessarily inferior to the other with respect to the virtue of practical wisdom.

In some cases, agents may be *unable* to assign any definite weights to these measures of the reason-providing values. For these agents, the conflicting values could be incommensurable. It is a good question what we should say about such cases of incommensurability, but we need not pursue this question here. At all events, such incommensurability is far from ubiquitous. In many cases, we have no difficulty in judging the relative importance of competing reason-providing values.

Undoubtedly, much further investigation is required into the questions of which ways of weighting these measures of the values are reasonable and which are not, and why these weightings are reasonable in this way. But there is no reason to assume that there is no illuminating account that could in principle be given.

7. OBJECTIONS TO THIS ADDITIVE CONCEPTION

The idea that the aggregation of reasons is essentially additive is often criticized – especially by those who are sympathetic to a "holistic" or "particularistic" view of reasons, such as that of Jonathan Dancy (2004). In fact, however, the conflict between the "Reasons Aggregation Theorem" and this holistic or particularistic view is much less stark than it may initially appear to be.

Within this framework of this theorem, the particularistic or holistic phenomena that Dancy appeals to may arise at three levels:

a. The way in which the specific values $\{V_1, V_2\ldots\}$ of the alternatives *A*, *B*, … that are available to the agent in the relevant situation are determined by the *naturalistic* facts about this situation may be sensitive to the exact constellation of all these naturalistic facts.

b. The precise list of values $\{V_1, \ldots, V_m\}$ that count as "reason-providing" in the agent's situation may be determined holistically by the complete list of *all* the values $\{V_1, \ldots, V_n\}$ that are non-trivially exemplified by the available acts.

c. Finally, the precise *weighting* of the relevant measures of these values may also be determined holistically by the complete list of available acts and of the values that are non-trivially exemplified by those acts.

At the first level (a), for each of these specific values V_i, the way in which these naturalistic facts determine the degrees to which the available acts are good in terms of V_i may defy all attempts to summarize in terms of a simple principle. This will account for most of the particularistic phenomena that Dancy appeals to.

For example, consider the huge range of naturalistic facts that are relevant to determine whether an act *A* is an act of infidelity. It is not sufficient that if the agent does *A*, the agent will not fulfil a promise: it is also relevant whether the promise was given under duress, whether it was even feasible for the agent to keep the promise, and whether the promise was morally permissible in the first place. Even more naturalistic facts are relevant to determining *how much* worse, in terms of the value of fidelity, this act *A* is than the alternative acts that do not count as acts of infidelity. (Was it a solemn vow or a casual promise? How seriously is the promisee relying on the promise's being kept? And so on.)

At the second level (b), the complete list of values $\{V_1, \ldots, V_n\}$ that are non-trivially exemplified in the agent's situation may result in some of the values on this list – $\{V_n, \ldots\}$ – *failing* to count as "reason-providing" – even though in other situations, in which a different complete list of values is non-trivially exemplified, these values *do* count as reason-providing. This will account for the way in which reasons can be "silenced" or "disabled" or "excluded" in the way that was famously explored by Joseph Raz (1990, 41).

For example, it may be that the value of *family loyalty* – which in many situations is a genuinely reason-providing value – may be "silenced" or "disabled" from being reason-providing in certain situations. For example, perhaps the value of family loyalty is not just outweighed, but totally silenced, when one is acting on behalf of an institution – say, on a committee that is tasked with awarding government contracts, or making academic appointments. In this case, the fact that the alternatives differ in terms of the value of institutional fairness may in effect disable the value of family loyalty from counting as reason-providing at all.

The way in which it is determined which values count as "reason-providing" in the agent's situation may be important in order to avoid a kind of *double-counting*. The fact that an action would be admirable may seem to count as a reason in favour of the action. But the degree to which an action is admirable depends on the degree to which it manifests many other talents, skills, and virtues. If these virtues were included among the reason-providing values alongside the degree to which the action is admirable, the contribution that these virtues make to how much reason ATC one has for each of the available acts would in a way be counted twice over. To avoid this kind of double-counting, the list of reason-providing values needs to be determined in such a way that none of the reason-providing values depends on any of the others in this fashion.

Finally, at the third level (c), the permissible weightings of the measures of the values may depend, at least in part, on the available options and the list of values that are at stake. This clearly allows for different weightings to be permissible in different situations. Formally, it could be that in situations in which the only values at stake are V_1 and V_2, V_1 is weighted heavily, so that it normally overrides V_2, but in situations in which a third value V_3 is also at stake, V_2 comes into its own and is more often able to override V_1. In these three ways, the account proposed here can accommodate the phenomena that have motivated theorists like Dancy to opt for a holistic and particularistic approach.

I should stress that I am not endorsing the claim that any of these three holistic phenomena genuinely occur. I simply wish to emphasize that the additive conception of the aggregation of reasons is entirely compatible with all plausible claims that have been made about these phenomena.

The advantage of this conception, however, is clear. The aggregation of reasons is not an inscrutable mystery. It is a phenomenon that clearly justifies the metaphor of "weighing" – namely, an essentially additive relationship between the specific reason-providing values and the overall value of how much reason ATC there is against the available acts.[8]

[8] Earlier versions of this essay were presented at the following gatherings: a conference on Ethics and Epistemology at the University of Shandong in Jinan, China, in 2018; an Ethics Workshop at the Croatian Inter-University Centre in Dubrovnik in 2019; and the University of Arizona's Normative Ethics Workshop in 2021. I am grateful to those audiences, to Jacob Nebel, and to an anonymous reviewer, for extremely helpful comments on those earlier drafts.

REFERENCES

Bolker, Ethan (1967). “A simultaneous axiomatization of utility and subjective probability”, *Philosophy of Science* 34: 333–40.

Broome, John (1992). *Weighing Goods*. (Oxford: Blackwell).

Broome, John (1999). *Ethics out of Economics* (Cambridge: Cambridge University Press).

Dancy, Jonathan (2004). *Ethics without Principles* (Oxford: Oxford University Press).

Finlay, Stephen (2014). *Confusion of Tongues: A Theory of Normative Language* (Oxford: Oxford University Press).

Fogal, Daniel (2016). “Reasons, Reason, and Context”, in Errol Lord and Barry Maguire (eds.), *Weighing Reasons* (Oxford: Oxford University Press).

Harsanyi, John C. (1955). “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility”, *Journal of Political Economy* 63, no. 4 (August): 309–321

Hurley, S. L. (1989). *Natural Reasons* (Oxford: Oxford University Press).

Jeffrey, Richard C. (1965 / 1983). *The Logic of Decision*, 1st / 2nd edition (Chicago, Illinois: University of Chicago Press).

Joyce, James M. (1999). *Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press).

Mongin, Philippe, and Pivato, Marcus (2016). “Social Evaluation under Risk and Uncertainty”, in *The Oxford Handbook of Well-Being and Public Policy*, ed. Matthew D. Adler and Marc Fleurbaey (Oxford: Oxford University Press).

Moore, G. E. (1903). *Principia Ethica* (Cambridge: Cambridge University Press).

Nair, Shyam (2016). “How Do Reasons Accrue?”, in Errol Lord and Barry Maguire (eds.), *Weighing Reasons* (Oxford: Oxford University Press).

Raz, Joseph (1999). *Engaging Reason* (Oxford: Oxford University Press).

Raz, Joseph (1990). *Practical Reason and Norms*, 2nd edition (Princeton: Princeton University Press).

Savage, Leonard (1954). *Foundations of Statistics* (New York: John Wiley and Sons).

Schroeder, Mark (2007). *Slaves of the Passions* (Oxford: Oxford University Press).

Sen, Amartya (2017). *Collective Choice and Social Welfare*, Expanded Edition (Cambridge, MA: Harvard University Press).

Snedegar, Justin (2018). *Contrastive Reasons* (Oxford: Oxford University Press).

Von Neumann, John, and Morgenstern, Oskar (1944 / 1953). *Theory of Games and Economic Behavior*, 1st / 3rd edition (Princeton, New Jersey: Princeton University Press).

Wedgwood, Ralph (2009). “Intrinsic Values and Reasons for Action”, *Philosophical Issues* 19: 342-363.

Wedgwood, Ralph (2013). “The Weight of Moral Reasons”, *Oxford Studies in Normative Ethics*, Vol. 3, ed. Mark Timmons (Oxford University Press, 2013): 35–58.