

Infinite Aggregation

Hayden Wilkinson

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
at the Australian National University

June 2021

© Copyright by Hayden Wilkinson 2021

All rights reserved

Except where otherwise indicated, this thesis is my own original work.

Hayden Wilkinson

24 June 2021

Abstract

Suppose you found that the universe around you was infinite—that it extended infinitely far in space or in time and, as a result, contained infinitely many persons. How should this change your moral decision-making? Radically, it seems, according to some philosophers. According to various recent arguments, any moral theory that is ‘minimally aggregative’ will deliver absurd judgements in practice if the universe is (even remotely likely to be) infinite. This seems like sound justification for abandoning any such theory.

My goal in this thesis is simple: to demonstrate that we need not abandon minimally aggregative theories, even if we happen to live in an infinite universe. I develop and motivate an extension of such theories, which delivers plausible judgements in a range of realistic cases. I show that this extended theory can overcome key objections—both old and new—and that it succeeds where other proposals do not. With this proposal in hand, we can indeed retain minimally aggregative theories and continue to make moral decisions based on what will promote the good.

Contents

Abstract	iii
List of Figures	ix
1 Introduction	1
1.1 Outline	6
2 Expanded addition	9
2.1 Introduction	9
2.2 Preliminaries	10
2.2.1 Locations	10
2.2.2 Identity/counterpart relations	12
2.2.3 Desiderata	14
2.3 Isn't spatiotemporal position morally irrelevant?	16
2.3.1 Pareto – don't take it personally	17
2.3.2 Impartiality	22
2.4 Expansionism	26
2.5 Problem cases	34
2.5.1 Spatial shifts	34
2.5.2 Temporal shifts	37
2.5.3 Chaotic effects	39
2.6 Solution	42
2.7 Remaining problem cases	49
2.8 Summary	53

2.9	Appendix: Subvaluative expansionism	54
3	The argument from chaos	57
3.1	Introduction	57
3.2	Chaos	60
3.2.1	Identities	61
3.2.2	Weather patterns	63
3.2.3	Planetary motion	64
3.2.4	The problem of cluelessness	66
3.3	Chaos, with infinite value added	71
3.3.1	Strongly anonymous theories	72
3.3.2	Weakly anonymous theories	76
3.3.3	What's left?	81
3.3.4	Position-dependent theories	83
3.4	Summary	86
4	Problems of relativity	89
4.1	Introduction	89
4.2	The basic problem	91
4.2.1	Galilean relativity	92
4.2.2	Special relativity	95
4.3	Existing solutions (and their problems)	99
4.3.1	Double lightcones	99
4.3.2	The spacetime metric	102
4.4	A better solution	106
4.5	General relativity	110
4.6	Summary	114

5	Problems with risk	117
5.1	Introduction	117
5.2	Comparing worlds: Additivity	118
5.3	Comparing lotteries: Local expectations	121
5.4	The problem	124
5.5	My proposal: Expectations of Differences	127
5.5.1	Version 1	128
5.5.2	Version 2	131
5.5.3	Weaknesses of Expectations of Differences	137
5.6	Lotteries over chaotically different worlds	139
5.7	Summary	140
5.8	Proofs	141
6	The question of fanaticism	147
6.1	Introduction	147
6.2	The case against fanaticism	152
6.2.1	The argument from intuition	152
6.2.2	The argument from impossibility	153
6.2.3	The argument from tolerance	155
6.2.4	The argument from mortality (or YOLO)	157
6.3	Background assumptions	159
6.4	A continuum argument	161
6.5	A dilemma for the unfanatical	163
6.6	Egyptology and Indology	169
6.6.1	The Egyptology Objection	169
6.6.2	The Indology Objection	177
6.7	Summary	180

7 Conclusion	181
Bibliography	183

List of Figures

2.1	W_1 in a temporal version of <i>Firing Line</i>	27
2.2	W_2 in a temporal version of <i>Firing Line</i>	27
2.3	Expansions over W_1 , centred at $\mathbf{p}$	28
2.4	Expansions over W_2 , centred at $\mathbf{p}$	28
2.5	W_1 in <i>Christmas</i>	35
2.6	Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$	36
2.7	W_1 in <i>Public Holiday</i>	38
2.8	Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$	38
2.9	W_1 in <i>Writing or Netflix</i>	40
2.10	Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$	41
2.11	Typical cumulative sums for $W_1 \succ W_2$	43
2.12	Representative cumulative sums for $W_1 \simeq W_2$	44
2.13	Cumulative sums for W_1 and W_0 in <i>Christmas</i> , with expansions starting from $\mathbf{p}$	45
2.14	Cumulative sums for W_1 and W_0 in <i>Public Holiday</i> , with expansions starting from $\mathbf{p}$	46
2.15	Cumulative sums for W_1 and W_0 in <i>Writing or Netflix</i> , with expansions starting from $\mathbf{p}$	48
2.16	W_1	50
2.17	Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$	51
2.18	W_1	52
2.19	Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$	52

2.20	W_2	55
2.21	Cumulative sums for W_2 and W_0 , with expansions starting from $\mathbf{p}$	56
3.1	The difference in total value between W_1 and W_2 , given by $1 + \sum_{i=1}^n (Y_i - X_i)$	69
3.2	Probability distributions for some random variables	74
3.3	The size of the population obtaining each possible level of value	75
3.4	A typical sum of $1 + \sum_{i=1}^n (X_i - Y_i)$	84
3.5	A typical sum of $1 + \sum_{i=0}^{r_j} (X_i - Y_i)$, as well as the area under that sum	85
4.1	$W_{\text{moving}} - W_{\text{stationary}}$	93
4.2	Cumulative sum of $W_{\text{moving}} - W_{\text{stationary}}$, with expansions starting from $\mathbf{p}$	93
4.3	$W_{\text{moving}} - W_{\text{stationary}}$, with the regions centred on $\mathbf{p}$ as above, but plotted in a coordinate system under which the velocity of the inhabitants of W_{moving} is treated as at rest.	94
4.4	$W'_{\text{moving}} - W_{\text{stationary}}$	96
4.5	$W'_{\text{moving}} - W_{\text{stationary}}$, with regions centred on $\mathbf{p}$ as above, but plotted in a coordinate system under which the velocity of the inhabitants of W'_{moving} is treated as at rest.	97
4.6	$W'_{\text{moving}} - W_{\text{stationary}}$ with a sequence of double lightcones around $\mathbf{p}$	100
4.7	$W'_{\text{moving}} - W_{\text{stationary}}$ with the same sequence of double lightcones, viewed from the perspective of the inhabitants of W'_{moving}	101
4.8	A sequence of regions each comprised of points within some fixed $ s^2 $ of $\mathbf{p}$	103
4.9	$W'_{\text{moving}} - W_{\text{stationary}}$	104
4.10	$W_{\text{right}} - W_{\text{left}}$	105
4.11	$W_{\text{right}} - W_{\text{left}}$	108
4.12	A sequence of regions each comprised of points within some fixed g' of $\mathbf{p}$	112

6.1	Cumulative probability graphs for L_1 and L_2 ; here and below, $\mathcal{O}_{\preccurlyeq_O O}$ denotes the set of outcomes in $\mathcal{O}$ as which O is at least as good.	173
6.2	Cumulative probability graphs for L_{risky} and L_{safe}	174
6.3	An example of background uncertainty B —a Cauchy distribution	174
6.4	Cumulative probability graphs for $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$	175
6.5	Probability distribution of B	176

Introduction

Suppose you found that the universe around you was infinite – that it extended infinitely far in space or in time and, as a result, contained infinitely many persons. How should this change your moral decision-making? Radically, it seems, according to some philosophers.

Note first of all that we each have some moral reason to make the world better than it would otherwise be, at least somewhat and sometimes. This need not be because some form of pure consequentialism is true. Perhaps the true moral theory upholds other considerations—constraints, prerogatives, or reasons from other sources. But any plausible theory will, in at least a large class of moral decisions, recognise *some* reasons to bring about the best available outcome.¹ For instance, when one devotes some of one's resources to aid strangers through charity, it seems that one has some reason to use those resources as effectively as possible—to bring about the best available outcome.²

On top of this, it is plausible that (in at least some decision contexts) one outcome can be at least as good as another if and only if it contains at least as great a total aggregate of value, impartially construed.³ Again, this need not be because *only* total

¹Rawls (1971, p. 30) himself claims that any plausible deontological theory satisfies this criterion: "It should be noted that deontological theories are defined as non-teleological ones, not as views that characterise the rightness of institutions and acts independently from their consequences. All ethical doctrines worth our attention take consequences into account in judging rightness. One which did not would simply be irrational, crazy."

²For a defence of that claim, see Pummer (2016).

³This definition excludes any narrow person-affecting theory of value. It also excludes any *monistic* theory of value under which value does not admit an additively separable representation (e.g., egalitarianism, maximin, averageism). This exclusion is not because such views escape the infinitarian worries described below—typically, they don't—but for brevity. And it doesn't exclude views with a total utilitarian axiology

value matters; some form of pluralism may be true. A range of the plausible moral theories are *minimally aggregative* in this sense: in a large class of comparisons, they evaluate one state of affairs as better than another only if the former contains a greater total aggregate of value; and they say that we have reason to bring about that better world.

Here and throughout I will be interested in this broad category of minimally aggregative moral theories. This entire category appears to face serious problems if our universe is infinite in the relevant sense.

Here is a simple case in which minimally aggregative theories run aground. Suppose you have two actions available to you. You can either rescue one person from death, or rescue five others from death (each of whom would be better off not dying). And, either way, the universe will contain infinitely many other persons with quite good lives. We can represent the outcome of saving one with W_1 , and that of saving five with W_5 . Each outcome, or world, contains an infinite plurality of persons $\{p_a, p_b, p_c, \dots\}$. And here the moral value of each person's life is represented on an interval scale as 0 (if they die now) or 1 (if they get to enjoy the remainder of their life).

	p_a	p_b	p_c	p_d	p_e	p_f	p_g	p_h	p_i	p_j	$\dots$
W_1 :	1	0	0	0	0	0	1	1	1	1	$\dots$
W_5 :	0	1	1	1	1	1	1	1	1	1	$\dots$

For these worlds, we cannot say that either has the greater total sum. Even *if* we extend the real numbers by defining ∞ (or even by defining the transfinite ordinals or cardinals as well), the total sum of these two worlds will still be the same infinite number; their totals will be equal. If our only means of evaluating worlds is by their total, then we cannot say which of these worlds is better. But W_5 is clearly better! Five people are

or a prioritarian axiology in the style of Parfit (1997). Nor does it exclude pluralist views which give any weight at all to either of those axiologies.

better off, and only one worse off (by the same amount). Intuition says that W_5 is better and that, absent any other considerations, we ought to save the five. But minimally aggregative theories cannot say this.

More generally, suppose that our universe does have infinite spatial volume or temporal duration. And suppose further that it contains infinitely many moral subjects, infinitely many of whom have quite good lives (with positive value greater than some fixed $\epsilon > 0$ on some interval scale) and infinitely many of whom have quite bad lives (with value less than $-\epsilon$). Then the total sum of value over all good lives minus the disvalue of all bad lives is *undefined*.⁴ And so it will remain even if you make some great change to the world—if you like, add, remove, or improve any finite number of lives. The total sum of value in the world will be unchanged—it will still be undefined (or, in the earlier case, infinite). If the universe is indeed like this, no matter what actions you or I might take, then we cannot ever say that the outcome of any available action will contain greater total value than any other. And so, by minimally aggregative views, none can be better than any other: they are all incomparable to one another. So such views will give us no reason to bring about any of them over any other. And this extends to a large class of decisions in universes like this: perhaps all pure rescue cases, or all decisions of how to allocate resources for charitable purposes.⁵

If you accept a minimally aggregative moral theory, you might then hope that our own universe is not infinite in this way. But it seems that you are not so lucky. Contemporary physics suggests that it is indeed infinite in the relevant sense. In particular, two leading theories of cosmology each independently imply that our universe is guaranteed to have an infinite population. The *flat-lambda* model—a cosmological model accepted

⁴The situation is analogous to that faced by an expected utility maximiser with an option to play the Pasadena game (see Nover and Hájek, 2004; Hájek and Nover, 2006). In that case, the expected utility of the option is the sum of infinite positive value and infinite negative value, giving an undefined total. Here, the total value of each outcome forms such an undefined sum.

⁵Something akin to this argument appears in each of: Nelson (1991), Hamkins and Montero (2000), Bostrom (2011), and Askill (2019).

by a majority of physicists—predicts that our universe will tend towards a stable state and will then remain in that state for infinite duration (Wald, 1983; Carroll, 2017). And the *inflationary view*—also widely accepted—posits that our world is spatially infinite, containing infinitely many other ‘bubble’ universes beyond our cosmic horizon (Guth, 2007). Under either theory, the universe will contain a great many of any given physical phenomenon. This includes things that we tend to think have moral value, e.g., a human brain experiencing the thrill of reading philosophy for a given duration, or a person at some time holding a preference for some events to occur as they in fact do. That thrill and that preference-holding can be instantiated physically, with the right number of the right particles at the right energy levels and in the right places relative to one another. And it can be repeated at other locations to produce the same additional value no matter how many such thrills or preferences already exist (according to minimally aggregative theories). Each of the flat-lambda model and the inflationary view each independently predict that our universe, in its infinite volume, will contain infinitely many⁶ such small-scale physical phenomena (Garriga and Vilenkin, 2001; Linde, 2007; de Simone et al., 2010; Carroll, 2017)⁷, and so infinitely many such thrills. If we sum up the value of all of those thrills, we have an infinite total. So, our world will contain infinite (or, if we include events with negative value, *undefined*) total value, at least by those physical theories. And if either of those theories are true, then minimally aggregative theories fail us.

But even if we are uncertain whether those physical theories are true, minimally aggregative theories are still under threat. Given that reputable physicists endorse these theories, as good epistemic agents, we should assign some non-zero subjective probabil-

⁶The set of such token events will be *countably* infinite. For any given phenomenon, I’ll assume, each token must occupy some (exclusive) non-zero volume of spacetime. For illustration, a human brain to experience a given quantity of pleasure, it requires some non-zero spatial volume and some non-zero, finite duration. So we can only fit a countably infinite number of those token events into the world.

⁷Physicists are hesitant to speak of infinite quantities (exceptions include Garriga and Vilenkin, 2001; Gott, 2008), as these can never be empirically confirmed. Nonetheless, this conclusion is entailed by both of the above views.

ity⁸ to at least one of them being true. And any non-zero probability is enough to cause problems in practice. Without a plausible method for comparing infinitely-valued outcomes, agents will also typically be unable to compare lotteries that include some probability of such outcomes. And so, with some probability of infinitely valued-outcomes resulting from their actions, they will be unable to compare any pair of available actions in terms of their *instrumental* betterness. So even in this situation, agents need a method of comparing infinitely-valued outcomes.

Indeed, problems arise even if we are certain that those physical theories are false. After all, physical laws are contingent. It is possible, both metaphysically and epistemically, that they had been different, and that our universe had been infinite. In that hypothetical, minimally aggregative theories would still deliver absurd results. And this is reason to reject them—we tend to think that true moral theories are *necessarily true*, and so would have been true even if contingent physical facts were a little different. If minimally aggregative theories fail in some possible worlds, then that failure gives us reason to reject them even in our own world.

Given such failure, what are we to do? Must we abandon anything resembling a minimally aggregative theory?

In this thesis, I will argue that, no, we need not. We can extend minimally aggregative theories to give plausible judgements in an infinite universe. And we can do so in a way that avoids various other serious objections.

Admittedly, other extensions have already been proposed. Each of Vallentyne (1993), van Liedekerke and Lauwers (1997), Vallentyne and Kagan (1997), Lauwers and Vallentyne (2004), Bostrom (2011), Arntzenius (2014), and Jonsson and Voorneveld (2018) offer an alternative (see Askill 2019 for a survey). But, as I'll argue, these alternatives are all unsatisfactory. They may allow us to make *some* comparisons of infinite outcomes, but many do not allow us to compare the outcomes we may realistically face (see Chapter 3),

⁸Likewise, the *evidential* probability that at least one of these theories is true is non-zero.

or they imply absurdity when we consider other properties of our universe (see Chapter 4), or they do not allow us to compare *lotteries* over infinite outcomes in a plausible way (see Chapter 5).

I have what I consider to be a successful solution to the problems of infinite aggregation—a proposal that faithfully extends minimally aggregative theories to the infinite context, that compares infinite outcomes in a plausible way, and that overcomes the most serious objections that can be thrown at it. Given the existence of such a solution, we can indeed retain minimally aggregative theories; in our actual universe and in cases we may actually face, we can still make decisions based on what will promote the good.

1.1 Outline

This thesis proceeds as follows.

In Chapter 2, I begin by introducing the basic formal framework I will be working with. I then describe my proposal for comparing infinite outcomes: a so-called *expansionist* theory. I also explain several key differences between my proposal and those in the literature, justify those differences, and head off some immediate objections.

In Chapter 3, I present what I consider the single most compelling argument in favour of my proposal: that, unlike almost all other proposals, it can still compare outcomes when those outcomes differ by long-lasting *chaotic* effects. Of course, our universe contains many chaotic systems, and agents like us often interact with them. So we must often choose between such outcomes. If our theory of infinite aggregation cannot compare them, then it will often fail to deliver comparisons in practice.

In Chapter 4, I address a serious objection for many theories of infinite aggregation—one that my own proposal falls to as well, if not modified further. The verdicts of proposals like mine are sensitive to the positioning of valuable events in space and time, but in our universe the position of an event in space and time is in some sense relative to the

perspective of the observer. But it would be absurd for moral betterness to be relative to the perspective of different observers. In Chapter 4, I show that my basic proposal can be modified to avoid such problems.

Up to this point, the discussion focuses on comparing *outcomes*, not on comparing *lotteries* over outcomes. But in practice we are never certain of which outcome will be brought about by one action or other, so this is insufficient. We need a rule for comparing lotteries over infinite outcomes as well. In Chapter 5, I present such a rule. I also demonstrate a serious problem for other such rules from the existing literature.

But my proposal for comparing lotteries invites a possible objection. It is *fanatical*: it implies that it is better to produce a tiny probability of infinite gain over a certainty of finite gain, no matter how enormous the finite gain or how tiny the probability of infinite gain. You might well find this implication implausible. I address this objection in Chapter 6 and show that, as it turns out, there are compelling arguments in favour of allowing such fanatical verdicts. To deny all such verdicts is to bring on even more severe objections to one's theory.

Chapter 7 is the conclusion.

Expanded addition

2.1 Introduction

In an infinite universe, minimally aggregative views appear to face a dire problem: in any decision scenario that agents like us might ever face, they appear unable to say that any outcome will be better than any other. It appears that they rule out agents having any moral reasons based on which acts will bring about better consequences.

But perhaps not all minimally aggregative theories must fail in this way. Others have proposed alternative theories that they claim avoid the problem (e.g. Vallentyne, 1993; van Liedekerke and Lauwers, 1997; Vallentyne and Kagan, 1997; Lauwers and Vallentyne, 2004; Bostrom, 2011; Arntzenius, 2014; Jonsson and Voorneveld, 2018). In this chapter and throughout this thesis, I focus on one approach that appears particularly promising: *expansionism*.

Specifically, I propose and defend a *spatiotemporal* version of expansionism.¹ This contrasts with the previous version given by Vallentyne and Kagan (1997), which is agnostic as to what the basic locations of value are. Those locations might be individual persons, or positions in spacetime, or something else. Without knowing which, we often cannot apply the rule in practice. And if we adopt the obvious choice of locations—persons—then it turns out that we cannot give plausible judgements (or any judgements

¹When discussing expansionism, Askill (2019, pp. 71-87) also focuses on the spatiotemporal version. I agree that this is the most compelling version, as I argue in §2.3, Chapter 3, Wilkinson (2021), and Wilkinson (n.d.(a)).

at all) in even some fairly mundane cases (see §2.3 below). So I argue for a *spatiotemporal* version of expansionism: one which adopts spacetime positions as our basic locations, and uses their natural structure to give more plausible judgements than we otherwise could.

Approaches like mine face objections (see §2.3). One is that they seem to abandon a crucial feature of aggregation—its *impartiality*, including over properties such as *where* each person happens to be. Another is that they give counterintuitive judgements in key cases (see Cain, 1995). And an objection of my own is that, in some fairly mundane cases some of which we might actually face, my approach seems to fall silent. I'll argue that these objections are not decisive. For one, we will see that basic tenets like impartiality do not rule out spatiotemporal views entirely. For two, we must accept counterintuitive judgements in some cases to avoid far more implausible judgements in other (more realistic) cases. And, for three, with a bit of work we can construct at least some spatiotemporal views which *do* give plausible judgements even in many problem cases.

2.2 Preliminaries

2.2.1 Locations

We want to compare worlds on the basis of the value they contain. We cannot compare their totals, so we look to the individual instances of value – *local values*. We could individuate and identify these across worlds in different ways, such as: by who the *person* is whose life contains the value; or which person-time-slice obtains it; or in which generation² it is obtained; or at which *position* in space and/or time they arise; or perhaps some other way. For any of these, local value is associated with token entities of some

²Although generations are commonly used as locations in the social welfare literature, I agree with Askill (2019, p. 12-3) that they are typically underspecified and, when precisified, bring implausible results.

common type which exist (or have counterparts) across different worlds. I'll call those tokens *locations*. For reasons given in Section 3, I focus on the two most plausible types of locations: persons, and spacetime positions.

Whichever locations we adopt, each world W_i contains a set of them: $\mathcal{L}_i = \{l_1, l_2, l_3, \dots\}$.³ The subscripts j for each l_j may be arbitrary or may reflect some natural, essential structure of locations (as for spacetime positions). The worlds we compare will often contain the same (or counterpart) locations, in which case we can use just the one set $\mathcal{L}$. (More on this below.)

For each world W_i , some value function $V_i : \mathcal{L}_i \rightarrow \mathbb{R}$ maps each location to its value. So, in W_1 , the local value at location l_a is given by $V_1(l_a)$. I assume that local values admit at least a cardinal representation on the reals—that is, a representation which is unique at least up to affine transformations. (This means that, for any two alternative representations v and v' of the same value, $v = a \cdot v' + b$ for some real b and positive a .) But I will not assume that local values have a natural zero or a natural unit (as is necessary for ratio-scale or translation-scale representations, respectively).

I also assume that locations—whatever type they might be—can be positioned in space and time within each world. If locations themselves are spacetime positions, that's easy enough—they each have an essential position. If instead they are persons, they still each occupy spacetime positions. For simplicity, let's treat each person's position as a single point: perhaps the point of their birth, or the midpoint of all positions they ever occupy. (Nothing below hangs on which we choose.) So, whichever form our locations take, they are assigned a point in spacetime $\mathbf{x} \in \mathbb{R}^4$, or (x, y, z, t) . This is represented in

³This set may be countable or uncountable. If the latter, we won't be able to list our locations with natural-numbered subscripts. But much of what follows does not depend on which it is.

If our locations are persons, we can assume that the set will be countable. After all, persons require some positive, finite volume and duration. We could set some minimum physical volume and duration for a person to matter morally, and only a countable infinity of persons will fit into any spacetime (whether Euclidean or not).

If our locations are spacetime positions, we can assume that the set will be uncountable. After all, these are positioned densely.

a Cartesian coordinate system on four dimensions and is unique up to translation and scalar multiplication.⁴

2.2.2 Identity/counterpart relations

My strategy for comparing worlds—the strategy of almost everyone in the literature to date—is to take sets of individual locations and to compare their local values across worlds. But, to compare local values across worlds, we need the same locations (or, alternatively, their *counterparts*) to exist across worlds. How do we match up locations with each other?⁵ If we cannot answer this, then my strategy will fail immediately. I'll describe what I think are the most plausible methods of identifying locations with themselves or their counterparts, whether the relevant type of locations is persons or spacetime positions.

Suppose that locations are individual *persons*. In this case, I propose that we stick to our theory of personal identity, whichever is correct—that our identity or counterpart relation use criteria analogous to those used by the identity relation between a person's past, present, and future selves. If all that's required for one's past and future selves being the same person is that they share some bundle of qualitative properties, let that be sufficient for a person in one world to be the same as a person in another world. Alternatively, the identity relation between past and future selves might require physical continuity. If so, then let our transworld relation hold between two persons if and only

⁴I am assuming for now that spacetime is *Euclidean*: that points have a unique coordinate representation up to translation and scalar multiplication; and that every two points have a unique inner product and geometric distance between them (up to scalar multiplication). This distinguishes the spacetime I'm working with from Minkowskian and pseudo-Riemannian spacetimes which are more accurate representations of our universe and which each admit different coordinate representations under which inner products and geometric distance are not fixed. I accept that the assumption of Euclidean spacetime is unrealistic, but necessary to avoid taking us too far afield. I wish to demonstrate that my method overcomes basic problems in even the Euclidean setting before attempting to deal with problems introduced by relativistic spacetimes. In §4.5, I show that these further problems can be resolved as well.

⁵Note that I am not committing myself to a particular position in the debate—between Lewis (1971) and, for instance, Kripke (1980, p. 45)—on whether transworld identities exist. The following relation works just as well whether it's an identity relation or just a privileged counterpart relation, so long as it forms an equivalence relation.

if they share some common history of identical events. (To determine which events are identical, we can use the following proposal.)

Alternatively, suppose that our locations are *spacetime positions*. Then we may not have such a theory of identity. But that is no trouble—there is an obvious identity/counterpart relation, at least in any pairs of worlds we'll ever need to compare. Since our actions necessarily cannot change the past, any such worlds will share the same past events (at the same positions). In worlds like these with common histories up to the present, let us map all past positions to those occupied by the same events (e.g., we can map the position of Runnymede in 1215 in one world to the same position in every other world, as that's where the signing of the Magna Carta occurs in all of them). Then we can map future positions too: each such x is uniquely specified by its spatial and temporal distance from (any four) past points⁶, so we can specify its transworld identities/counterparts as the positions which are also those same distances from those same points.

This matches our intuitive sense of the 'same' positions (e.g., "1 metre in front of me, 1 second in the future" will be the same point no matter what I do). It also preserves the distances among all spacetime points. And it ensures that any two worlds we must ever compare—the outcomes of two actions available to us—will share the same set of positions, which will prove useful in what follows. The same cannot be said for worlds which do not have some shared history, but in practice we'll never need to choose among such worlds.⁷

⁶This clearly holds in Euclidean spacetime. In relativistic spacetimes, the counterpart relation becomes more complicated (see Chapter 4).

⁷Note also that this proposal does not require that counterpart positions are metaphysically identical, nor that they have essential properties. I propose merely that, in our moral evaluations, we treat points as equivalent when they satisfy these relations. This is compatible with both spacetime substantivalism and relationism.

2.2.3 Desiderata

To compare worlds, we want an ‘at least as good as’ relation $\succsim$ on the set of (metaphysically possible) worlds $\mathcal{W}$. I assume that this relation is reflexive and transitive.⁸ The asymmetric component ($\succ$) will be our strict betterness relation, and the symmetric component ($\simeq$) our equality relation.

Defining a satisfactory betterness relation is no easy feat. We must make some hard choices. From the social welfare literature, we have a collection of nasty impossibility results – in particular, from Zame (1999) and Lauwers (2010).⁹ They show that it is impossible to construct¹⁰ a version of $\succsim$ which is transitive, complete, and does not violate either *Pareto* or *Finite Anonymity*—both of which are basic requirements for a plausible betterness relation.¹¹

Pareto holds that, if two worlds contain precisely the same locations all with precisely the same local values, then those worlds are equally good; and if we made one of those worlds better at some locations, then it would then be better than the other. And this seems highly plausible, for any type of location! It can be stated more formally as follows.

Pareto (over locations): For any worlds W_1 and W_2 containing the same locations $\mathcal{L}$, if $V_1(l) \geq V_2(l)$ for all $l \in \mathcal{L}$, then $W_1 \succsim W_2$.

If, as well, $V_1(l_i) > V_2(l_i)$ for some $l_i \in \mathcal{L}$, then $W_1 \succ W_2$.

Meanwhile, *Finite Anonymity* is necessary for our aggregation to be genuinely impar-

⁸Transitivity of moral betterness has its critics, e.g., Temkin (2014). It has also received compelling defences from, e.g., Broome (2004); Huemer (2008); Nebel (2018); Dreier (2019). I find it overwhelmingly plausible so, in keeping with the infinite aggregation literature to date (within both moral philosophy and welfare economics), I will assume without argument that it holds.

⁹Lauwers shows that the following impossibility holds even for the domain of worlds with local values of only 0 and 1.

¹⁰It has been shown that such relations do exist, but cannot be constructed (see Svensson, 1980). This is unsatisfactory for a moral rule, by which we would like to explicitly define the criteria for betterness.

¹¹Note that both Zame and Lauwers only prove this result for worlds that are *infinite utility streams*: worlds in which the locations form a totally ordered set, starting from some l_1 . Such worlds are usually interpreted as taking locations to be entire *generations* of people (and as applying Pareto and Finite Anonymity over generations). But the results can be trivially extended to the broader setting of all worlds with a countably infinite number of locations, and to any type of locations (see Askill, 2019, p. 39-44).

tial. If two worlds only differ by swapping the local values of two locations, then those worlds must be equally good.

Finite Anonymity (over locations): If for some $l_a, l_b \in \mathcal{L}$ we have $V_1(l_a) = V_2(l_b)$, $V_1(l_b) = V_2(l_a)$ and $V_1(l) = V_2(l)$ for all other $l \in \mathcal{L}$, then $W_1 \simeq W_2$.

Since $\succsim$ must be transitive, Finite Anonymity implies that we can permute *any* finite number of local values and the resulting world remains equally good. But note that it does not imply that the result of an *infinite* number of permutations—shuffling infinitely many local values—will be equally good. (More on this below.)

Like Pareto, Finite Anonymity seems highly plausible. I find both Pareto and Finite Anonymity hard to deny, at least for *some* type of location (if not multiple types). Perhaps they hold for persons, or perhaps for positions. (Note that, in finite worlds, they hold for *all* types.) Whichever it is, the Zame-Lauwers impossibility will apply. Does this mean that any proposal for a betterness relation over infinite worlds is doomed from the start? I don't think so. To avoid the impossibility while hanging onto these principles, we can simply allow that $\succsim$ is *not complete*. This would mean that there is some W_1 and W_2 in $\mathcal{W}$ such that neither $W_1 \succsim W_2$ nor $W_2 \succsim W_1$. This may seem counterintuitive, but it is better than the alternative: having $\succsim$ violate either Pareto or Finite Anonymity (for every type of location), or else fail to be transitive. So, in this chapter and throughout this thesis, I won't be seeking a complete relation. No such relation would be plausible.

But a moral theory still needs to give verdicts in mundane cases that agents like us *actually* face. Even if we cannot demand that the $\succsim$ relation be *complete* (over the set of all metaphysically possible worlds), it must still be '*minimally complete*'. By this I don't mean anything as simple as completeness over the set of all epistemically possible worlds, or all physically possible worlds. Instead, I mean this: for any decision-making agent who will *actually* exist, for any decision they will actually face, and for any two worlds which would be brought about if they took one or another action in that decision, those worlds

must be comparable by the $\succsim$ relation. If so, then $\succsim$ is minimally complete. Or, in short: $\succsim$ must never *actually* fall silent on us. It's fine for it to be silent in exotic problem cases—such as when deciding between worlds which don't share the same history, or between highly gerrymandered worlds. But, if it falls silent in decisions I actually need to make, that's a serious failing.

Note that it would be impossible to prove that any given $\succsim$ relation is minimally complete. Instead, I will simply show that some proposed $\succsim$ relations (in the next section and Section 5) fail to be.

To sum up, my desiderata for infinite aggregation are that:

- the betterness relation is reflexive and transitive;
- it appears minimally complete, and gives plausible judgements in those cases it covers;
- it satisfies Pareto *over spacetime positions*; and
- it satisfies Finite Anonymity *over spacetime positions*.

Why over spacetime positions but not (also) over persons? And why Finite Anonymity rather than a stronger, infinite form of anonymity? I'll justify these choices below.

2.3 Isn't spatiotemporal position morally irrelevant?

My view gives verdicts which are sensitive to where value is positioned in spacetime, as do almost all views proposed in the social welfare literature. In this section, I will pre-emptively (and all too briefly) address the most common objections to such views. For an extended defence, see Chapter 3.

2.3.1 Pareto – don't take it personally

We cannot endorse both Pareto *over spacetime positions* and Pareto *over persons*. The two principles are incompatible. (Likewise, Pareto over any location-type will be incompatible with Pareto over almost any other.) But Pareto over persons seems highly plausible, so this constitutes a serious objection to my view and, indeed, to many views in the literature.

To see why the two principles are incompatible, consider *The Shuffle*.¹²

Example: The Shuffle

You are hosting a line dancing event, to which infinitely many people have shown up—many more than you expected! Fortunately, you have two dances prepared which can accommodate infinitely many dancers. Each dance involves the dancers standing in single file. And, for both, the moves required of each dancer depend on their position in the line, so dancers will find some positions more strenuous and less pleasant than others. If you ask the dancers to perform one dance or the other then the moral value obtained by each dancer (and at each position) is given by W_1 or W_2 , respectively.

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	$\cdots$
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\cdots$
$W_1 :$	1	0	1	0	1	0	1	0	1	0	$\cdots$

	p_1	p_2	p_4	p_6	p_3	p_8	p_{10}	p_{12}	p_5	p_{14}	$\cdots$
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\cdots$
$W_2 :$	1	0	0	0	1	0	0	0	1	0	$\cdots$

¹²Cases like this, which demonstrate the disagreements between different Pareto principles, originate with Cain (1995, p. 401-3). Cain considers them a decisive objection to spatiotemporal views like mine. But, as we'll see below, views that endorse Pareto over persons face problems that you might consider even worse.

Here, the worlds W_1 and W_2 contain precisely the same persons ($p_1, p_2, \dots$) and the same spatiotemporal positions ($x_1, x_2, \dots, x_i, \dots$ where x_i is the i th position in line). But many of those persons are at different positions in each world—they're *shuffled*. Compared to W_1 , every even-numbered person from p_4 onwards is positioned earlier in the line. And every odd-numbered person from p_3 onwards is positioned later.

Which dance would bring the better outcome? Look at the spacetime points: all points do at least as well in W_1 , and some do strictly better ($x_3, x_7, \dots, x_{4n+3}, \dots$). Pareto *over positions* says that $W_1 \succ W_2$.

Look at the persons: every person bears the same amount of value in W_1 as in W_2 , they're just positioned less densely. Pareto *over persons* says that $W_2 \simeq W_1$, in conflict with Pareto over positions.

The judgement from Pareto over persons may seem compelling – no person would be worse off in W_2 , so what's the harm in choosing that dance? As a person myself, I wouldn't mind finding myself in W_2 . And you might think that improving outcomes for people is what matters morally, rather than simply increasing value in the abstract. You might think, as Askill (2019, p. 84) does, that "...Ethics is concerned with people. It is not particularly concerned with the pattern of utility across spacetime." I have sympathy with this thought, but we also have reason to reject it. I provide only a brief argument here, but see Chapter 3 for a full defence.

The reason is this: unlike Pareto over positions, Pareto *over persons* implies *radically* widespread incomparability among worlds, even in quite mundane cases. So any theory that satisfies Pareto over persons fails to be minimally complete.

To argue this, I must first make two assumptions explicit. The first: $\succ$ is a *qualitative* relation. It ranks any pair of worlds (W_1, W_2) the same way as it would rank any other world pair (W_3, W_4) with the same qualitative properties. Note that it need not also be a *qualitative internal relation*, which means that it must rank *worlds* W_1 and W_2 as equally

good if they are qualitative duplicates. No, that would violate Pareto immediately. Instead I am assuming merely that a *pair* of worlds will be ranked the same way as any qualitative duplicate of that *pair* of worlds. In effect, if we take a pair of worlds and swap around (or replace) the identities of their inhabitants, in exactly the *same* way in *both* worlds, then we don't change the ranking of the resulting worlds.¹³

The second: we can change many of the qualitative features of persons however we want (including at least their positions) without changing their identity (or counterpart relationships). This implies that worlds W_1 and W_2 above are both metaphysically possible – we can shift infinitely many of the even-numbered persons leftwards by any distance, and odd-numbered persons rightwards, while preserving their identities. Both this and the previous assumption are hard to deny, particularly if we think that such problem cases are possible.

Now consider *Firing Line*.

Example: Firing Line

Infinitely many persons are lined up in front of you, extending infinitely far into the distance. You have a choice to make. Either let every *second* person in the line be killed, or let every *fifth* person be killed. Their lives will otherwise be quite pleasant, and equally so.

Here are the local values for *Firing Line*. The i th person in the line corresponds to p_i , at position x_i . And note that the beneficiaries in W_2 are more *densely* positioned over this sequence of locations than the beneficiaries in W_1 . In an important sense, there are *more*

¹³This assumption is often called *Relative Anonymity* in the social welfare literature, e.g., by Asheim et al. (2010, p. 10).

beneficiaries in W_2 . (More on this below.)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	$\dots$
	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	$\dots$
W_1 :	1	0	1	0	1	0	1	0	1	0	1	$\dots$
W_2 :	1	1	1	1	0	1	1	1	1	0	1	$\dots$

From the assumptions above, there are (metaphysically) possible worlds W_3 and W_4 with the following local values, and which share every qualitative feature with W_1 and W_2 , respectively.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	$\dots$
	p_1	p_{15}	p_3	p_{35}	p_5	p_{55}	p_7	p_{75}	p_9	p_{10}	p_{11}	p_{95}	p_{13}	p_{115}	p_2	p_{135}	$\dots$
W_3 :	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	$\dots$
W_4 :	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	$\dots$

W_3 and W_4 have been modified from W_2 and W_1 , respectively, by a slightly complex procedure. Most persons have been moved to new positions: the persons in the sets $\{p_{15}, p_{35}, p_{55}, \dots\}$ and $\{p_2, p_4, p_6, p_8, p_{12}, \dots\}$ have each swapped positions with a member of the other set; while the remaining persons stay put. And some persons ($p_{15}, p_{35}, p_{55}, \dots$) have been given different values (1 and 0 in W_1 and W_2 , respectively, but 0 and 1 in W_3 and W_4). Those persons do strictly better in W_3 than in W_2 , and strictly worse in W_4 than in W_1 . All other persons have the same local values in W_3 as in W_2 , and in W_4 as in W_1 . Assuming Pareto over persons, this means that $W_3 \succ W_2$ and $W_1 \succ W_4$.

But the persons in these worlds have been rearranged so that W_1 is qualitatively identical to W_3 , and W_2 to W_4 (or so we can assume, changing each person's other qualitative properties such that they are qualitatively identical to the person inhabiting the same position in the other world). So the pairs must have the same ranking: $W_1 \succcurlyeq W_2$ iff

$W_3 \succcurlyeq W_4$.

Now suppose $W_2 \succcurlyeq W_1$. Then $W_2 \succcurlyeq W_1 \succ W_4 \succcurlyeq W_3 \succ W_2$. So this implies a violation of transitivity. So $W_2 \not\succcurlyeq W_1$.

But if we suppose that $W_1 \succcurlyeq W_2$, we can construct a pair of additional worlds W_5 and W_6 by a similar process such that $W_1 \succcurlyeq W_2 \succ W_3 \succcurlyeq W_4 \succ W_1$. That will lead to a violation of transitivity too. So $W_1 \not\succcurlyeq W_2$.

Therefore, in *Firing line*, W_1 and W_2 would be incomparable. Not equally good, but *incomparable*. But it was not so exotic a case. As it happens, we face a roughly analogous case whenever we influence infinitely many future persons, benefiting one set or another (with or without preserving their identities – see *ibid.*: 114-35). If we maintain Pareto over persons, incomparability would arise in (a close variant of) every case raised in the remainder of this chapter.¹⁴ (Note that the same implication doesn't follow from Pareto over *spacetime positions*—we cannot construct the W_3 and W_4 needed for the argument, as altering the spatiotemporal positioning of value makes the worlds qualitatively different.)

This is one of several reasons to reject Pareto over persons. With the assumptions above (which are hard to reject), it implies that we face incomparability in even mundane cases. It implies that the vast majority of the outcomes we produce are no better nor worse than the alternatives. And this is implausible. Our moral theory must say *something* in cases like *Firing line*.

On top of that, our moral theory must say the *right* thing which, at least according to intuition, is that it's worse to let every second person die than to let every fifth person die. (Or, for an analogous and even more compelling case, it would be worse to let every second person die than to let every 99th person die!) Intuitively, the set of every second person seems larger than the set of every fifth: in particular, take any interval of three or more consecutive people and there are more of them. And of course the more people

¹⁴The preceding argument is loosely adapted from Askill (*ibid.*), although Askill accepts this incomparability rather than abandon Pareto over persons.

are killed, the worse the outcome is. And Pareto over persons rules that we cannot give this verdict. But Pareto over *positions* does not stop us, so I am quite happy to seek a $\succsim$ relation which endorses Pareto over positions but not persons.¹⁵

2.3.2 Impartiality

Here is another common objection to views like the one I'll present below. If we take the spatiotemporal position of value into account, we abandon a fundamental part of aggregative views—we fail to be *impartial*.

Impartiality is defended by aggregationists (Mill, 1861; Sidgwick, 1907; Singer, 1972; Parfit, 1984), in roughly the following form: it is morally no better for one to obtain the good than for another to. So too, it is no better for one group of persons—beneficiaries—to obtain the good than for another to. This applies whenever the amounts of the good obtained by the beneficiaries would be the same for either group, and the *number* of those beneficiaries is the same. And it applies regardless of *where* those beneficiaries are positioned and *who* the beneficiaries are.

But recall Finite Anonymity. If two worlds contain the same persons at the same positions, and they differ only by a rearrangement of any finite number of values from one set of beneficiaries to another, then Finite Anonymity (over *either* persons or positions) guarantees that the worlds are equally good. In practice, if we could either benefit n people by some amount(s) or n others by the same amount(s), and they start with the

¹⁵Despite the conflict between the two forms of Pareto, Pareto over persons does still have some intuitive draw. Given this, it may seem desirable to have $W_a \succsim W_b$ (or $W_a \succ W_b$) whenever Pareto over persons says so *and* Pareto over positions does not deny it. But Firing Line demonstrates that this can bring problems. In Firing Line, Pareto over persons says that $W_3 \succ W_2$ and $W_1 \succ W_4$ and Pareto over positions doesn't deny it, but that leads us to judging W_1 and W_2 as incomparable.

Given this, we must be cautious of following the verdicts of Pareto over persons even in narrow circumstances. But, I will briefly note, one situation in which it seems harmless to do so is when, across the two worlds we are comparing, only finitely many persons have their spatiotemporal positions differ between worlds. And my own proposal below satisfies Pareto over persons in those cases.

Another situation in which it seems harmless is when Pareto over positions (or some strengthening of it) says that two worlds are equally good. If so, and if Pareto over persons says that one world is strictly better, we might break the tie according to whatever Pareto over persons recommends. My own proposal doesn't do this, but it does seem a plausible modification to make.

same amount of the good, then we should be indifferent between the two options.

Likewise, we might take two worlds with the same persons at the same positions and we rearrange their local values, and *also* rearrange which person is at which position. Still, we can rearrange any finite number of values and any finite number of persons and Finite Anonymity (over *either* persons or positions) guarantees that the worlds are equally good. So it seems that, if $\succsim$ satisfies Finite Anonymity (even just over positions), it preserves the impartiality we want.

Not all aggregation methods do satisfy Finite Anonymity. Discounting does not (e.g. Koopmans, 1972). It involves summing local value in the usual manner, but only after *discounting* that value by some function of how far it is from the agent's position. This will often allow us to compare infinite worlds, since a harsh enough discount rate can bring the total value down to a finite number. But it violates Finite Anonymity – it says that distant value is worth less (and very distant value almost *worthless*). It says that we should prefer to benefit our neighbours over benefiting distant strangers. Discounting may be a convenient solution to the problem, but I agree with Parfit (1984, p. 480-6) and the others that it lacks the impartiality we want and so is morally indefensible. And that indefensibility is captured by its violation of Finite Anonymity.

But Finite Anonymity only requires indifference up to rearrangements of *finitely* many local values. What about rearrangements of *infinitely* many local values? Does impartiality require that we adopt a stronger principle of *Infinite* Anonymity? No, it doesn't. To see why not, consider these worlds. They'll be familiar from above, but this time have the same persons (p_i) at the same positions (x_i).

	p_a	p_b	p_c	p_d	p_e	p_f	p_g	p_h	p_i	$\dots$
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	$\dots$
$W_1 :$	1	0	1	0	1	0	1	0	1	$\dots$
$W_2 :$	1	0	0	0	1	0	0	0	1	$\dots$

W_1 is a rearrangement of W_2 – a rearrangement of infinitely many local values, over both persons and positions. We could take all of the 0-values in W_1 and bring them forward to occupy x_2, x_3, x_4 and then x_6, x_7, x_8 , etc. And we could take the 1-values and spread them out to $x_1, x_5, x_9, \dots$ and so on. If we were indifferent to even *infinite* rearrangements of value – if our evaluation was entirely independent of *who* and *where* the beneficiaries are—then we’d be indifferent between these two worlds.

But this is implausible. W_1 is clearly better, and Pareto (over either persons or positions) says so. So we cannot be wholly indifferent to infinite rearrangements of both persons and positions.

This doesn’t conflict with the core claim of impartiality: that it is no better to for one set of beneficiaries to obtain some values than another. It only holds when there is the same *number* of beneficiaries in either case. But there are infinitely many beneficiaries in both W_1 and W_2 . In response to this, we might say that the sizes of the sets of beneficiaries are simply undefined, so impartiality tells us nothing.

But we *can* define the sizes of infinite sets. How? We might do so by using their *cardinality*. Two sets have the same cardinality if and only if a bijection exists between them – for locations, if one set of locations can be rearranged (perhaps in infinitely many places) to precisely replace another set, just as we can do above with the 1-valued locations in W_1 and the 1-valued locations in W_2 . That case demonstrates that, in this *ethical* context, cardinality is not a plausible way to count beneficiaries. Impartiality would then force us to reject all forms of Pareto.

To satisfy Pareto while remaining impartial, we can turn to alternate notions of size for infinite sets. One is *containment*: a set A is larger than another, B , if $A \supset B$. That is, if A is a proper superset of B , and so *contains* it. And in the example above, the set of locations with value 1 in W_1 does contain the set with value 1 in W_2 . By containment, the set of beneficiaries is not the same size in both worlds—it’s strictly larger in W_1 . And so

impartiality will not imply, implausibly, that they are equally good. Nor must we violate Pareto.¹⁶

And there is another Pareto-compatible version of size as well—*density*. Compare, for example, the set of natural numbers to the set of (positive) even numbers. We might say that the natural numbers form the larger set because they contain the evens. But as we count along the naturals, from 1 up, it is also true that we encounter natural numbers more frequently than evens—in other words, the naturals have greater *density* over that ordered sequence. Likewise, in the example above, the locations with value 1 occur more densely over the sequence $(x_1, x_2, x_3, \dots)$ in W_1 than they do in W_2 —in this sense, we can say there are *more* of them.

We can use density to measure subsets of locations more generally, when our locations have some essential properties which generate the right sort of structure. Whenever those properties are naturally represented in a coordinate structure (of any number of dimensions), we can say that one subset of locations occurs more densely over the space of possible values than another.¹⁷ For instance, the participants/positions in *Firing Line* above each had the property of where they were positioned in the line, and this was represented by a real number on a single coordinate. And we can say that the set of every second person occurred much more densely over that structure than the set of every fifth person. In general, if we treat spacetime positions as the relevant type of locations, they have an essential property which always generates such a structure: their coordinates in spacetime. So it can make sense to say that one infinite set of spacetime positions is larger than another if the first set occurs more densely over spacetime. Alternatively, we might say that one set of *persons* is larger than another because it occurs more densely over spacetime, but this is a little strange—the size of the set could change if we moved those persons around, since a person's position isn't essential to their identity.

¹⁶We might instead use a stronger notion of containment: a set A is larger than B if $A \setminus B$ has greater cardinality than $B \setminus A$.

¹⁷This is an application of *asymptotic density* to this particular context (see Halberstam and Roth, 2012).

The key takeaway from all of this is that we have multiple candidate methods for measuring the size of a set of beneficiaries, not just cardinality.¹⁸ And the others don't imply that we must be indifferent to rearrangements of infinitely many local values, since sets of locations need not be equally large just because we can biject them. To claim that impartiality *does* require indifference to such infinite rearrangements is to assume that cardinality is the relevant notion of size. And this requires independent justification, lest we beg the question.

So it is not at all obvious that views which are sensitive to the position of value in spacetime are immediately impartial, or that they constitute a fundamental departure from aggregation. I see no compelling reason to disregard spatiotemporal views like mine, especially not when they give far more plausible verdicts (such as in *Firing Line* above).

2.4 Expansionism

If we give up Pareto over persons, and we don't ignore the spatiotemporal position of value, then we can compare many more pairs of worlds. Here is one way we might do so, by way of example.

Take worlds W_1 and W_2 (which resemble *Firing line*). Both worlds contain a sequence of persons spread over time, each holding the same position in both worlds. In W_1 , some harm befalls every second person. In W_2 , that harm befalls every fifth person. Those harmed bear value 0; those not harmed bear value 1.

We can depict these worlds on spacetime diagrams, as below. Each person is positioned in spacetime, so we can assign them some coordinates x in four dimensions. Here, the vertical axis represents time and the horizontal axis represents just the one spatial dimension, for simplicity.

¹⁸Another candidate is *numerosity*, as described by Benci and Nasso (2003), which is an extension of containment.

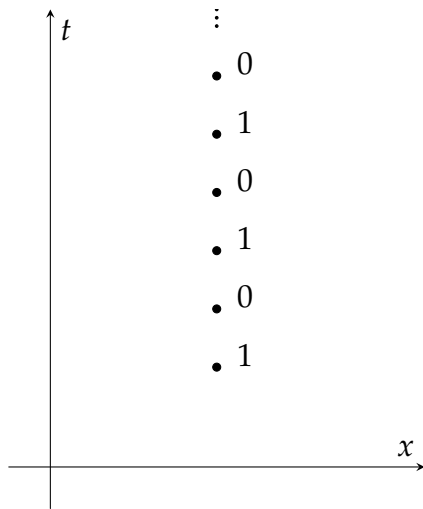


Figure 2.1: W_1 in a temporal version of *Firing Line*

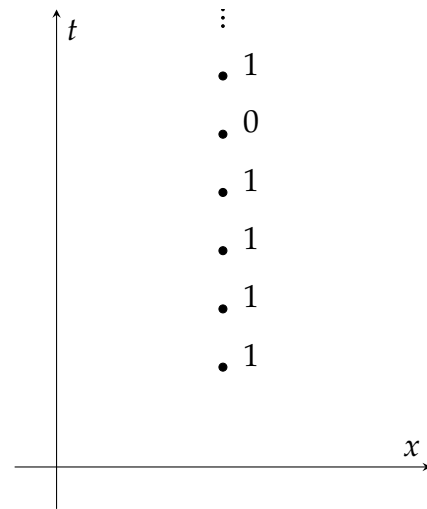


Figure 2.2: W_2 in a temporal version of *Firing Line*

Which world is better? To determine this, let's gradually sum up the value in each world, but in a particular order. Start at some point $\mathbf{p}$, as illustrated below. Sum the local values in the order of how far they are from $\mathbf{p}$. For all locations within a given distance from $\mathbf{p}$, those will lie within a circular region centred on $\mathbf{p}$ (or, in four dimensions, a hypersphere). To sum values in order of their distance from $\mathbf{p}$ is to sum them over some sequence of expanding circles of increasing radius.

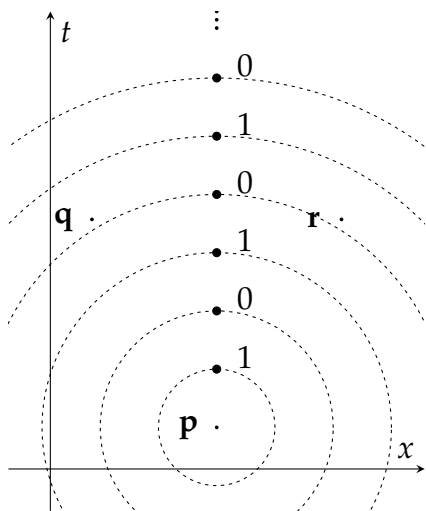


Figure 2.3: Expansions over W_1 ,
centred at $\mathbf{p}$

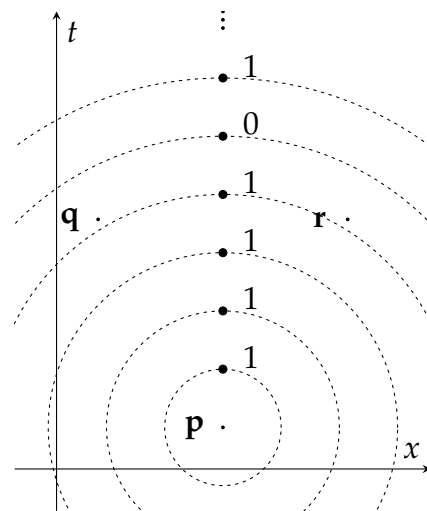


Figure 2.4: Expansions over W_2 ,
centred at $\mathbf{p}$

As we sum value over this sequence, we obtain a sequence of cumulative sums. These are the sums of local values within each circular region (the stage in the *expansion*). We can represent these on a diagram too, as a function of the radius r of each circle.

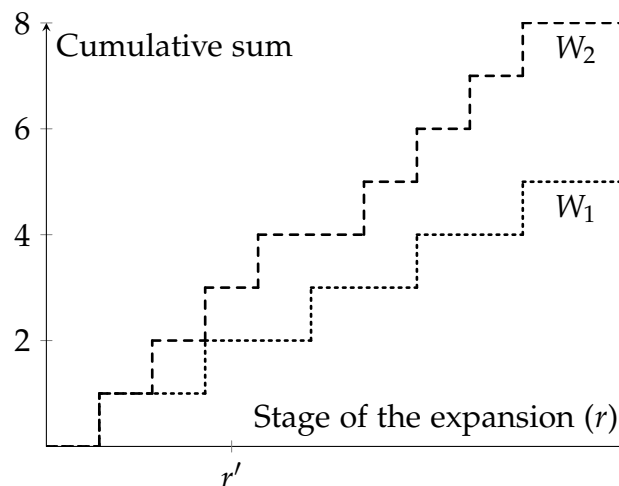


Figure 5: Cumulative sums in W_1 and W_2 , with expansions starting from $\mathbf{p}$

Notice that the cumulative sum for W_2 overtakes that of W_1 . There's some stage in

the expansion (r') at which the cumulative sum of W_2 is higher than that of W_1 . And, beyond that stage, it remains higher.

Here, this happens no matter where we start the expansion. We could put the center of those circles in line with the values themselves (e.g., p), or somewhere off to the left or right (at q or r in Figures 3 and 4). We could start at any real coordinates, and our expansion would still reach a stage at which W_2 has greater cumulative sum and, beyond which, W_1 won't ever catch up.

That's how we determine that W_2 is the better world: every sequence of expanding circles we could start from any point agrees that W_2 has more value. That's more value according to its cumulative sum, no matter how far out we take that sum.

This is expansionism in a nutshell. More precisely, it's a *spatiotemporal* version of expansionism—it uses the natural arrangement that locations (whether persons or positions) have in spacetime.

In general, expansionism consists in a principle like the following, from Vallentyne and Kagan (1997, p. 17).¹⁹

SBI3 (Strengthened Basic Idea 3): W_1 is better than W_2 if

- (1) W_1 and W_2 have exactly the same locations, and
- (2) for all bounded regions of locations there is a bounded uniform expansion, such that, relative to all further bounded uniform expansions, W_1 is better than W_2 .

To clarify, a *bounded region* is best interpreted as a set of finitely many locations with non-zero value (*ibid.*: 13). An *expansion* is an infinite sequence of bounded regions, each a superset of the previous region. A *bounded uniform expansion* is an expansion, each

¹⁹Arntzenius (2014) accepts a similar expansionist principle, designed to deal with cases of uncertainty. In cases of certainty, his principle is equivalent to SBI3.

region in which is obtained from the previous region by "... [adding] a band of constant width..." in all directions to the boundary (as in the example above) (*ibid.*: 16).

The principle I defend, and which was demonstrated in the example above, will deviate from Vallentyne and Kagan's SBI3 in several important ways. Here is how, and why, it differs. (Readers uninterested in the motivation for these changes may skip ahead to the definition of SE1.)

- (1) It uses the natural arrangement of locations in spacetime (whatever those locations are) to define regions and bounded uniform expansions, and so deals only with locations which *have* spatiotemporal positions.

Vallentyne and Kagan's principle does not necessarily use spatiotemporal positions to make judgements. It can be applied with other arrangements (and corresponding notions of boundedness and uniform expansions). For instance, using persons as locations, we could use the arrangement of those persons given by when we rank them by height. Or we might use their position in an alphabetical list of their names. In either case, we could start at any height (e.g., 170 centimetres) or any name (e.g., Xavier), and we could specify expanding 'regions' of persons over intervals of height or name (e.g., the interval [169 cm, 171 cm] or [Xander, Xena]). But we face problems in practice if we use such arrangements to make moral judgements. According to the physical theories cited above, our universe contains infinitely many tokens of every local phenomenon. This includes infinitely many persons who go by Xavier, Xander, or Xena, as well as infinitely many people with heights falling within any interval $[a - d, a + d]$. A non-zero proportion of them will have valuable lives, so these expansions would give an infinite (or undefined) cumulative sum at every stage. But not so if we use the spatiotemporal arrangement—for most plausible forms of moral value (such as pleasure), it is only physically possible to fit a finite amount of value into finite spacetime. This is one reason to expand over the spatiotemporal arrangement. Another is simply that it seems far more plausible

that moral evaluation depends on spatiotemporal position than that it depends on the heights, names, or other properties of persons or locations.

(2) The principle applies only if local value is additively separable.

This is an assumption which I expect Vallentyne and Kagan would agree with, but I want to make it explicit. It is also well-justified. For one, it is implied by all minimally aggregative views in the finite setting, and these are the views that we want to extend to the infinite setting. It is also very useful in practice—it allows us to simplify cases enormously, as we’ll see in the following sections. Since local value is additively separable, we can add or subtract the same value to the same positions in each of two worlds, even infinitely many positions, and we wouldn’t affect the ranking of those worlds. This means that, when comparing worlds, we can cancel out the local values that they have in common. For instance, if $W_1 \succ W_2$, then $W_1 + W \succ W_2 + W$ for all $W \in \mathcal{W}$, and vice versa (where the world $W_a = W_b + W_c$ is the world defined by $V_a(l) = V_b(l) + V_c(l)$ for all $l \in \mathcal{L}$ and $\mathcal{L}_a = \mathcal{L}_b = \mathcal{L}_c = \mathcal{L}$).

(3) Expansions begin with points rather than regions.

This is equivalent to restricting initial regions, and expansions, to circles on the two-dimensional diagrams above—or, in four dimensions, hyperspheres—rather than any nasty old shape one can imagine. This is a costless strengthening of the principle—it allows us to judge in more cases without, as far as I can tell, giving incorrect judgements in any. (Note that the problems described in the following section emerge with or without (4), but their solution requires it.) It is also in the spirit of Vallentyne and Kagan’s approach: their regions were already uniform in the way one expands to the next; now each region is uniform in its boundary as well.

In effect, this means that each region in the expansion is the set of all points within some fixed distance r of a central point $\mathbf{p}$: a ball of radius r centred on $\mathbf{p}$.

- (4) The distance between locations is defined by geometric distance.

For expansions to be uniform, we need to define distance, which brings us to (5). The natural way to do this in the spatiotemporal setting is geometric distance d , given by $d^2 = \Delta x^2 + \Delta t^2$ for any two endpoints separated by spatial distance Δx and temporal distance Δt . On the two-dimensional diagrams used above, this corresponds to the distance given by placing a ruler on the page. And it results in each uniform expansion of a circle (or hypersphere) also being a circle (hypersphere).²⁰

- (5) The principle provides a sufficient condition for equality ($\simeq$) between worlds, rather than just strict betterness.

Vallentyne and Kagan's principle only stated a sufficient condition for strict betterness $\succ$. I want to add one for $\simeq$, along the same lines.

With those modifications in hand, here is what I take to be an initially plausible principle of spatiotemporal expansionism (SE).

SE1: For worlds W_1 and W_2 with the same set of locations²¹, $W_1 \succ W_2$ if, for

²⁰Geometric distance may seem the most natural way to do this, but it is not the only way. Since time and spatial distance measure fundamentally different quantities, and use different units, there are many different ways we might choose to combine the two into a common metric. But to allow us to construct useful expansions, such a metric does need to satisfy three conditions. i) For every pair of points, the distance between them must be defined (or else expansions from one point will never reach the other point). ii) If we start with a bounded region, any uniform expansion of that region must also be bounded (or else our cumulative sums may be infinite). And (iii) Distance must increase for increasing changes in space and increasing changes in time, so that the expansions become larger in the natural sense. Together, these conditions imply that our distance metric must satisfy $d^p = |\frac{\Delta x}{a}|^p + |\frac{\Delta t}{b}|^p$ for some positive, real a, b and p . (For instance, Arntzenius (2014, p. 44) describes an approach which effectively uses $a = b = 1$.) If we apply an additional condition—iv) that regions of fixed distance are invariant under rotation—then we obtain geometric distance.

Note that, in a non-Euclidean space, there may be no definition of distance which satisfies (i)-(iii), let alone (i)-(iv), so we would need to abandon one of them. So in Chapter 4 I will move away from this definition of distance and, in doing so, sacrifice requirement (iii). But even with the distance metric I describe there, we obtain the same verdicts in the problem cases below. Likewise, if we adopted any metric which satisfies $d^p = |\frac{\Delta x}{a}|^p + |\frac{\Delta t}{b}|^p$, we would obtain the same verdicts, so we can simplify the discussion by using geometric distance.

²¹In practice, these sets will always be the same if we use positions as locations. If we use persons merely positioned in spacetime, they may not be.

all starting points $\mathbf{p}$, there exists $r' \in \mathbb{R}$ such that for all $r > r'$,

$$\sum_{\{\mathbf{x} | d(\mathbf{x}, \mathbf{p}) \leq r\}} (V_1(\mathbf{x}) - V_2(\mathbf{x})) > 0$$

And $W_1 \simeq W_2$ if, for all $\mathbf{p}$ and all $r > r'$ the sum equals 0.

SE1 matches the method we used at the start of this section. When comparing any two worlds, start at any point you like ($\mathbf{p}$). Generate larger and larger circles (hyperspheres)—circles with increasing r —centred at that point. Those are the regions $\{\mathbf{x} | d(\mathbf{x}, \mathbf{p}) \leq r\}$. Take the cumulative sum of value within each of those circles (hyperspheres). Is there a stage at which one world takes the lead (where the difference in cumulative sums is greater than 0)? Does it remain in the lead for all greater r ? And does this happen no matter where you placed $\mathbf{p}$? If so, that world is better. Or is there a stage beyond which the cumulative sums remain equal? If so, the worlds are equally good.²²²³

So we have a way of comparing infinite worlds which isn't clearly implausible. In the remainder of this chapter, I'll test SE1 on examples that demonstrate that it needs to be stronger. That stronger version will come in §2.6.

²²With SE1, we can extend minimally aggregative moral theories to make comparisons of infinite worlds. But does that extension itself count as minimally aggregative? Recall that a minimally aggregative theory must say that a world is at least as good as another if and only if it contains at least as great a total aggregate of value (impartially construed). If we were to interpret 'total aggregate' as being a real-valued, one-dimensional sum of value, then no, a theory satisfying SE1 would not be minimally aggregative. But a total aggregate can be any combination of the local values in a world, so long as every such value is incorporated into the aggregate and is incorporated in an impartial manner. (I take this impartiality as requiring only that any such value could be switched with any other and the aggregate would not be increased or decreased.) And SE1 does indeed determine betterness on the basis of such a total aggregate: the cumulative sum of value over a given expansion. This cumulative sum is a more complicated aggregate than the one-dimensional totals we can use in the finite context, but it forms an aggregate nonetheless. (The same holds for the SE2 principle introduced below.)

²³Note that SE1, and other expansionist approaches, can be expressed in terms of linearly ordered abelian groups and filters. Pivato (2014) shows this for all $\succsim$ relations which satisfy basic conditions resembling Finite Anonymity and Pareto. I won't attempt to express SE1 in such terms here, but note that it is not necessarily inconsistent with approaches which use such representations (e.g. Bostrom, 2011, p. 20-4).

2.5 Problem cases

In the last section, I gave a seemingly plausible principle of spatiotemporal expansionism: SE1. Now, I'll test it on three fairly mundane cases, more complex versions of which would not be uncommon in our own universe. As it turns out, SE1 fails to give judgements in these cases; it falls silent.

Why is silence a problem? After all, SE1 is guaranteed to give us a $\succsim$ relation which is incomplete. Otherwise it would violate Pareto or Finite Anonymity. But incompleteness comes in degrees—an incomplete relation may still be minimally complete, and deliver verdicts in cases which agents face in real life. And some agents will likely face cases resembling the three I describe below (absent some of my simplifications). So a principle which falls silent in these cases is unsatisfactory. This is why SE1 isn't quite enough, and why I'll strengthen it further in the next section.

2.5.1 Spatial shifts

Consider *Christmas*.

Example: Christmas

Each year the Consistent family celebrates Christmas. The Consistents always celebrate at their family home (at the exact same spatial location) and continue to celebrate at that home over successive generations. Celebrating Christmas brings joy to the Consistents, producing a brief, localised quantity of value. And the Consistents are lucky enough to have their family line persist forever, so there will be an infinite sequence of Christmases celebrated at their home.

One generation of Consistents have the option of moving to a new address. If they move, every future Christmas will be celebrated at the new address. But they're a consistent family—they'll still celebrate at the same time, and

produce the same amount of value. Everything else in the world will remain unchanged.

Staying at their current address can be represented by the zero world, W_0 , with local value 0 at all locations. The option of moving would then be represented by W_1 , pictured below. W_1 contains a sequence of local values -1 where they moved from, and values +1 where they moved to. And, since local values are additively separable, we ignore all other value in the world and just record the differences.

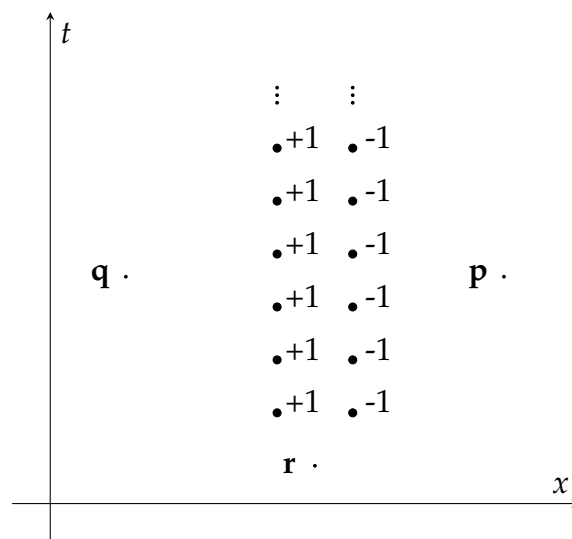


Figure 2.5: W_1 in *Christmas*

Let's apply SE1. Starting our expansion at point p , we get the following cumulative sum for W_1 . W_0 has a cumulative sum of 0 at all stages of the expansion.

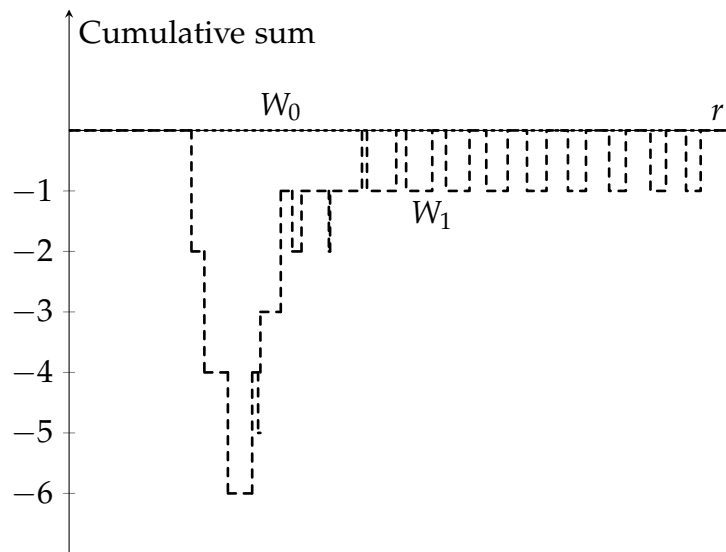


Figure 2.6: Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$

Starting at $\mathbf{p}$, our expansion always reaches the next -1 before its partner +1.²⁴ The expansion never reaches a stage at which the cumulative sum of W_1 remains greater than, less than, or equal to that of W_0 ; it will always alternate among these. And likewise if we start at point $\mathbf{q}$ – the cumulative sum of W_1 will still alternate, but between greater than and equal to.

So SE1 is silent. If our betterness relation is fully described by SE1, these worlds remain incomparable. And that's a problem, because this case seems fairly mundane. We're just shifting a sequence of local value in space by a fixed amount. All the more mundane is that we didn't specify the distance we shifted it – it could be just a millimetre, and SE1 would still be silent. It seems plausible that agents really could make such an arbitrarily subtle change to the future, at least in addition to countless other changes, which would make the case all the more complicated. So I find it implausible that we cannot compare such worlds. Further, I find it implausible that these two worlds aren't

²⁴This holds *even though* the delay between reaching the -1 point and the partner +1 point approaches 0. To see this, simply compare the distance from $\mathbf{p}$ to a -1 point to the distance from $\mathbf{p}$ to its partner +1 point – the latter is always strictly greater.

equally good, as intuition suggests. Since SE1 cannot compare the two, nor justify our indifference between them, we have a problem.²⁵ But worry not—we'll see below that it can be solved.

2.5.2 Temporal shifts

Consider *Public Holiday*.

Example: Public Holiday

A prophecy reveals that the kingdom of Alethkar will persist forever. To celebrate this, the queen will establish a new public holiday. This holiday will be accompanied by grand festivities and will be held on the same date, forevermore. Once it's established, its date will never be changed.

There are two dates to choose from, half a year apart. No matter when the holiday is held, it produces a fixed amount of value in every year except the first—in the first year, the delay will result in a slightly more enjoyable festival. The choice of date has no other effects.

The world in which the queen chooses the earlier date can be represented by the zero world W_0 . The world in which the queen chooses the later date is then W_1 (pictured below). W_1 differs from W_0 in two ways: there is local value -1 at every point the holiday

²⁵Could SE1 be modified to avoid this verdict, while leaving intact its core features? Yes, but such modifications bring on other nasty implications.

One option is to abandon the requirement that expansions be *uniform*. For instance, we might replace them with sequences of square-shaped (hypercube-shaped) regions, as Askill (2019, p. 208-11) does. This alternative may avoid the problem in *Christmas*, but it still faces the problems below. I should also note that this solution also appears somewhat arbitrary – it remains sensitive to *some* changes in Δx and Δt but not others.

Here is another option, indeed the only option if we hold fixed the uniform expansions and the dependence of $\succsim$ on the cumulative sums over those expansions. When making judgements, do not require that *all* starting points agree on judgements. Instead of *supervaluating* over starting points, we could *subvalue*. After all, in the *Christmas* case, there is some starting point ($\mathbf{r}$) for which the cumulative sums of W_0 and W_1 are equal for all expansions (and there are *none* which give a conflicting verdict). Why not then judge them as equally good? I address this approach in an appendix at the end of this chapter, and show that it results in an intransitive $\succsim$ relation.

would otherwise have happened (once per year); and there is local value 1 at every point the holiday happens instead, half a year later, with the exception of the first holiday in W_1 which is made slightly better by the delay. All of this happens at the same spatial position where the festivities traditionally take place.

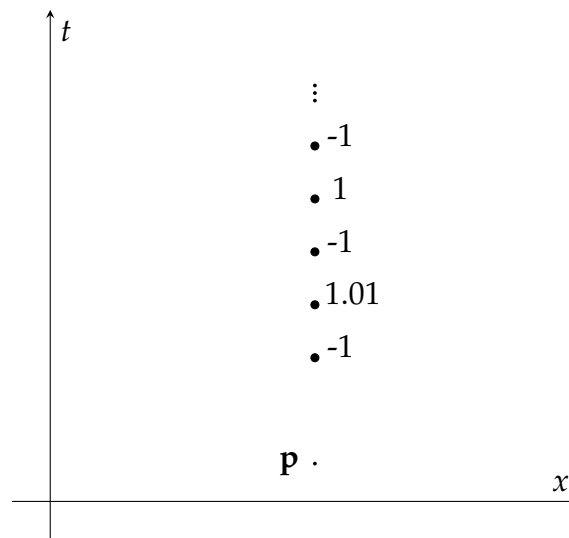


Figure 2.7: W_1 in *Public Holiday*

If we start the expansion at $\mathbf{p}$, these are our cumulative sums:

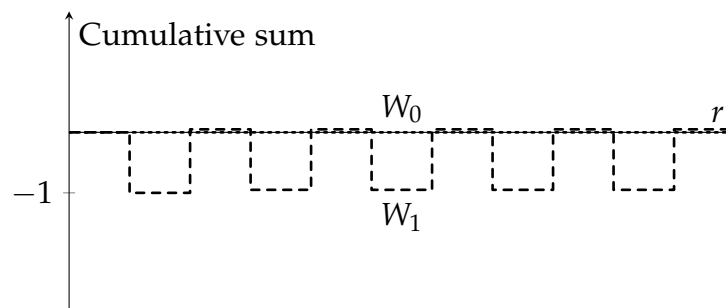


Figure 2.8: Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$

This pattern repeats indefinitely. So, no matter how far we expand, there's always a later stage at which W_0 has the greater cumulative sum, and one at which W_1 does. And this happens no matter where we start our expansion. So SE1 is silent in this case (as is

SE1* from above) – these two worlds are incomparable.

This is odd. We only made a fairly mundane change to the world, one that agents might realistically achieve. But we still get incomparability? Again, we have a problem. (And again, as we'll see, it can be solved.)

2.5.3 Chaotic effects

In practice, cumulative sums will almost always alternate, as in *Christmas* and *Public Holiday*. To see why, consider *Writing or Netflix*.

Example: Writing or Netflix

Rita has a choice to make. She can either spend the next hour writing the last of an important paper on ethics, or spend that hour watching a television show. If she finishes the paper in the next hour and submits it for publication, it will be read by a policymaker. The policymaker will increase funding to some international aid intervention and thereby save k people from painful deaths. If Rita doesn't finish her paper within the hour, those k people will die. She knows all of this.

Unlike above, her actions have lasting, chaotic effects on the world. Saving those k people will, e.g., produce random effects stretching indefinitely into the future (described in detail in Lenman, 2000; Greaves, 2016, and Chapter 3 below). If Rita takes one action rather than the other, this will cause random changes to many future events, continuing infinitely far into the future.

We can represent the actual outcome of Rita watching television by the zero-world by W_0 and the actual outcome of her finishing the paper by W_1 , as below. This world W_1 contains all of the differences between the two outcomes.²⁶ These differences consist in:

²⁶As above, this simplification is allowed by the assumption that value is additively separable.

the initial benefit of saving k lives, proportional to k ; and an infinite sequence of random differences $X_1, X_2, \dots$ (more on these shortly). For simplicity, I'm modelling the effects of Rita's actions as changing the value at one location per unit time, rather than at all times. I'm also pretending that these changes all occur at the same spatial position, rather than spread across her future as is more realistic. These simplifications won't change the verdicts below.

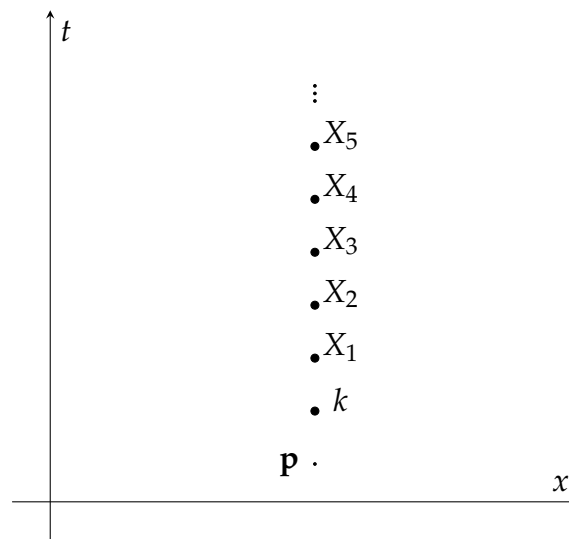


Figure 2.9: W_1 in *Writing or Netflix*

Since we are assuming that Rita's actions make chaotic, seemingly random changes to the world, the differences X_i are given by *random variables* which are independent, identically distributed, and symmetric about 0.

Those variables' probability distributions may reflect any of: Rita's subjective uncertainty about the exact effects of her actions; the evidential probabilities of the different outcomes; or their objective chances, whichever is morally relevant. Either would work here. But note that, despite the mention of probabilities, we are not suddenly comparing *lotteries* over outcomes. This is still a comparison of *ex post* outcomes—of what would *actually* end up happening if Rita took a given action. But let's pick out outcomes typical

of this situation, given the uncertainty and chanciness present. We can do this by generating local values using the random variables X_i , and comparing whatever they give us. Due to the way they're generated, these X_i will be random but still determinate. And, as we'll see below, using these random variables allows us to draw much more general conclusions.

To compare the outcomes, let's run our expansions, starting from $\mathbf{p}$ (or any other point, which will give the same result). Here is our cumulative sum for a typical, randomly-generated W_1 .

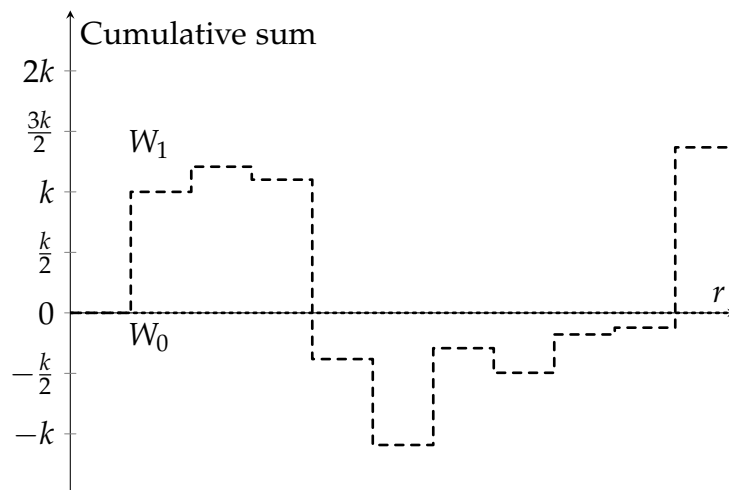


Figure 2.10: Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$

Within this diagram, the cumulative sum of W_1 returns to 0 twice. But it will keep doing so: no matter how large r gets, it is guaranteed to return to 0 eventually. Why? The cumulative sum $(\sum_{\{\mathbf{x} | d(\mathbf{x}, \mathbf{p}) \leq r\}} (V_1(\mathbf{x}) - V_0(\mathbf{x})))$ forms a symmetric random walk over the reals, with starting point k and each step in the walk given by X_i . This random walk has an unfortunate property: it's *recurrent* (Chung and Fuchs, 1951). For any r , no matter the value of the cumulative sum of W_1 at that stage, it will return to 0 again eventually. (It has probability 1 of doing so.) And this holds no matter how big k is.

What does this mean for *Writing or Netflix*? It means that SE1 won't be satisfied for

this W_1 and W_0 —it will remain silent. Nor will SE1 be satisfied in any situation like Rita’s with values generated the same way, with probability 1. And that means that, whenever our actions cause lasting and independent chaotic changes to the world, the outcomes of our actions are practically *guaranteed* to be incomparable by SE1.²⁷

This is a problem. In real life, the differences between actual effects of our actions will often be widespread and, to epistemically limited agents like us, will seem random. As I argue in the next chapter, in almost any realistic decision scenarios, these random differences will continue endlessly into the future. And so most moral decisions faced by agents like us will resemble *Writing or Netflix*—the worlds that the agent might produce differ randomly and independently at infinitely many locations. And SE1 won’t be able to compare them.

2.6 Solution

Fortunately, all three problems are solvable. We can replace SE1 with a stronger, modified principle that gives all of SE1’s judgements and more.

But first, to help make sense of that principle when I present it, consider a typical case in which SE1 says that $W_1 \succ W_2$ (illustrated below). For SE1 to say that, the graph of W_1 ’s cumulative sum must end up above that of W_2 at some point, and then stay above it for all greater values of r . The graph for W_2 might be above that of W_1 for some values of r (shaded with vertical lines), but only for finitely long—if we added together the lengths of every interval for which W_2 is in the lead, we’d get a finite length. Meanwhile, since W_1 will stay in the lead for all r above some finite bound, the total length that W_1 spends in the lead will be infinite. Note that the total *area* between the curves when W_1 is in the lead will be infinite too.

²⁷I am indebted to Christian Tarsney for raising this problem.

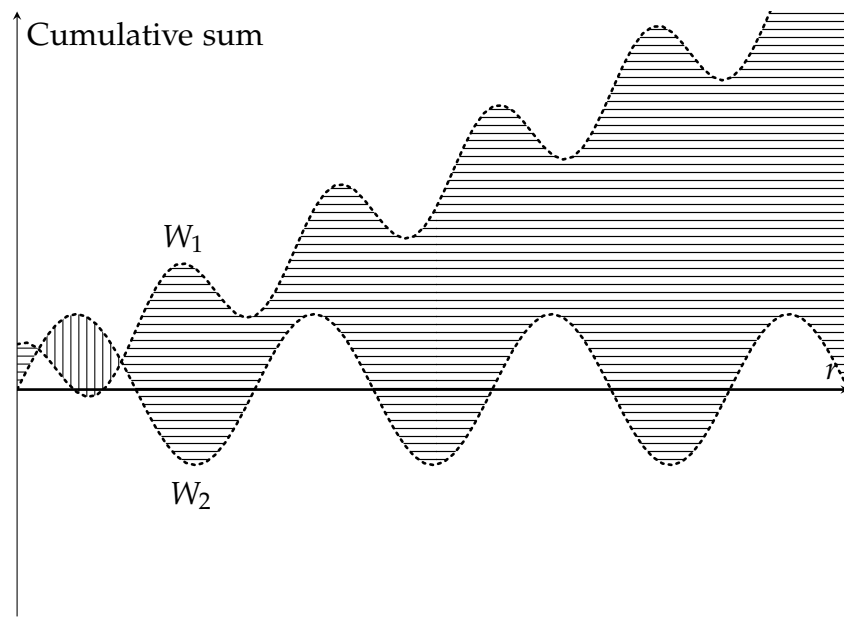


Figure 2.11: Typical cumulative sums for $W_1 \succ W_2$

Consider also some W_1 and W_2 which SE1 judges as equally good, as below. Beyond some finite value of r , their cumulative sums must be equal. So there can be at most a finite total length for the intervals during which W_1 is in the lead (shaded with horizontal lines) and likewise for the intervals during which W_2 is in the lead (shaded with vertical lines). Note also that the total area between the curves for the intervals during which W_1 is in the lead is finite, as is the total area for when W_2 is in the lead.

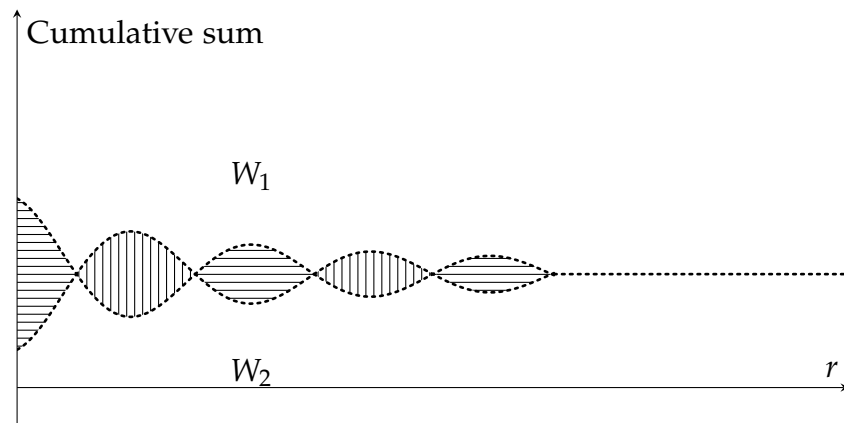


Figure 2.12: Representative cumulative sums for $W_1 \simeq W_2$

We could generalise this to worlds which SE1 fails to compare. We could propose: take all the intervals of r values for which W_1 has the greater cumulative sum, and the intervals for which W_2 does; sum up the lengths of those intervals for each; if both total lengths are merely finite, then W_1 and W_2 are equally good; and, if the total length is infinite for W_1 and merely finite for W_2 , then W_1 is the better world.

This proposal would strengthen SE1 substantially. For instance, it gives us the correct judgement in *Christmas*. Below we have graphs of the cumulative sums for both worlds from that case. On those graphs, we frequently have intervals during which W_0 has greater cumulative sum (shaded with horizontal lines), but the lengths of the intervals become shorter and shorter as r increases. As it happens, the length of these periods shortens quickly enough that their total length converges to a finite sum. Meanwhile, there are no intervals during which W_0 has greater cumulative sum, so they have total length 0, which is finite too.

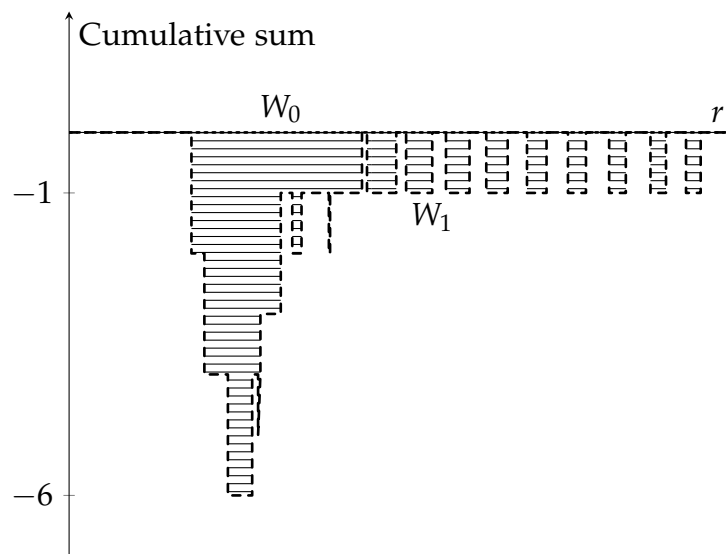


Figure 2.13: Cumulative sums for W_1 and W_0 in *Christmas*, with expansions starting from

p

And the same holds for all starting points in that case. For all of the intervals of r values during which either world has the greater cumulative sum, their total length is finite. So we might say that the two worlds in *Christmas* are equally good. And this is the correct answer, by intuition – the only difference between the two worlds is that we’ve shifted a sequence of values across in space slightly, which surely makes the world no better or worse.

But we’re still in trouble in *Public Holiday*. As illustrated below, we have regular intervals of r values during which W_0 is in the lead (vertical lines) and regular intervals during which W_1 takes the lead. Since the cumulative sums repeat in this pattern without end, both sets of intervals will have infinite total length. So we fall silent yet again.

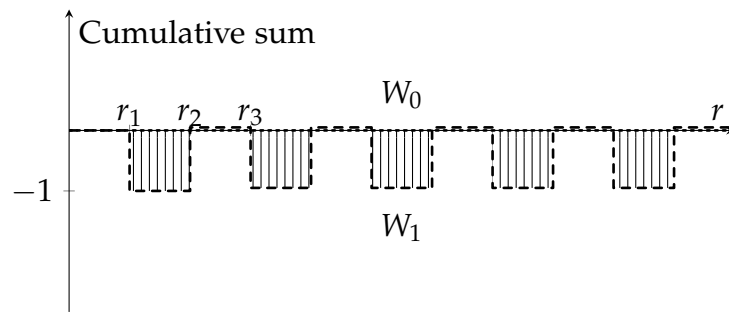


Figure 2.14: Cumulative sums for W_1 and W_0 in *Public Holiday*, with expansions starting from $\mathbf{p}$

This seems odd. We may just as frequently have intervals during which W_1 leads, but its lead is always tiny. And recall from above that the very first holiday in W_1 obtained *slightly* higher value than the others. If we reduced its value to 1, like all the others, then W_1 's cumulative sum would never take the lead. We could then say that W_0 is better, but we cannot because of this tiny addition.

But we can overcome this. It seems important not only how frequently each cumulative sum is in the lead, but also how *great* that lead is. The occasional lead of W_1 (such as between r_2 and r_3) is one hundredth as big as the lead that W_0 takes (between r_1 and r_2), so it seems plausible that it should count for one hundredth as much. And this correlates exactly with the area between the curves during those intervals—the area of each interval during which W_1 is one hundredth the area of each interval during which W_0 leads. If we were to take a cumulative sum of the *area* between W_0 and W_1 , subtracting the area for intervals when W_1 leads, that sum would rapidly approach positive infinity. And likewise, for the exemplar comparison in Figure 12, the area between the curves when W_1 leads minus the area when W_2 leads approaches positive infinity. Likewise for all cases in which SE1 says that one world is better than another. Meanwhile, in *Christmas* and the exemplar comparison in Figure 13, that sum of areas would be merely finite.

We might thereby say which world is better in *Public Holiday* and other cases, by

considering the cumulative sum of *area* between the curves. In *Public Holiday*, that sum rapidly approaches positive infinity (and it does so no matter where we put our starting point). So we could say that W_0 is better—the world in which the queen delays the public holiday is worse than the world in which she doesn’t. This verdict may be counterintuitive, but I think it is the right one. After all, if the holiday were delayed by an entire year, then the world would only differ from the undelayed world in two ways: it would have one fewer holiday at the start (valued at -1), and a slight increase of value to the first holiday (0.01); the rest of the holidays would be identical between worlds. That would clearly be a worse outcome; roughly, it just involves removing value from the world. And since delaying a full year would be worse, it makes sense that delaying by less than a year would also make the world worse.

We can formalise this approach with SE2. Here, the sum on the right is the very same as what appeared in SE1: the difference in cumulative sums between W_1 and W_2 . But we want the area, so we need to multiply each of those sums by the length $r_{i+1} - r_i$ that it spends at that level. The start and finish of those intervals are given by $\{r_1, r_2, r_3, \dots\}$, the distances between the starting point $\mathbf{p}$ and each point $\mathbf{x}$ at which there’s a difference in value between W_1 and W_2 . And summing the areas between the curves over each such interval $[r_i, r_{i+1}]$, from r_1 out towards $+\infty$, we can see whether the total area is bounded or infinite.²⁸

SE2: Let W_1 and W_2 be worlds with the same set of locations. For any starting point $\mathbf{p}$, let $\{r_1, r_2, r_3, \dots\}$ be the strictly increasing sequence of distances between $\mathbf{p}$ and each $\mathbf{x}$ such that $V_1(\mathbf{x}) - V_2(\mathbf{x}) \neq 0$.

Then $W_1 \succ W_2$ if the following sum diverges unconditionally to $+\infty$.

²⁸SE2 can be stated more generally to allow for value spread densely over a region rather than concentrated at discrete points. To do so, we need only replace the sum with an integral: $\int_0^\infty \sum_{\{\mathbf{x} | d(\mathbf{x}, \mathbf{p}) \leq r\}} (V_1(\mathbf{x}) - V_2(\mathbf{x})) dr$.

$$\lim_{r \rightarrow \infty} \sum_{i=1}^r (r_{i+1} - r_i) \left(\sum_{\{\mathbf{x} | d(\mathbf{x}, \mathbf{p}) \leq r\}} (V_1(\mathbf{x}) - V_2(\mathbf{x})) \right)$$

And $W_1 \simeq W_2$ if the sum is bounded both above and below.

Based on the same reasoning as above, SE2 gives the verdict in *Public Holiday* that it's better to schedule the holiday sooner rather than later. Likewise, it gives the same verdict in *Christmas*. And it implies SE1, so it implies all of the verdicts of it too, such as in *Firing line*. It also maintains the advantages of SE1 that it implies Finite Anonymity and Pareto over spacetime positions.

It also resolves our one outstanding problem case: *Writing or Netflix* (illustrated below).

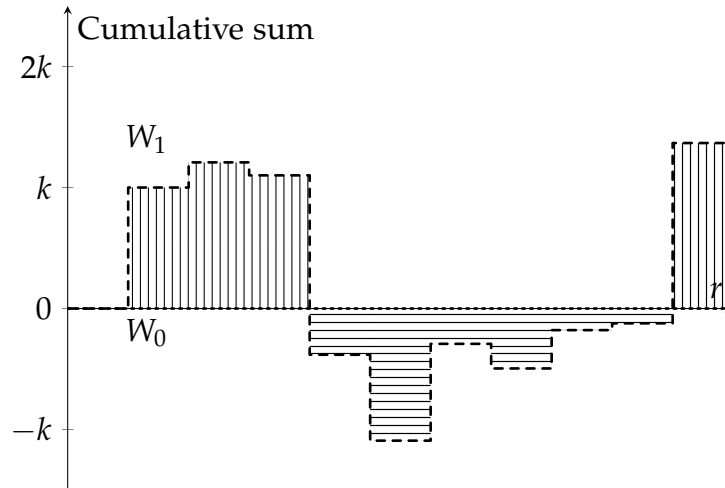


Figure 2.15: Cumulative sums for W_1 and W_0 in *Writing or Netflix*, with expansions starting from $\mathbf{p}$

The problem here was that the cumulative sum of W_1 formed a symmetric random walk over the reals, which was guaranteed to be recurrent and so guaranteed to pass through 0 over and over without end. But the area between the cumulative sums forms no such random walk. Rather, it's guaranteed to *not* be recurrent—it's guaranteed to diverge unconditionally to $\pm\infty$. To see this, take any finite upper and lower bounds: it

can be shown that the probability that the area under a symmetric random walk is within those bounds approaches 0 as $r \rightarrow \infty$ (Lipkin, 2003). So the probability that the integral approaches $\pm\infty$ as $r \rightarrow \infty$ is arbitrarily close to 1. And the same holds no matter where we start the expansion. So W_1 is *guaranteed* to be comparable to W_0 by SE2. In fact, it's guaranteed that one of the worlds is strictly better.²⁹

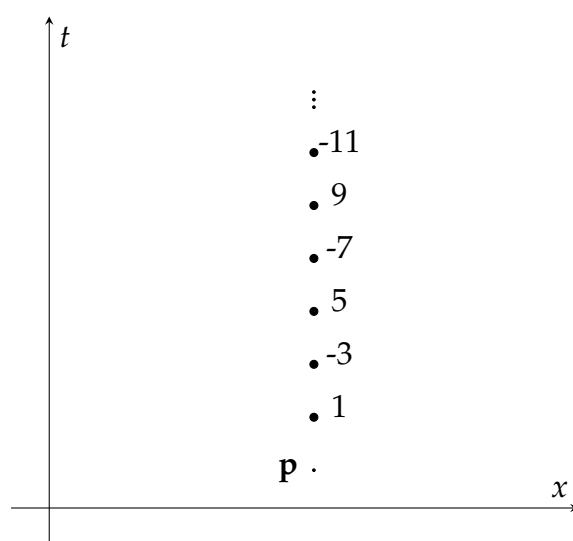
Thus, SE2 resolves all three cases. Spatiotemporal expansionism is back on firmer ground.

2.7 Remaining problem cases

I leave you with two more problem cases, lest you accept spatiotemporal expansionism too quickly.

The first involves local values alternating between positive and negative over time. But the cumulative sum also alternates, as does its integral. Compare the following W_1 to the zero world W_0 .

²⁹Admittedly, there's a probability of $\frac{1}{2}$ that W_0 is the better world, so Rita would still be very uncertain of which act would turn out better. But this is just the same conclusion we face in the finite case (Greaves, 2016).

Figure 2.16: W_1

Here are the cumulative sums for W_1 and W_0 . As you can see, the areas between the curves grow from segment to segment. Each of them exceeds the total area before it. And they grow so quickly that the magnitude of the area grows too. The area will not approach positive or negative infinity, nor will it remain bounded above or below. SE2 says nothing about this pair of worlds.

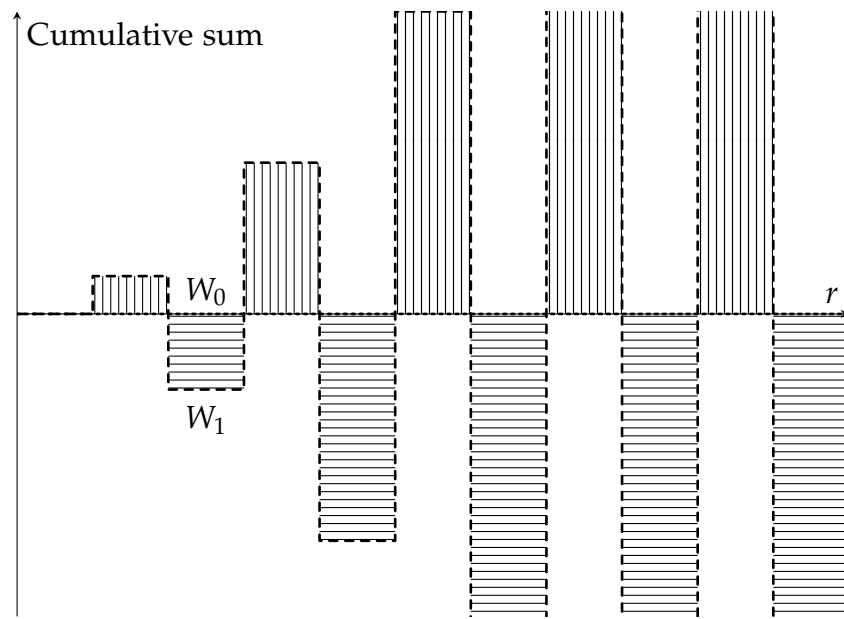


Figure 2.17: Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$

But this is a silence I can live with. This pattern of differences between worlds is carefully gerrymandered and, I would claim, we are unlikely to face such a pattern in real life. To achieve a world like this, we would need the universe to be neatly arranged into regions of overwhelmingly positive value alternating with regions of overwhelmingly negative value, of increasing length or intensity, in some sequence which is *just* right. And from the previous section we know that, if we model local values as random, this is guaranteed to not happen.

I am happy to put aside cases this exotic. We knew that SE2 wouldn't cover all cases, thanks to impossibility results. I would be happy for it exclude only cases like this.

Here is a more puzzling case. Suppose you have a sequence of local values at a single spatial position. This might be the Consistent family's Christmas celebrations. And suppose you can either leave them where they are or set them moving at a constant speed, maybe by launching their household through space (represented by W_1 below, compared to the zero world W_0). The same amount of value is obtained at the same time

intervals, but the path of the sequence will be diagonal rather than vertical.

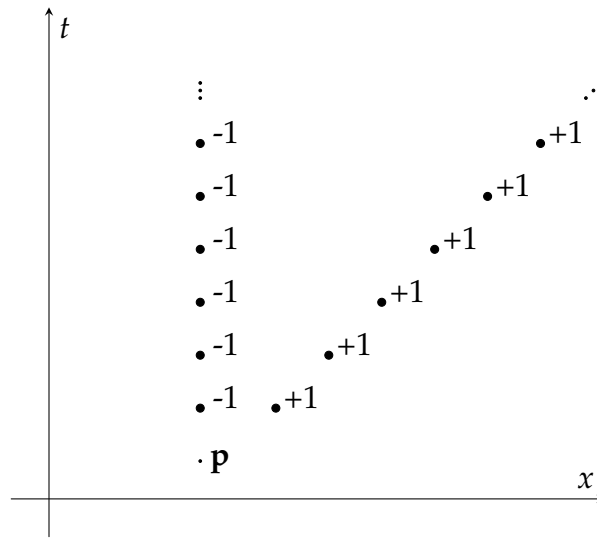


Figure 2.18: W_1

Starting our expansions at $\mathbf{p}$, we obtain the following cumulative sums (with similar results for all other starting points.)

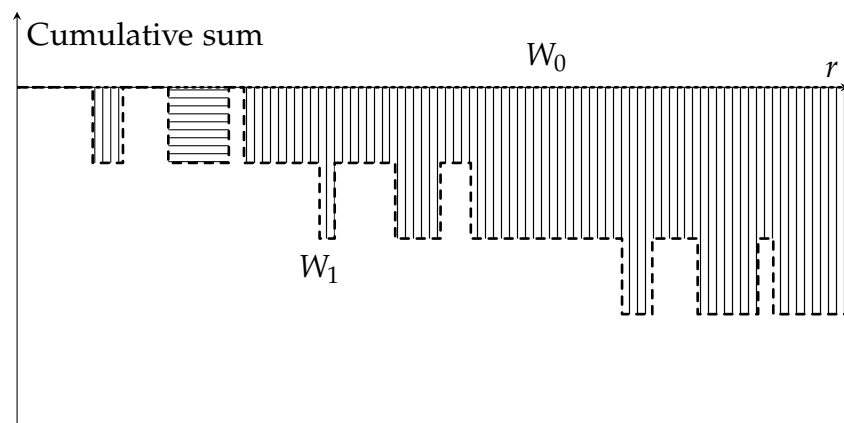


Figure 2.19: Cumulative sums for W_1 and W_0 , with expansions starting from $\mathbf{p}$

This is puzzling. Spatiotemporal expansionism (either SE1 or SE2) states that W_0 is better – the world is better if you leave the Consistents where they are rather than sending them into space at constant speed. But, intuitively, it doesn't seem like putting them in

motion makes any moral difference. Worse still, in general, SE1 and SE2 will weigh stationary sequences of value more heavily than moving sequences of value, and slower moving sequences more heavily than faster moving sequences. This is counterintuitive.

Worse again, it is a matter of perspective which sequence is moving and which is stationary. In this coordinate system, relative to which I've defined geometric distance, it appears that the Consistents are in motion. But for an observer standing in the Consistents' new house rocketing through space, watching their old house shrink into the distance, it will appear that the old house is the one in motion. They are no less justified in choosing a coordinate system (with corresponding distance measure) which says that their new address is stationary. We have no principled reason to adopt one coordinate system or the other.

Does that mean that the true betterness relation relies on an arbitrary choice of coordinate system?³⁰ Or that it does not hold at all in the real world, since no unique coordinate system exists?

I will return to this problem in Chapter 4. But, for now, I will just briefly note that SE2 can indeed be modified to avoid such relativistic implications. With a more carefully chosen coordinate system and distance measure than I gave here, we can reach the verdict that the worlds W_0 and W_1 above are equally good—we can judge that it is no better to leave the Consistents be than to send them rocketing off into space.

2.8 Summary

The basic challenge we faced was finding a way to extend minimally aggregative views to compare worlds containing infinite or undefined total value (and, ideally, to do so in a remotely plausible manner). We now have at least one way of doing so: with spatiotemporal expansionism. And we have seen that spatiotemporal expansionism can

³⁰Cain (1995) and Arntzenius (2014) make similar points.

deal with various basic cases, at least if we choose our expansionist principle carefully.

But you might think that spatiotemporal expansionism is an odd view to adopt. It gives ethical significance to the spatiotemporal position of value. It even implies that it may sometimes be better not to make Pareto improvements to persons' wellbeing. I hope to have convinced you that these objections are not decisive—the circumstances are dire enough, and the alternatives disturbing enough in their implications, that a spatiotemporal view can still be a live option. In the next chapter, we will see an argument for the claim that such spatiotemporal views are the *only* plausible options.

2.9 Appendix: Subvaluative expansionism

As we saw in §2.5.1, the principle SE1 fails to give a plausible verdict in the *Christmas* case. As pointed out in a note, we can give the desired verdict without strengthening our principle all the way to SE2, by modifying SE1 to *subvalue* over expansions from different starting points. I'll show here that such a modification gives an intransitive $\succ$ relation and so we should reject it.

Here is the subvaluative principle that we might adopt instead of SE1.

SE1*: For worlds W_1 and W_2 with the same set of locations, $W_1 \succ W_2$ if, for *some* starting point $\mathbf{p}$, there exists $r' \in \mathbb{R}$ such that for all $r > r'$,

$$\sum_{\{\mathbf{x} | d(\mathbf{x}, \mathbf{p}) \leq r\}} (V_1(\mathbf{x}) - V_2(\mathbf{x})) > 0$$

and there is no $\mathbf{p}$ or $r'' \in \mathbb{R}$ such that this sum is negative.

And, $W_1 \simeq W_2$ if the sum is 0 for all $r > r'$.³¹

³¹There is an alternate version of SE1*, in which we use the $\succ$ relation ($W_1 \succ W_2$ if there's *some* starting point such that W_1 's cumulative sum $\geq$ W_2 's cumulative sum beyond some stage of the expansion). This alternate version gives the correct judgement in *Christmas*. Unlike SE1*, it also gives the correct judgement (of equal goodness) in *Christmas 2* below. But it then fails if we increase any one of those instances of value by even a small amount. It then fails in a manner analogous to SE1* in *Christmas 2*.

Good news: SE1* does give us a verdict in *Christmas*, and the right one at that – that the worlds are equally good. (See Footnote 32.) But it doesn't in *Christmas 2*.

Example: Christmas 2

As in *Christmas*, the Consistents may either move to a new address or stay where they are. But, if they move, the Persistent family will be inspired to move to a new address as well. If they stay, the Persistents will stay too.

The Persistents are similar to the Consistents. They celebrate every year at their family home, successive generations maintain this tradition, and their family persists forever, but they happen to enjoy Christmas twice as much as the Consistents.

The world in which both families stay can be represented by the zero world W_0 . And the world in which they move by W_2 , pictured below. W_2 is much like W_1 from above, but features a second spatially-shifted sequence.

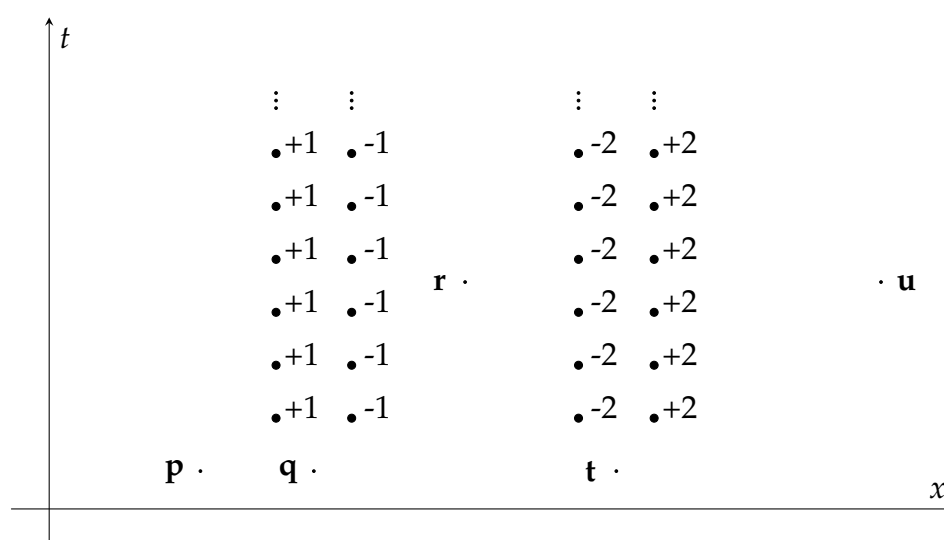


Figure 2.20: W_2

The cumulative sum of W_0 will always be 0, wherever we start our expansion. But in

The argument from chaos

3.1 Introduction

Consider a seemingly simple moral decision. While out walking, you see a child stepping onto the road to cross to the other side. Until recently, this was a one-way street. And it seems the child does not know that traffic now runs both ways, because they are looking in just one direction. Meanwhile, a long line of traffic bears down on them from the other side. If the child continues their crossing, it is clear that they will be squashed by the leading car. (Perhaps you can see that the driver is rummaging in their glove-box, oblivious to the child before them.) But, fortunately, you are not far from the child. If you dash forward, you can grab them in time to stop them stepping into the path of traffic. If you do, they will be saved. If you do not, they will die.

It seems obvious that the outcome in which the child survives will be better than that in which they die. And we might expect minimally aggregative views to confirm this—it seems that the former will contain more moral value in total, resulting from that child going on to enjoy the rest of their (presumably good) life.

But minimally aggregative views cannot say this, *even if* we suppose that the world will contain only finite total value. Why not? Because saving the child will have causal ramifications far beyond producing just one additional life that is worth living. In particular, it will have large-scale, lasting, and seemingly random effects, via the ways it influences physical systems that are *chaotic*. Small interventions on such systems typi-

cally result in wildly different outcomes in the future. One such system is the traffic flow in a sufficiently dense city (see Lighthill and Whitham, 1955). By allowing or preventing the child being hit by that car, you would allow or prevent a temporary standstill in traffic on that street. Over the following hours, this would alter traffic flow throughout the system, momentarily delaying or advancing countless other drivers. Depending on the size and density of your city, this would change which other traffic accidents occur and where. And beyond this, saving the child means that they will continue their life, with all the street-crossing and driving they will ever do, and so you will alter the traffic flow on countless other evenings too. Likewise for the street-crossing done by every one of the child's descendants. Effectively, by saving that child, you have caused and prevented many, many other traffic accidents and the resulting deaths, likely far outweighing the value of the child's life. And that's merely through disrupting traffic, rather than the disruptions you cause to many other chaotic systems (see §3.2 below). So you cannot say with much confidence at all that saving the child will result in an outcome of greater total value—you are *clueless* as to whether the long-run effects of your action will do more harm or good overall. So you will be clueless as to whether it is better to save the child or not (Greaves, 2016; MacAskill and Mogensen, n.d., see also). And this is deeply counterintuitive—our intuition suggests that it is clearly better to save the child.

Chaos poses one problem for comparing outcomes by their total values. And we know from above that we may face another if our universe is infinite: if the universe will contain infinitely many valuable lives whether you save the child or not, it may be that neither outcome has a finite total sum of value; and so, at first glance, it seems that minimally aggregative theories may be unable to compare them. But, fortunately, we already have plenty of proposals for comparing outcomes with infinite or undefined total values, such as that in the previous section.

In this chapter, I discuss a further problem, which has for the most part been over-

looked in the literature. If our universe is both infinite *and* widely chaotic, we find ourselves in a more difficult spot. Our solutions to the problem of chaos fail if outcomes also contain infinite value. And most of our solutions to the problem of infinite value fail too—they can only compare highly idealised pairs of outcomes, not those that differ by lasting chaotic effects, *ad infinitum*, such as every pair of outcomes we ever must compare in practice. In the decision of whether to save the child or not, most of our existing methods say that the outcomes are incomparable. Neither is better. And, if the value of outcomes is all that weighs in favour of either choice, we have no reason to choose one rather than the other. And the same holds in many of the moral decisions you or I will ever make. Effectively, chaos and infinite value together undermine a large swathe of our moral decision-making.

This problem is avoided by some, but not all, theories of infinite aggregation. Specifically, it is avoided by select spatiotemporal theories, including SE2 from the previous chapter. All extant *non*-spatiotemporal theories fail in chaotic settings. And so we have a compelling argument in favour of accepting a spatiotemporal theory.

Alternatively, you might interpret the argument given in this chapter as a *reductio ad absurdum* against minimally aggregative theories. You might think it absurd that whether one outcome is morally better than another depends on where value is positioned in space and time in each outcome. And so, if minimally aggregative theories require such dependence, you might conclude that such theories must all be false. It seems to me that this goes too far—abandoning even minimal aggregation in one’s moral theory seems too great a cost to pay to avoid dependence on spatiotemporal position. But you may disagree, depending on which conclusion you consider more absurd. I will leave it to you to decide which conclusion to draw.

The chapter proceeds as follows. In §3.2, I explain in greater depth how our actions almost always result in chaotic, long-lasting changes to the world—this goes far beyond

just affecting traffic flows. Then, in §3.3, I lay out the further problem of chaos *ad infinitum*, demonstrating how the problem arises for each of several classes of aggregative theories tailor-made for the infinite setting: *strongly anonymous theories* (§3.3.1); *weakly anonymous theories* (§3.3.2); and *position-dependent theories* (§3.3.4). I will demonstrate that, of the existing proposals, only position-dependent theories can overcome the problem. Indeed, accepting position-dependence is by far the most plausible way to overcome the problem. Section 5 concludes.

3.2 Chaos

Many of the physical systems in our universe are *chaotic*. Chaos has a technical definition (see Feldman, 2019, p. 83) but, for our purposes, we can focus on just one of the conditions necessary for a system to be chaotic: its states at all later times are highly sensitive to its initial states.¹ They are highly sensitive in that if we were to ever-so-slightly alter the properties of the system in the earlier state—perhaps in an *arbitrarily* small way—the later states would change in a much more significant way. For instance, consider the system of a pendulum, with the initial state given by where and when you release it to begin swinging. If the pendulum is a chaotic system, releasing it from a millimetre higher may result in it swinging in a wildly different, seemingly random manner.² But this need not be because the pendulum system is indeterministic—systems may be governed by deterministic laws (and indeed must be, for the system to meet the technical definition of chaos), but those laws and the factors they depend on may still be extremely complex,

¹Of course, a system need not be chaotic for its later states to be highly sensitive to initial conditions. Systems which are not chaotic may even be more sensitive to those conditions—as Gleick (1987, p. 292) points out, given the attractor states observed in chaos theory, chaotic systems often follow similar large-scale patterns regardless of what happens at a small scale. But the systems I describe here are all genuinely chaotic ones. And, although their chaotic nature is not necessary for demonstrating the problem I have in mind, it is sufficient. (Less importantly, the phrase ‘sensitive dependence to initial conditions’ is also far less catchy than simply ‘chaos’.)

²This is exactly what occurs with double-rod pendulums, even though they are governed by quite simple, deterministic laws. See Rubinsztein (2018) for a helpful demonstration.

resulting in chaotic and unpredictable behaviour.

Three such chaotic systems are: the composition of the human population; the Earth's atmosphere; and the positions of the planets in our solar system. And our actions frequently change all of these, now and indefinitely far into the future.

3.2.1 Identities

Recall the case in which you can either save or not save the child from walking into traffic. If you save the child, they will live out the rest of their life. Conditional on that life being sufficiently long, they may have children of their own, who then may have children of their own, and so on. If they would have a long line of descendants, your choice changes whether or not those descendants come to exist. You would thereby change the composition of the human population well into the future.

How likely is it that that child would spawn a long line of descendants? Very likely, it turns out. At least in the United Kingdom, 97.9% of 5-year-old children live to at least 45 (Office for National Statistics, 2019). And, by the age of 45, 82% of people in England or Wales have had at least one child (Office for National Statistics, 2018). Supposing that the child is 5 years old and that assuming that current rates hold³, it can be shown that the child's lineage will last forever, in expectation. Not just that, but the probability that they have at least *some* genetic descendants at a given time (conditional on humanity surviving until then) converges to 0.7 as the time elapsed approaches infinity.⁴ So there

³Current rates may not hold, but they are still a reasonably good predictor of future rates. We can anticipate that UK birth rates will be a particularly good predictor, as the mean number of births is very close to 2. In the long run, for humanity to not die out from too low a birth rate, the long-run mean will need to be roughly 2.

⁴From ONS (*ibid*), by the age of 45, 18% of Britons have had no children, 18% have had 1 child, 37% have had 2, 17% have had 3, and 10% have had 4 or more. And from ONS (2019), 97.4% of Britons live to the age of 45. For simplicity, I'll assume that no one has children after the age of 45. And, conservatively, I'll assume that no one who dies before 45 has any children, everyone who has 4 or more children just has 4, and that the child's descendants aren't any more likely to have (many) children just because their forebears did. Then, conditional on that first child living to 45, we can represent the number of their descendants in each subsequent generation n by $Z_n = \sum_{i=1}^{Z_{n-1}} X_i$, where the X_i s are independent, identically-distributed random variables with $P(X_i = 0) = 0.201$, $P(X_i = 1) = 0.175$, $P(X_i = 2) = 0.360$, $P(X_i = 3) = 0.166$, and $P(X_i = 4) = 0.097$.

is a probability of at least 0.7 that your choice in that situation will change the genetics of some portion of humans for as long as humanity exists.

And this will change the events that happen in those descendants' lives and the lives of those they interact with. With differences in genetics come countless differences in heritable traits, including height, body mass index (Wainschtein et al., 2019), intelligence (Plomin and Deary, 2015), empathy (Melchers et al., 2016), and plenty of others. At least some of these traits affect the quality of a person's life to a significant extent, and at least some (e.g., empathy) affect the quality of the lives of others. So significantly changing the genetic composition of humanity millennia into the future will also, at least occasionally, affect the moral value that particular humans obtain and the times at which they obtain it.

But this is not the only way your choice will affect the future of humanity. As noted earlier, saving that child would prevent a delay to the vehicle that would have hit them, as well as those that follow it. This will change the positions of those vehicles at each time throughout the remainder of their journeys. As well, due to the dynamics of traffic flow, this will change the positions of countless vehicles throughout your city for the rest of the afternoon. At least a small number of the lucky drivers on the road that afternoon would be on their way home to conceive a child. And you will almost certainly delay or advance the moment at which their child is conceived, by delaying or advancing their parent's journey. But, as pointed out by Parfit (1984, ch. 16) and Greaves (2016), delaying or advancing conception by even a fraction of a second will likely change which sperm fertilises the egg, and so too the genetics of the resulting child. And just like the child you saved (or failed to save), that child is likely to spawn an indefinitely long line of

Readers familiar with stochastic processes may notice that this is a *Galton-Watson* process and that, as such, the probability of any descendants remaining by generation n is given by $P(Z_n > 0) = 0.201 + 0.175 \times P(Z_{n-1} > 0) + 0.36 \times P(Z_{n-1} > 0)^2 + 0.166 \times P(Z_{n-1} > 0)^3 + 0.097 \times P(Z_{n-1} > 0)^4$ (Grimmett and Stirzaker, 1992, §5.4). That probability converges to p as $n \rightarrow \infty$ if there is some p such that $P(Z_n > 0) = P(Z_{n-1} > 0) = p$ solves that equation. Here, it does, for $p = 0.715$. Multiplying that by the probability that the first child lives to 45 (0.979), we have a probability of (at least) 0.700 that the child has descendants for as long as humanity survives.

descendants.

So too, similar changes will be made by that child you saved (or failed to save) throughout their life, and by every one of their descendants. Every time they delay traffic by crossing a road, or encourage (or dissuade) their friends from having children, or introduce someone to their future co-parent, they make similar changes to genetic composition of humanity indefinitely far into the future. And so the scale of the effects of your choice is amplified many times over.

In this sense, the composition of the human population is a chaotic system. By making the tiniest of changes—adding or removing just one person—you will typically make widespread, unpredictable, and permanent changes to the system as a whole.

3.2.2 Weather patterns

Another archetypal chaotic system, and one which we constantly interact with, is the Earth's atmosphere. It was this system, and its behaviour under small perturbations, that inspired Edward Lorenz' work on chaos theory (e.g. Lorenz, 1963).

Suppose you save the child from walking into traffic. By doing so, you will move your body and disturb the air around you, such that the air flow around you will be slightly different from how it would have been had you not dashed forward. The oncoming traffic will continue unhindered, and disturb the air even more. Likewise, your action will result in the child continuing to go about their life for many years, disturbing the air around them with every movement they make. If they end up ever driving a car themselves, they will make even greater disturbances.

But even quite small disturbances to the air within the Earth's atmosphere results in major, unpredictable changes to weather patterns, indefinitely far into the future (*ibid*). Lorenz (1972) claimed, iconically, that changing even an event as trivial as whether a butterfly in Brazil flapping its wings at a given time and place is often enough to change

whether and when a tornado occurs in Texas. As Broome (2019) points out, Lorenz may be wrong—it is unclear whether the disturbance of a mere butterfly flap is dampened out by the air's viscosity or not—but it is clear that larger-scale disturbances do have this effect (Palmer et al., 2014). Some such disturbances are: each human breath, which releases 140 times as much energy as a butterfly's wing flap; and running an electric fan for a single second, which releases 10 million times as much.⁵

Your choice of whether to save the child will change whether, when, and where countless human breaths, uses of electric fans, and other atmospheric disturbances take place. And so you will affect the timing and location of tornadoes, cyclones, wind gusts, rain-fall, and so on indefinitely far into the future. And likewise by affecting the lives of the child's descendants—they too will take breaths, use fans, and so on. But tornadoes and other weather events have significant effects on the lives of many people. So, by changing the timing and location of these events, you affect the moral value that particular humans obtain at particular times indefinitely far into the future.

3.2.3 Planetary motion

Yet another chaotic system is the position of the planets in our solar system under the gravitational influence of one another (Malhotra et al., 2001).⁶

Suppose yet again that you save the child from walking into traffic. By doing so, you cause the child's body to continue moving about the Earth's surface for the remainder of their life, rather than perhaps remain stationary in a cemetery. (Likewise, you change the positions of yourself and a great many cars in your city.)

Like all objects with mass, a human body casts a gravitational field. Move a human body and that field changes. At the most extreme, you might move that body from one

⁵These ratios are calculated from figures given in Dasgupta (2008, p. 78-9), Putensen and Wrigge (2007), Tang et al. (2013), and the packaging of an electric fan I had on hand.

⁶In fact, it was while studying a similar system that Henri Poincaré (1890) inadvertently laid the foundations for chaos theory.

side of the Earth to the other. By doing so, you would slightly change the distribution of mass of the Earth and its inhabitants. But this distribution is important—since the Earth’s mass occupies more than a single point, the strength of the gravitational field from other planets and celestial bodies will differ from one side of the Earth to the other. So changing the distribution of mass can change the gravitational force applied to the Earth and its inhabitants. Just from the gravitational field exerted by Mercury (the planet most often closest to the Earth), moving an 80 kilogram human body from one side of the Earth to the other has an effect equivalent to changing the Earth’s mass by 1.6 milligrams.⁷ Over the course of 24 hours, that results in the Earth’s position changing by 4×10^{-30} metres (or 40 ‘attometres’).

That distance may seem trivial. But the celestial bodies in our solar system form a chaotic system: small disturbances bring large changes, at least over long enough periods of time. If the Earth’s position is shifted by some distance now, it is overwhelmingly likely that its future position is shifted by an even greater distance, although we cannot predict exactly where it will end up. Typically, if you shift the Earth by some distance now, its position in 10 million years will end up shifted by approximately ten times that distance (Laskar, 1989). If you shift the Earth’s position by 4×10^{-30} metres today, you will typically cause a 4 metre difference in the Earth’s position in 300 million years—still trivial, perhaps. But wait another 120 million years and the difference will typically grow to about 4 *billion kilometres*. For reference, that is roughly the distance of Neptune from

⁷The calculation for this is somewhat complex, but follows from the following facts: the gravitational force exerted on one object with mass m_1 by another of mass m_2 and at distance r is given by $F = 6.674 \times 10^{-11} \times \frac{m_1 m_2}{r^2}$; the Earth has a mass of 5.972×10^{24} (kilograms); Mercury has a mass of 3.285×10^{23} and lies at an average distance of 7.7×10^{10} (metres) from the Earth; and the Earth’s radius (and so the distance of the 80kg human from the centre) is 6.371×10^6 . Assume that Mercury lies at its average distance from us, that the human body will be positioned at either at the point on the Earth’s surface nearest to Mercury or the point furthest away, and that the system of the Earth and the human body will have the same centre of gravity regardless of the human’s position. Using the formula for gravitational force, it is straightforward to show that the force applied to Mercury (and, likewise, to the joint body of the Earth plus the human body) increases when the human body is placed on the opposite side, equivalent to increase in mass of 1.569 milligrams. Applying the equation for acceleration, $F = ma$, this gives us a change in acceleration of $9.713 \times 10^{-40} \text{ ms}^{-2}$. Integrating twice for the difference in position, and letting time be the number of seconds in a day, we obtain a difference in position of 3.625×10^{-30} metres.

the sun. And, as you might guess, this would have some serious effects on what occurs here on Earth—for comparison, the full range of temperature variations we experience each year are a result of varying our distance from the sun by a mere 5 million kilometres.

That's what happens if we move a human body from one side of the planet to the other. But similar effects occur if the child you save (or any of their descendants) spends a day in another city or country. The initial difference in the Earth's position will be smaller but, over long enough time spans, it will grow ever larger. And this applies not just to the Earth's position—the bodies in our solar system also exert gravitational fields on neighbouring systems and beyond. Over long enough time spans, these changes would ripple out to affect other planets, any civilisations which inhabit them, and any moral value which arises there. By this route, your choice of whether to save the child will affect events indefinitely far into the future and into remote space. From some time onward, your choice will change the moral value which arises for almost every person at every future time.⁸

3.2.4 The problem of cluelessness

Why are any of these effects a problem for our moral evaluations?⁹

Let us stick with the seemingly clear-cut decision of whether to stop the child from walking into traffic. If you let the child die, the future of the world will proceed in one particular way. If you save them, the future will proceed very differently, now and

⁸This continues to hold even if the future lasts infinitely long and contains infinitely many persons. On the constant cosmological constant model mentioned in Footnote 1, infinitely many future persons will be the product of quantum fluctuations which arise after our universe enters its 'heat death' era. The manner and timing of these fluctuations will be affected—although not predictably—by even subtle changes in gravity (as by the Hawking effect), electric field strength (as by the Casimir effect), and, plausibly, many other factors too. And we can easily influence the gravitational properties of an infinite region of points in the post-heat-death vacuum—by radically changing the positions of planets and star systems over millions of years, as described above, we would produce gravitational changes that ripple outwards in spacetime, infinitely far into the future. So the choice to save the child from traffic, for instance, would continue to have ramifications infinitely far into the future.

⁹The following argument originates with Greaves (2016) and, before that, was hinted at by Lenman (2000).

continuing far into the future: the genetic composition of much of humanity will be altered; the Earth's weather patterns will be altered; and the trajectory of each planet in our solar system will diverge further and further from where they would otherwise be, indefinitely into the future.

To compare these outcomes on an aggregative ethical theory, we look to the total value of each. Assume for now that each has a finite total sum of value, and so too does each interval of time—that those values can be represented on a cardinal scale with a finite real number. (Infinite totals can wait until the next section.) Then we can obtain the total value of each outcome by summing up the value obtained over each consecutive interval of time.¹⁰ For some outcome W_1 , we can write the values for those intervals like this.

$$\begin{array}{cccccccccccc} t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & \dots & t_N \\ W_1 : & X_0 & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & \dots & X_N \end{array}$$

Suppose W_1 is the world that results from letting the child die. Apart from the death of the child, we will, and should, be uncertain about many of the events from then on. So we will be uncertain of the moral value at each later time. This means that, when making judgements about the *actual* outcome of your action, each X_i will be a random variable with some probability distribution.¹¹ (The probability distributions will differ for different times—different X_i s—based on what you *do* know about the future.) And when you say that the total value of the world will probably be such and such, you effectively say that the sum of all X_i s will probably be such and such, if we draw each X_i randomly from their respective probability distributions.

¹⁰If the totals did not admit a finite cardinal representation, we would not necessarily be able to obtain the correct totals by summing over a partition of *times*. As Cain (1995) shows, with non-finite totals, summing over one partition rather than another may give a different total. So we then must ensure that we choose the right partition.

¹¹These distributions may be determined by either your subjective or evidential probabilities. All of what follows can be read in terms of either.

Let's use W_2 to denote whatever happens if you save the child. This outcome too will contain a random value Y_i at each time t_i . This outcome will be better if and only if the sum of all of those Y_i s is greater than the sum of all of the X_i s in W_1 . Or, equivalently, if the sum of $Y_i - X_i$ over every time t_i is positive. But what would those differences be? To simplify matters, we can suppose that the only difference in value at t_0 is whether the value in the life of that child—represented as value 1—occurs or not. And then the difference at each future t_i is at least somewhat unpredictable—your choice will result in seemingly random differences in value at each future time via changes in traffic flow, in the genetic makeup of humanity, in the weather, and in planetary positions. And, since we have no reason to think that each of these differences will favour one outcome or the other, we know that each Y_i is distributed identically to its corresponding X_i . So each difference $Y_i - X_i$ will be drawn from a probability distribution that is symmetric about 0.

Adding up all of those differences, the difference in total value between W_1 and W_2 will look like this:

$$V(W_2) - V(W_1) = 1 + \sum_{i=1}^n (Y_i - X_i)$$

And, by this equation, the difference follows a random walk, and a *symmetric* one at that. As we sum the differences $Y_i - X_i$ up to some large t_n , it will typically look like this.

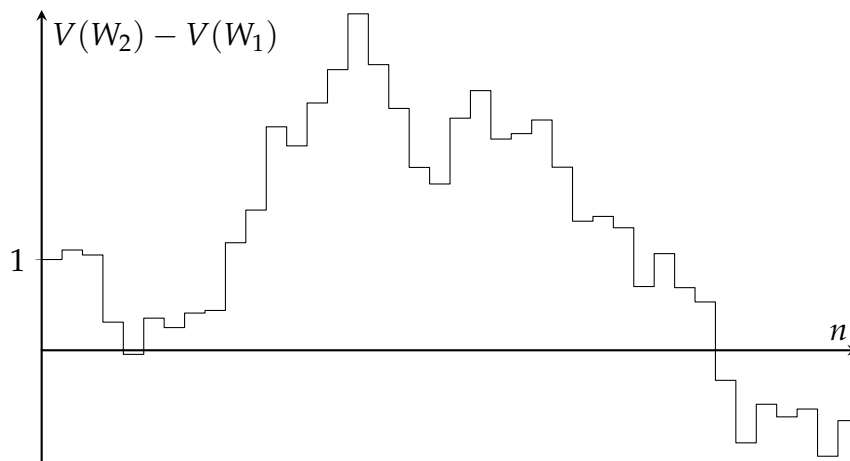


Figure 3.1: The difference in total value between W_1 and W_2 , given by $1 + \sum_{i=1}^n (Y_i - X_i)$

Notice that this walk passes above and below 0 sporadically. In fact, any random walk generated from any sum of *symmetric, independent, identically-distributed* random variables will continue to return to 0 no matter how far along the walk we go—no matter how great a t_n we choose, we know that such a walk will return to 0 at some later time, and that it is then just as likely to rise above 0 as it is to dip below it (Chung and Fuchs, 1951). This entails that, for large n , we cannot predict whether it will be above or below 0. We have no clue. The theory of random walks tells us that the probability that $V(W_2) - V(W_1)$ is positive (or negative) approaches 0.5 as n approaches infinity. In fact, the probability already approximates 0.5 for modest n of even 10 or 100, at least in cases like this where we are certain of just one small, initial difference. But no matter that initial difference, over a long future, it is as likely as not that either outcome turns out worse.

As it happens, the variables $Y_i - X_i$ may not be entirely independent and identically-distributed. After all, the scale of differences between the two outcomes may grow over time. And some later differences will be (at least partially) dependent on earlier differences—e.g., if you cause the Earth’s orbital path to pass nearer to the sun and the resulting increase in temperature does harm, its effects at later times are more likely to

be harmful too. But the above results will still hold—over long enough time spans, the walk will be guaranteed to return to 0, and the probability of being above (or below) 0 will approximate 0.5.¹² So it will still be as likely as not that either outcome turns out worse than the other.

This seems like bad news for the aggregationist. Even in seemingly simple decisions like this—as well as many of the decisions we ever face in practice—we are radically uncertain as to whether an action will actually turn out better than its alternatives. And if we make decisions based even partly on our reasons to promote (actual) moral value, then those reasons *seem* to give no guidance at all.

But this uncertainty isn't fatal for the aggregationist project. We are deeply uncertain about which action will *actually* turn out better, but we may not be uncertain about which action looks better *in expectation*, given our uncertainty—we cannot say which action produces a better world, but we might still be able to say which action produces a better *lottery* over possible worlds. And we can indeed do so for cases like this, as Greaves (*ibid*: 8-9) demonstrates. Take the expectation of the total value for each world, and compare them, as below.

$$\begin{aligned}
 & \mathbb{E}(V(W_2)) - \mathbb{E}(V(W_1)) \\
 &= \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \mathbb{E}(Y_3) + \dots - \mathbb{E}(X_1) - \mathbb{E}(X_2) - \mathbb{E}(X_3) - \dots \\
 &= \mathbb{E}(1) + \sum_{i=1}^n \mathbb{E}(Y_i - X_i) \\
 &= 1
 \end{aligned}$$

Perhaps unsurprisingly, the difference is positive; saving the child brings the greatest expected total value. This is because all we have to tell the two lotteries apart is that we're certain we'll save the one child under one, and certain that we won't under the

¹²For a survey of what are termed *reinforced* random walks like this, see Pemantle (2007).

other. What of the future events, with values represented by X_i and Y_i ? We know nothing about the effects of our actions on them; we place the same probabilities on each possible value of X_i as we do on each possible value of Y_i . And so of course the difference in their expected values is 0, since they're identical. And so, in expectation, the value at every future time cancels out, and all we are left with is the guaranteed value of the child's life.

So, if we switch from comparing actions by the whatever total value their *actual* outcomes contain, to comparing actions by the total *expected* values of their lotteries *over outcomes*, we can say that it is better to save the child. This is the approach defended by Jackson (1991) and other *subjective aggregationists*, who would say that the outcome is *subjectively better* if you save the child, and that this is the normatively relevant sense of betterness. We may be uncertain of which outcome is *objectively* better, but that need not trouble us. So, on at least some aggregative views, we can still say that saving the child is better in a relevant sense and that you *should* save them, despite the presence of widespread chaos.¹³

3.3 Chaos, with infinite value added

As we saw above, if the universe is finite then chaos poses a problem for aggregation. Unpredictable chaotic effects lead to uncertainty over which act will actually have the better outcome. But that need not stop us from making practical judgements—after all, we still know that *some* outcome will be better (or that they will be equally good). We just need to respond to uncertainty over which outcome that will be. And we can do that, by taking the expectations of their (finite) total values.

But if the universe is infinite, chaos poses a greater problem. As we will see, in the infinite context we are not just uncertain of which act will turn out better. Rather, we

¹³The situation is different for *objective* aggregationist views, such as that defended by Railton (1984). Such objective aggregationists will still find themselves clueless as to what they should do.

are *certain* that neither act will. We can say with certainty that the outcomes of any two available acts will be *incomparable* or else equally good. Those are the verdicts given by almost every aggregative view that has been proposed in the infinite setting. So, by those views, we again cannot say which is better—to save the child from walking into traffic, or not to save them.

How we reach those verdicts will differ according to the aggregative theory we adopt in the infinite context. We can separate the available theories into three categories: *strongly anonymous theories*; *weakly anonymous theories*; and *position-dependent theories*. I'll explain why each of these commits us to implausible verdicts, in turn.

3.3.1 Strongly anonymous theories

A *strongly anonymous* theory is one which says that, when comparing one world to another, the comparison depends only on *how many*¹⁴ persons obtain each level of value. It does not matter *which* persons obtain value in each world, nor their positions in space and time, their hair colour, their height, nor any other such qualitative properties except for whether they exist and the level of value they obtain (see §3.3.3 for further discussion). So if two worlds look identical after we censor the identity and those other qualitative properties of each and every person, then those worlds must be equally good.¹⁵

To put this precisely, call a view strongly anonymous if it satisfies the following condition.

Strong Anonymity: Let W_a , W_b , and $W_{a'}$ be any possible outcomes. If $W_{a'}$ can be obtained from W_a by replacing the identities and qualitative properties of

¹⁴Specifically, I mean the *cardinality* of each such set of persons, rather than another notion of size like *containment* or *density*.

¹⁵Strictly speaking, this last claim requires not just Strong Anonymity but also that the $\succsim$ relation is transitive.

Take the worlds W_a and $W_{a'}$ from the definition of Strong Anonymity and let W_b be any world which is equally good as W_a . Strong Anonymity implies that it is equally good as $W_{a'}$ too. Then, by the transitivity of $\succsim$, $W_a \simeq W_{a'}$.

some subset of persons in W_a , holding fixed the value they each obtain, then

$W_a \succcurlyeq W_b$ if and only if $W_{a'} \succcurlyeq W_b$.

When all possible outcomes contain only finitely many persons, Strong Anonymity is satisfied by all plausible aggregative views. But it is less widely endorsed in the infinite setting. This is because, in the infinite setting, Strong Anonymity clashes with other highly plausible principles, e.g., the *Pareto Over Persons* principle (van Liedekerke, 1995), which I'll discuss below. As a result, many views reject Strong Anonymity, including all of the weakly anonymous and position-dependent theories that I mention below. But at least two authors do still endorse Strong Anonymity in this setting: Bader (n.d., p. 85-90) and Clark (n.d.).¹⁶

What do strongly anonymous theories say about realistic cases involving chaos? It turns out that they say cannot say that any outcome is better than any other.

Returning to the case of saving or not saving the child from walking into traffic, we can represent the two outcomes as below. In each outcome, each person obtains some amount of value. For the child p_1 , they obtain value 0 if they die, and 1 if you save them. Then we have some (perhaps finite) number of persons p_2, p_3, p_4 , and so on who are unaffected by your choice and so who obtain the same value in both outcomes. For those persons, you are (at least somewhat) uncertain of the values they each obtain, given by random variables X_2, X_3 , and so on, but these are the same in both outcomes. Then we have infinitely many persons p_i, p_{iii} , and so on who exist in both outcomes but who are affected by your choice. They also obtain random values but, importantly, different random values in each outcome (e.g., X_i as against X'_i). And then we have two distinct sets of infinitely many persons— p_a, p_b , etc and p_α, p_β , etc—whose existence is affected

¹⁶Interestingly, Clark (n.d.) denies Pareto Over Persons as an axiological principle but does still endorse an analogous Pareto principle for *choiceworthiness*. This makes the view strongly anonymous in terms of axiology but only weakly anonymous in terms of choiceworthiness (under a definition of Weak Anonymity different to the one below, which replaces all mention of 'better than' with 'more choiceworthy than'). It also rules out Clark's proposal as a suitable axiology for a purely consequentialist moral theory.

by your choice. In one outcome, one of the two groups exists and obtains some random values. In the other, the other group exists.¹⁷

	p_1	p_2	p_3	p_4	...	p_i	p_{ii}	p_{iii}	...	p_a	p_b	p_c	...	p_α	p_β	p_γ	...
$W_1 :$	0	X_2	X_3	X_4	...	X_i	X_{ii}	X_{iii}	...	X_a	X_b	X_c	...	—	—	—	...
$W_2 :$	1	X_2	X_3	X_4	...	X'_i	X'_{ii}	X'_{iii}	...	—	—	—	...	X_α	X_β	X_γ	...

What do we know about each of these random X values which make up the bulk of the value in the universe? We don't know exactly what each of them will be, as each is random. But we do know that each will be drawn from some probability distribution, perhaps one of those illustrated below.

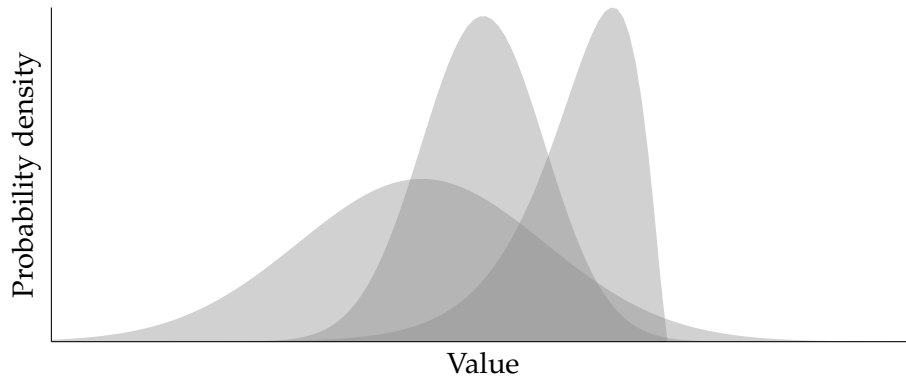


Figure 3.2: Probability distributions for some random variables

We know a few other things about these distributions. For one, we know that at least infinitely many of these X values will be drawn from distributions which give non-zero probability density to the values 0 and 1. After all, if these are possible values of the life of that child on the street, then these values are well within the realm of epistemic possibility for persons in the future

¹⁷Here, we have three infinite subgroups of persons in each world (and possibly four, if $\{p_2, p_3, \dots\}$ is infinite too). Why would these all be infinite? The chaotic effects of your action will extend infinitely far into the future. (See Footnote 9 above.) These include identity effects, so infinitely many people ($\{p_a, p_b, \dots\}$ and $\{p_\alpha, p_\beta, \dots\}$) will exist in each outcome but not in the other. But, as well, infinitely many persons ($\{p_i, p_{ii}, \dots\}$) will have their identities unchanged—for any given future person in one outcome, there is at least a *non-zero* probability that their identity will remain the same regardless of your action, even if the value they obtain is changed. With infinitely many such persons out there, infinitely many will therefore have their identities remain unchanged.

and anywhere else in the universe. For two, we know that the distributions in W_1 and W_2 will be similar—we should have the same distributions for X_i and X'_i and for X_a and some corresponding X_α (even their actual value turns out differently). After all, we have no reason to think that a particular one of the outcomes will be better for each person. For three, it then follows that, for any value v , if W_1 contains infinitely many X values whose distributions give non-zero probability density to v , then so too will W_2 . And for four, it also follows that neither outcome will contain merely *finitely* many X values with some probability of obtaining any given v .

And that gives us enough information to compare W_1 and W_2 . By the law of large numbers, if a world contains infinitely many X values with non-zero probability of v , then it also contains infinitely many persons who *actually* obtain value v (with probability arbitrarily close to 1). And so that number is infinite for every value v that could possibly be obtained in either outcome, including 0 and 1. If, for each outcome, we graph the number of people who end up actually obtaining each value, both graphs will look like this.

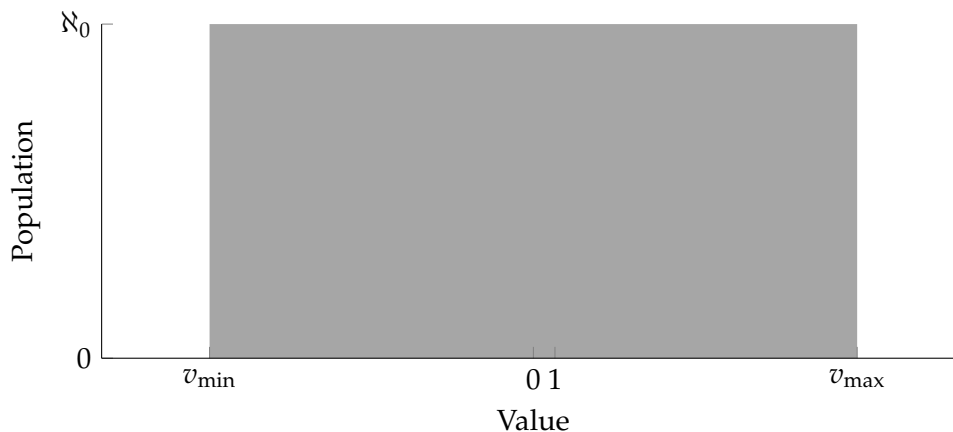


Figure 3.3: The size of the population obtaining each possible level of value

Here, $\aleph_0$ is an infinite cardinality.¹⁸ And $v_{\min}$ and $v_{\max}$ are whichever are the greatest and smallest possible X values, which will be the same for both outcomes W_1 and W_2 . In short, for every possible value a person can obtain, both outcomes contain the same infinite number of

¹⁸In fact, this is the one infinite cardinality which is possible for a set of persons. Why? Any metaphysically possible person must take up at least some minimum volume of space and non-zero duration of time. As spacetime is a four-dimensional manifold with at least one dimension which extends infinitely far, we can only fit countably many persons into spacetime (or $\aleph_0$ persons).

persons obtaining that value.¹⁹

But strongly anonymous theories compare worlds based only on how many persons obtain each level of value. In those terms, the worlds W_1 and W_2 are identical. And so strongly anonymous theories must say that they are equally good.

Thus, all such theories say that the outcome of saving the child from walking into traffic is no better than the outcome of not saving them. Indeed, the outcomes are equally good. And this carries over to almost every moral decision that agents like you or I will ever face in this universe—we can represent the outcomes as we did above and, by similar reasoning, reach the same verdict.

And that verdict is implausible. Surely the world in which the child is saved is better, or at least it could be. And likewise for every pair of available outcomes in almost every other moral decision we ever face. It is a failure on the part of strongly anonymous theories that they cannot say this, at least not in this chaotic, infinite universe of ours. I take this as a great enough failure that we should reject all such theories.

3.3.2 Weakly anonymous theories

But there are other theories of aggregation in the infinite setting to which we can turn. Another major category is *weakly anonymous* theories. These are the theories which satisfy the following.²⁰

Weak Anonymity: Let (W_a, W_b) and (W_c, W_d) be any two ordered pairs of possible worlds. If (W_c, W_d) can be obtained from (W_a, W_b) by replacing some persons across both W_a and W_b with other persons (and the same other persons in both worlds), or

¹⁹As stated, this argument assumes that there are finitely many possible levels of value, and therefore that possible levels of value are discrete. This assumption is plausible, as small differences in duration and length are discrete multiples of Planck time (or length), and so measures of value are likely to be discrete as well.

But a similar argument can still be made if possible values are continuous. In the continuous setting, it seems highly plausible that Strong Anonymity can be strengthened to say that two worlds must be equally good if, for any *arbitrarily* short interval of possible values, they have exactly the same number of persons obtaining a value within the interval. (This is an additional assumption but, I think, a modest one.) But all possible pairs of available outcomes will satisfy this—for any such interval somewhere between $v_{\min}$ and $v_{\max}$, infinitely many persons will have non-zero probability of obtaining value in that interval, in both outcomes. So infinitely many persons *will* obtain a value in that interval, in both outcomes. Therefore, the outcomes must be equally good, at least according to Strong Anonymity.

²⁰In the literature, principles equivalent to this are also called *relative anonymity* (Asheim et al., 2010, p. 10) and *qualitativeness* (Askill, 2019).

by altering some persons' qualitative properties across both worlds, then $W_c \succ W_d$ if and only if $W_a \succ W_b$.

This says something subtly (but crucially) different from Strong Anonymity. Worlds need not be equally good just because they have the same number of persons obtaining the same values. Instead, any two world *pairs* must be compared the same way if those *pairs* contain the same number of persons obtaining the same pairs of values across the two worlds. If two world *pairs* look identical once we censor the identities and those other qualitative properties for each person, but we don't censor the facts of which persons are the same *across worlds*, then those world *pairs* must be compared the same way.

To illustrate the difference between Strong and Weak Anonymity, recall the principle of *Pareto Over Persons* from the previous chapter.

Pareto Over Persons: Let W_a and W_b be any two possible worlds which contain precisely the same persons. If every person obtains at least as much value in W_a as in W_b , then $W_a \succcurlyeq W_b$. If, as well, there is some person who obtains more value in W_a , then $W_a \succ W_b$.

In other words, if two worlds contain precisely the same persons all obtaining the same values, then those worlds are equally good; and if we made one of the worlds better for some and worse for none, then we would make that world better. By intuition, this principle is hard to deny. But deny it you must if you endorse Strong Anonymity—the two principles are incompatible (van Liedekerke, 1995).²¹ And perhaps this shouldn't be surprising. After all, Strong Anonymity forbids us from considering how any specific people do in a world, even compared to another; so we cannot track identities between worlds. But Weak Anonymity does permit it—we can say that one world is better than another based on how individuals do in one compared to the other, just as long as we say the same whenever we face a structurally identical pair of worlds, which Pareto Over Persons always will. So whether they satisfy Pareto is one feature which distinguishes many weakly anonymous theories from strongly anonymous theories.

Like Strong Anonymity, both Weak Anonymity and Pareto are endorsed by all plausible aggregative views when our outcomes contain only *finitely* many persons. But neither is universally

²¹More precisely, they are incompatible if $\succcurlyeq$ is transitive.

accepted in the infinite setting. Like Strong Anonymity, each clashes with other highly plausible principles (such as Pareto Over Positions, from Chapter 2). But unlike Strong Anonymity, Weak Anonymity (and Pareto too) is widely endorsed e.g., in at least some of the proposals in Vallentyne and Kagan (1997, p. 11), Lauwers and Vallentyne (2004), Arntzenius (2014, p. 55), and Askill (2019).

So what do weakly anonymous theories say about realistic cases, in which the available outcomes are infinite and also differ by chaotic effects? Unfortunately, they stumble, following in the footsteps of their strongly anonymous counterparts. They cannot say that *any* outcome is better than any realistic alternative.

To see why, first consider again the same representations of W_1 and W_2 as above. To simplify matters, we can label the relevant subgroups of persons in each: A) those who obtain the same value in each world; B) those who obtain different random values but exist in both worlds; C) those who exist only in W_1 ; and D) those who exist only in W_2 .

	A					B				C				D			
	p_1	p_2	p_3	p_4	...	p_i	p_{ii}	p_{iii}	...	p_a	p_b	p_c	...	p_α	p_β	p_γ	...
W_1 :	0	X_2	X_3	X_4	...	X_i	X_{ii}	X_{iii}	...	X_a	X_b	X_c	...	—	—	—	...
W_2 :	1	X_2	X_3	X_4	...	X'_i	X'_{ii}	X'_{iii}	...	—	—	—	...	X_α	X_β	X_γ	...

We can construct a third possible world, W_3 , such that the pair (W_2, W_3) is structurally identical to (W_1, W_2) and so is (W_3, W_1) . (This construction is somewhat complex, so uninterested readers may wish to skip ahead to just after the comparison of W_3 and W_1 .)

	p_1	p_2	p_3	p_4	...	p_i	p_{ii}	p_{iii}	...	p_a	p_b	p_c	...	p_α	p_β	p_γ	...
W_3 :	—	X_2	X_3	X_4	...	—	—	—	...	0	X_β	X_γ	...	1	X_b	X_c	...

To see why this construction is useful, first note that each of the subsets of locations B, C , and D and their local values will end up being structurally identical to one another—for the same reasons as in the previous subsection, each contains an infinite number of persons obtaining each possible level of value. That includes infinitely many persons obtaining value 0, 1, and whatever else is possible. So, without loss of generality, we can let p_α be some person from the set D who

obtains value 0 in W_2 , and let p_a be some person from C who obtains value 1 in W_1 . We can also let $X_i = X_\beta$, $X_{ii} = X_\gamma$, and so on, as well as $X'_i = X_b$, $X'_{ii} = X_c$, and so on, since the labelling of persons within each set are entirely arbitrary. (Note that X_a and X_α have both been skipped, as those were already specified as 1 and 0.)

Now compare W_3 to each of W_2 and W_1 , with a slight rearrangement of how we list the persons. Both comparisons may look familiar.

	A					B'				C'				D'			
	p_α	p_2	p_3	p_4	...	p_β	p_γ	p_δ	...	p_1	p_i	p_{ii}	...	p_a	p_b	p_c	...
$W_2 :$	0	X_2	X_3	X_4	...	X_i	X_{ii}	X_{iii}	...	1	X_b	X_c	...	—	—	—	...
$W_3 :$	1	X_2	X_3	X_4	...	X'_i	X'_{ii}	X'_{iii}	...	—	—	—	...	0	X_β	X_γ	...

	A					B''				C''				D''			
	p_a	p_2	p_3	p_4	...	p_b	p_c	p_d	...	p_α	p_β	p_γ	...	p_1	p_i	p_{ii}	...
$W_3 :$	0	X_2	X_3	X_4	...	X_i	X_{ii}	X_{iii}	...	1	X_b	X_c	...	—	—	—	...
$W_1 :$	1	X_2	X_3	X_4	...	X'_i	X'_{ii}	X'_{iii}	...	—	—	—	...	0	X_β	X_γ	...

Each of these pairs is analogous to the pair (W_1, W_2) above. By replacing the specific identities in that pair of worlds (in the same way in both worlds), we can obtain the pair (W_2, W_3) . Likewise, we can obtain (W_3, W_1) . And so weakly anonymous theories must say that W_1 is at least as good as W_2 only if the same holds for those other pairs. And therein lies the problem. There is no way that W_2 can be better than W_1 , given that any plausible $\succ$ relation must be transitive. If W_2 were better, we would have a cycle: $W_2 \succ W_1 \succ W_3 \succ W_2$. And similarly if W_1 were better. To avoid cycles, we must accept that either $W_1 \simeq W_2$ or they are incomparable.

But both verdicts are implausible. Recall that these are the outcomes of saving the child from walking into traffic and of not saving them. We cannot say that it is better to save the child. As above, that is implausible. And so weakly anonymous theories are implausible.

Believe it or not, the problem gets even worse than this for most weakly anonymous theo-

ries. Most such theories endorse Pareto Over Persons, as defined above. And so they should, it seems—it's an overwhelmingly intuitive principle and, on the face of it, a compelling reason for abandoning strongly anonymous theories for their weakly anonymous brethren. Indeed, it is treated as axiomatic by each of the authors mentioned above. But if we accept Pareto Over Persons, then the worlds above cannot even be equally good; they must be incomparable. We cannot even evaluate them!

To see why, we can construct another pair of possible worlds, W_4 and W_5 . Again, this pair may look familiar. (And again the construction is complex, so uninterested readers may wish to skip the next paragraph.)

	A					B				C'''				D'''			
	p_1	p_2	p_3	p_4	...	$p_{i'}$	$p_{ii'}$	$p_{iii'}$	...	p_α	p_β	p_γ	...	p_a	p_b	p_c	...
W_4 :	0	X_2	X_3	X_4	...	X_i	X_{ii}	X_{iii}	...	X_a	X_b	X_c	...	—	—	—	...
W_5 :	1	X_2	X_3	X_4	...	X'_i	X'_{ii}	X'_{iii}	...	—	—	—	...	X_α	X_β	X_γ	...

As before, this pair of worlds are carefully constructed to be structurally identical to the pair (W_1, W_2) . But they are constructed to also ensure that each world is related to one of the previous worlds by Pareto Over Persons. This is achieved by first giving p_1 the same pair of values as in (W_1, W_2) . Then every person in A, B, C , and D obtains the same value in W_4 as in W_2 , and in W_5 as in W_1 . On the same reasoning as above, we can assume without loss of generality that $X_a = X_\alpha$, $X_b = X_\beta$, and so on for all of C and D . By similar reasoning, we also know that B can be rearranged to the above: for every possible value between $v_{\min}$ and $v_{\max}$, there are infinitely many persons in B obtaining that value; but since every pair of values X_i and X'_i are independent, there will also be infinitely many persons obtaining each *pair* of possible values across those two worlds; so, for any person p_i who obtains X_i and X'_i in W_1 and W_2 respectively, there is some $p_{i'}$ who obtains X'_i and X_i respectively. And so we can rearrange B as it has been immediately above, without changing the values obtained by any individual.

And so Pareto Over Persons compares W_4 to W_2 and W_5 to W_1 . For W_4 and W_2 , the same persons exist in both worlds; every person does just as well in W_2 as W_4 (everyone in A, B, C , and

D just equally well in both), and p_1 does strictly better; so Pareto implies that W_2 is strictly better. Similarly, Pareto implies that W_1 is strictly *worse* than W_5 .

From above, we know that weakly anonymous views allow only that either $W_1 \simeq W_2$ or they are incomparable. Suppose that they are equally good. Then, again by Weak Anonymity, W_4 and W_5 are equally good. But this generates a cycle once again: W_1 is equally good as W_2 , which is strictly better than W_4 , which is equally good as W_5 , which is strictly better than W_1 . This is impossible, since $\succsim$ is a transitive relation. And so W_1 and W_2 cannot be equally good. They must instead be incomparable.²²

Again, this is implausible, and even more implausible than previous implication: the disjunction that the two outcomes are *either* equally good or incomparable. Surely it is better to save the child from the traffic accident rather than not save them. And surely, at the very least, our moral theory can say *something* about how those outcomes compare. But weakly anonymous theories which endorse Pareto Over Persons—all of those I have seen seriously proposed—cannot compare them. So we should reject them, just as we did with strongly anonymous theories.

3.3.3 What's left?

We've ruled out both strongly anonymous and weakly anonymous theories. What's left? In a sense, we still have available the entire space of possible theories of aggregation which are not even weakly anonymous: those which provide comparisons of worlds which are at least occasionally dependent on the specific identities of the persons in the world pair, or dependent on some qualitative properties of those persons beyond those mentioned above. But many of those possible theories are clearly implausible.

Suppose we adopt a theory which allows dependency on the specific identities of the persons in the world pair but *not* on any further qualitative properties. On such a theory, whether some world W_a is better than W_b depends not just on whether it's better for some people (*de dicto*); it depends on whether it's better for *specific* people. For instance, some W_a may be better than W_b because it is better for Alexander, Bert, and however many other people. But, in the qualitatively identical pair of worlds $W_{a'}$ and $W_{b'}$ containing instead Alexandra, Beth, and however many

²²This argument is adapted from Askill (2019). Askill takes arguments like this as showing that outcomes are often incomparable, rather than as showing that Weak Anonymity and Pareto Over Persons must be false.

others, $W_{a'}$ is not better. And this is absurd.

So we are left with theories by which comparisons depend on some qualitative properties of the persons within them (beyond just the value they obtain and whether they exist). And persons have plenty of qualitative properties to choose from, e.g., their name, height, hair colour, ethnicity, gender, and so on. But, of course, making our moral evaluations sensitive to changes in any of these things is counterintuitive. Doing so seems to run counter to a basic motivation of aggregative theories: that they remain impartial in an important sense. But making our evaluations sensitive to changes in *more than one* of those properties is even more counterintuitive, and an even clearer failure of impartiality. So I will assume that, at worst, a theory of aggregation can have their comparisons depend on *just one* of these further qualitative properties.

But which such property is it plausible that moral evaluations may depend on? I can see only one: position in time and(/or) space.

Why? Suppose you have a world containing infinitely many persons each obtaining such and such values. And suppose we could change some qualitative property (and only that property) of some number of those persons. But the value each person obtains and whether they exist remain fixed. For which property is it least implausible that those changes would mean that those two worlds are no longer equally good?

Suppose we changed the hair colour of some (or even all) of those people. It is clearly absurd that that would produce a world not equally good as the original. Likewise for a great many properties, name, height, ethnicity, and gender included.

But suppose instead that we changed the positions of some (or even all) of those people in space and time. Perhaps we spread them out to be astronomically far from one another. If the original world has value 100 per unit of spacetime volume, the new world might have just value 1 for the same volume. The universe is thereby packed far less *densely* with value. It seems at least *somewhat* plausible that this makes the world worse, as a similar reduction in value density in any universe of finite dimensions would result in a reduction in value. (The same cannot be said of the changes in hair colour and so on, no matter how small the universe.)

And so I would suggest that the only *remotely* plausible theories of infinite aggregation which are neither strongly nor weakly anonymous are those which are *position-dependent*: those by which comparisons of worlds at least sometimes depend on the positioning of persons in space

and time. This is not to say, at least for now, that these theories are anywhere near as plausible as strongly or weakly anonymous theories; they are merely the most plausible of the remaining alternatives and, of those alternatives, the most plausible by far. So the one category of theory I will seriously consider here is position-dependent theories.

3.3.4 Position-dependent theories

Some, but not all, position-dependent theories have difficulty with chaos too.

We saw one such theory in the previous chapter: SE1, my earlier proposal for a plausible principle of spatiotemporal expansionism. For a simpler such theory that works in a similar way, consider *Overtaking* (originally from von Weizsäcker (1965)).

Overtaking: Let W_a, W_b be any possible outcomes, and let $V_a(t), V_b(t)$ be the corresponding functions for the value that arises at each time $t \in \mathbb{R}$.

Then $W_a \succ W_b$ if and only if there exists t such that, for all $t' > t$,

$$\sum_{0 \leq t \leq t'} (V_a(t) - V_b(t)) \geq 0$$

How does Overtaking fare with outcomes which differ by chaotic effects? As you might have guessed by now, it fares poorly.

Consider once again the decision between not saving and saving the child from walking into traffic, with outcomes W_1 and W_2 respectively. As we've seen, those outcomes differ by chaotic effects which continue forever. All that distinguishes them in the agent's eyes is that one contains that child living out the rest of their days, and the other does not. They will also contain a lot of the same events with the same values in both outcomes. And so too will they contain infinitely many events which differ in value in a seemingly random manner. As in §3.2, we know that the equation which appeared in the definition of Overtaking will be equivalent to:

$$\sum_{0 \leq t \leq t'} V_a(t) - V_b(t) = 1 + \sum_{i=1}^n (X_i - Y_i)$$

Here, the values of $X_i - Y_i$ are partially independent random variables that are all symmetric about 0. And so this sum has the same key property as we saw in §3.2: it forms a symmetric

random walk. So, as we sum it up over time, it will again typically look like this.

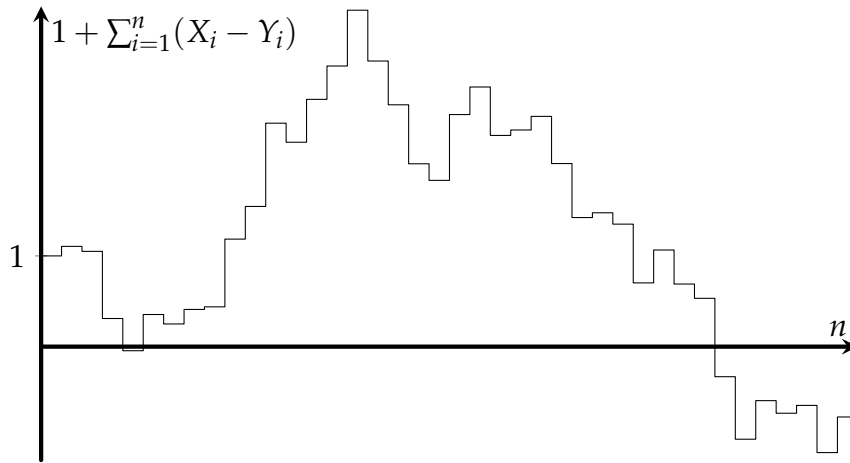


Figure 3.4: A typical sum of $1 + \sum_{i=1}^n (X_i - Y_i)$

As above, any such random walk will pass above and below 0 sporadically. It will never stop doing so, no matter how large n grows (with probability arbitrarily close to 1).²³ But we need the walk to stay at or above 0 from some point onwards (or else at or below 0); otherwise, Overtaking cannot say that either outcome is at least as good as the other. For outcomes which differ this chaotically, Overtaking implies that they are incomparable.

And so we have one of the same implausible verdicts we reached above. As we did with the previous views, we must reject Overtaking too. (Similarly, we must reject SE1.)

So chaos poses just as serious a problem for some position-dependent theories; some, but thankfully not all. There is at least one proposal which does succeed in comparing outcomes as badly behaved as W_1 and W_2 : namely, SE2. Recall the definition from the previous chapter.

SE2: Let W_1 and W_2 be worlds with the same set of locations. For any starting point $\mathbf{p}$, let $\{r_1, r_2, r_3, \dots\}$ be the strictly increasing sequence of distances between $\mathbf{p}$ and each $\mathbf{x}$ such that $V_1(\mathbf{x}) - V_2(\mathbf{x}) \neq 0$.

Then $W_1 \succ W_2$ if the following sum diverges unconditionally to $+\infty$.

²³Crucially, the walk will be *recurrent*: for any value it passes through, it has probability arbitrarily close to 1 of returning to that point, indeed infinitely many times (Chung and Fuchs, 1951).

$$\lim_{r \rightarrow \infty} \sum_{i=1}^r (r_{i+1} - r_i) \left(\sum_{\{\mathbf{x} | d(\mathbf{x}, \mathbf{p}) \leq r\}} (V_1(\mathbf{x}) - V_2(\mathbf{x})) \right)$$

And $W_1 \simeq W_2$ if the sum is bounded both above and below.

To use SE2 to compare outcomes analogous those above, set $\mathbf{p}$ at, for instance, the position of the agent making the decision. Much like the difference in value at a time t_i above, the difference in value at precisely distance r_i from $\mathbf{p}$ will be given by some random variable $X_i - Y_i$. So that long equation in the definition will become:

$$\lim_{r \rightarrow \infty} \sum_{r_j=0}^r (r_{j+1} - r_j) \left(1 + \sum_{r_i=0}^{r_j} (X_i - Y_i) \right)$$

Here, the term in the second set of brackets $(1 + \sum_{r_i=0}^{r_j} (X_i - Y_i))$ matches the equation used by Overtaking above, which we saw returned to 0 over and over again as r_j increased. The full equation here represents the *area* under that graph. While a random walk may return to 0 over and over again, the area under the graph of that walk typically will not. If we plot the area under the cumulative sum we obtained above, we see that that area rises sharply (as given by the black curve).

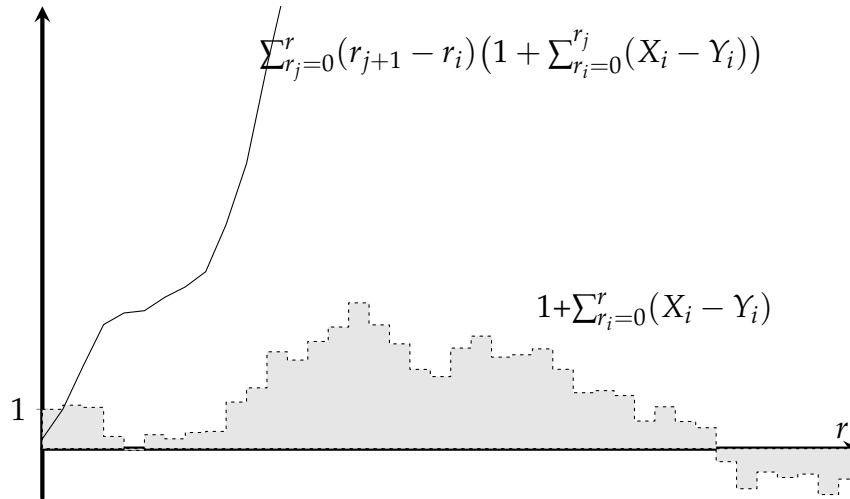


Figure 3.5: A typical sum of $1 + \sum_{r_i=0}^{r_j} (X_i - Y_i)$, as well as the area under that sum

As its graph suggests, this area under the random walk does not follow the walk itself in

returning to 0 over and over again. In fact, it's guaranteed to not do so—it has probability arbitrarily close to 1 of diverging unconditionally to either positive or negative infinity.²⁴ And, if it diverges to $+\infty$, then W_2 is strictly better than W_1 . Likewise, if it diverges to $-\infty$, then W_1 is better. So, for outcomes like these, SE2 is guaranteed to be able to compare them. In fact, it is guaranteed to say that one is strictly better. Good news: at last, we have a theory which can discriminate between such outcomes.

But there is also bad news. We know that the area will diverge to $+\infty$ or $-\infty$, but what is the probability that it goes to each? It is only 0.5. So, even if you adopt SE2, you will be radically uncertain as to which outcome will turn out better. We cannot confidently say that it is better to save the child from the traffic accident. As before, this is worrying—we may want our theory of aggregation to let us say *confidently* than one outcome is better than another.

But this result matches the situation we faced in the finite context (see §3.2.4). There too it was just as likely that one outcome would turn out better as it was for the other. As in that setting, we are left uncertain of which outcome is better but, since we know at least that *some* outcome is, the possibility is left open that we can apply some principle of subjective, instrumental betterness to give judgements under that uncertainty.

And, fortunately, this is easy to do. If we combine SE2 with even a very weak, uncontroversial principle for comparing lotteries, it can be shown that it is *subjectively* better to save the child than it is to let them walk into traffic. (See §5.6 for details.) So SE2 can easily be extended to say not only that *some* action turns out better, but also *which* of them does, at least in terms of subjective betterness. And that is as much as we could hope to be able to say, given that we cannot say more than that in finite chaotic cases (Greaves, 2016) or even in some much simpler cases (see Jackson, 1991).

3.4 Summary

We saw in previous chapters that minimally aggregative moral theories face difficulties if the future is bound to contain infinitely many valuable lives. The situation becomes all the more difficult when the universe is as widely chaotic as ours is: by almost all of the theories of infinite

²⁴To see this, take any finite upper and lower bounds. It can be shown that the probability that the area under a symmetric random walk is within those bounds approaches 0 as $r \rightarrow \infty$ (Lipkin, 2003).

aggregation we have at our disposal, we face the implication that, for almost every pair of actions ever available to us, neither will have a better outcome than the other. Even when the comparison of outcomes seems clear-cut, we must still almost always accept that the outcomes of those actions are either equally good or entirely incomparable.

But, unfortunately, both widespread chaos and an infinite future population are (to the best of our scientific knowledge) facts of life. So must we abandon all attempts to generate moral judgements based on the total value an action results in? No. Some theories of infinite aggregation still allow us to frequently say that one available outcome is better than another. We can still provide plausible moral verdicts.

But even the *most* plausible of those theories of infinite aggregation have a counterintuitive implication: that comparisons of outcomes can depend on the positions in space and/or time of the people within them. In any given outcome, move sufficiently many people around and you may make the outcome better or worse. That seems a bizarre implication for a moral theory. But, for a minimally aggregative theory to make even remotely plausible judgements in a physical setting like ours, we must bite the bullet of position-dependence. Otherwise, we must accept that the world cannot be made better by almost any action we can ever take, which is even more implausible.

But you can also draw a different conclusion from the argument in this chapter. After all, one philosopher's *modus ponens* is another's *modus tollens*. So too, one philosopher's bullet to bite is another's ammunition. You might think it more plausible that all aggregative theories are false than that moral evaluation depends on the positioning of value in spacetime. If so, you may take the argument above as showing that we should reject minimally aggregative theories entirely.

Problems of relativity

4.1 Introduction

At this point, spatiotemporal expansionism may seem to be on firm footing. Its most objectionable characteristic is that its rankings of worlds are position-dependent (and so will sometimes violate Pareto Over Persons). But, as we saw above, position-dependence is unavoidable. Deny position-dependence and we must accept that, for almost any pair of outcomes ever available to us, neither is better.

But position-dependent theories, for the most part, face a further problem that weakly and strongly anonymous theories do not.

Thanks to the physical properties of our universe, the position in space and time that we assign to an event is *relative*. Precisely how far in the future, and how far in the distance, an event is observed to occur can vary with the velocity of whoever does the observing (Cohn, 1904). So too, whether two events occur at the same time or not can depend on that velocity (Einstein 1905; see also Comstock 2010). If you were to simply speed up, the set of events that are happening ‘now’ (as well as the set happening ‘here’) for you would shift; it would include events that a slower observer would claim lie in the distant past or distant future. And that isn’t to say that fast-moving observers are wrong—we have no good reason to accept one observer’s perspective as more valid than another, since we have no scientific experiment which distinguishes one constant velocity over any other as ‘at rest’. There simply is no absolute perspective—no absolute time—by which we can measure the duration or distance between any two events, nor by which we can judge which events occur at the same time. And so the positions we assign to events may vary too.

This relativity of spacetime poses a serious problem for position-dependent theories like spatiotemporal expansionism. If observers can disagree about the positions of events, it seems they can disagree about the rankings of worlds as well. But such disagreement is implausible. Facts of moral betterness must be absolute. When evaluating worlds—rather than judging *acts*, where agent-relative considerations might be relevant—we are interested only in goodness *simpliciter*. And when evaluating outcomes based on their goodness *simpliciter*, we usually think that there can be only one correct evaluation for any pair of outcomes.¹ But we may have multiple, equally valid ways to assign positions to events, and multiple, equally valid rankings of worlds. If so, we will not be able to obtain a unique evaluation, at least not in an extremely arbitrary manner—there simply is no absolute perspective from which observers can apply their position-dependent theory, so there can be no absolute evaluation of outcomes. It seems we must abandon position-dependent theories entirely.

The relativity of spacetime isn't just a problem for position-dependent theories in the infinite setting. An analogous problem arises for some views of betterness in the finite setting. Take the view, common among economists, that value must be *discounted* purely by its remoteness in time from the agent—an outcome is at least as good as another if and only if its sum of *time-discounted* value is at least as great (Arrow, 1999; Dasgupta, 2008). Under such views, it is better for valuable events to occur sooner than to occur later; the sooner they occur, the less their value is discounted. But, in our universe, the duration that elapses between two events is relative to the observer's perspective. So there will often be many different, equally valid time-discounted total values that can be assigned to any given world. By this view, there will be no unique, non-arbitrary total, and so no absolute moral ranking of worlds.²

This problem has been raised before in the infinite setting, by Cain (1995) and Arntzenius

¹The problem raised above might be overcome if we simply allow that evaluations are relative to the perspectives of individual agents. This is rather more extreme than agent-relativity as it usually appears in ethical theory: for one, it would extend into evaluations of outcomes rather than just deontic considerations; and, for two, it would allow judgements to be relative to the agent's *velocity*, which seems absurd.

²Or consider the view of (Temkin, 2015) that it is better for a world to have a more equal *distribution* of value over time. By this view, a world with a more consistent level of value over time will be better than another with the same total sum of value but which brings long periods of time devoid of any value at all. But of course how we evaluate the distribution of value over time will, in our universe, be relative to the observer's perspective. It can be shown that this view too will fail to provide an absolute moral ranking of worlds. For examples of cases in which this view (as well as temporal discounting) fails to give absolute rankings, see §4.2. The cases I use there are just as problematic for these two views as they are for spatiotemporal expansionism.

(2014). Both Cain and Arntzenius show that forms of spatiotemporal expansionism, in particular, fall prey to it. Several solutions have been proposed as well: Arntzenius (*ibid.*) suggests two modifications of spatiotemporal expansionism that appear to avoid the problem. But neither of those modifications succeed, for reasons I present in §4.3 below. We still need a solution to the problem. And, fortunately, I have one, which I describe in §4.4 and §4.5.

In short, the relativistic nature of space and time in our universe poses a serious problem for position-dependent theories. Without a solution to this problem, we would have a compelling reason to abandon position-dependent theories. Given the argument of the previous chapter, this would also give us a compelling reason to abandon minimally aggregative theories entirely. But, as I show in this chapter, the problem can be solved—with some further modification, we can continue to use a form of spatiotemporal expansionism for our moral evaluations.

4.2 The basic problem

Recall the definition of SE2, my proposed version of spatiotemporal expansionism from above.

SE2: Let W_a and W_b be worlds with the same spacetime points. For any starting point $\mathbf{p}$, let $\{r_1, r_2, r_3, \dots\}$ be the strictly increasing sequence of distances between $\mathbf{p}$ and each $\mathbf{x}$ such that $V_a(\mathbf{x}) - V_b(\mathbf{x}) \neq 0$.

Then $W_a \succ W_b$ if the following sum diverges unconditionally to $+\infty$.

$$\lim_{r \rightarrow \infty} \sum_{i=1}^r (r_{i+1} - r_i) \left(\sum_{\{\mathbf{x} \mid d(\mathbf{x}, \mathbf{p}) \leq r\}} (V_a(\mathbf{x}) - V_b(\mathbf{x})) \right)$$

And $W_a \simeq W_b$ if the sum is bounded both above and below.

Recall also that the distance metric d used here is simply the geometric distance between two points, given by $d^2 = \Delta x^2 + \Delta t^2$. As it turns out, this distance metric will give radically different measurements depending on how we assign coordinates to each event in spacetime. Using different, equally-valid, methods of assigning coordinates to events, we can end up obtaining different rankings of worlds. And we cannot pick out just one ranking as correct without arbitrarily selecting one ranking over another.

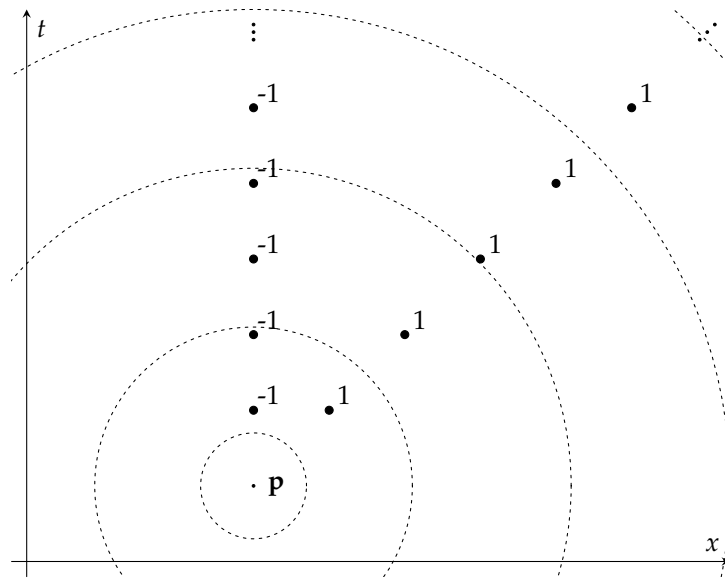
Note that, in what follows, I demonstrate the problem only for SE2. But the examples I use bring on the same problem for SE1 and for the proposals of Vallentyne (1993), Bostrom (2011, p. 16), Arntzenius (2014, p. 56), and Jonsson and Voorneveld (2018). So we cannot avoid these problems by simply switching to a different one of the position-dependent theories currently on offer in the literature.

4.2.1 Galilean relativity

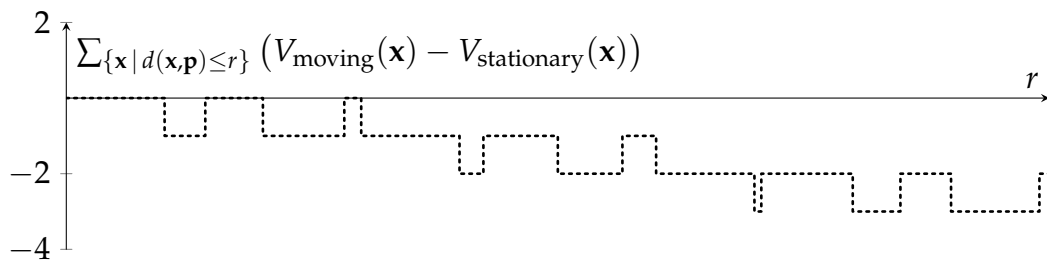
Problems emerge even if space and time are not relative in the Einsteinian sense. We can consider a simpler physical context in which Galilean relativity is present and still encounter serious problems.

To illustrate those problems, consider a seemingly straightforward case. Suppose there will exist an infinitely long-lived civilisation that produces constant amounts of moral value at regular intervals of time. You can either 1) leave that civilisation at a fixed spatial position, producing world $W_{\text{stationary}}$; or set the civilisation moving at some moderate (perhaps very slow) speed relative to yourself, producing W_{moving} . The same values will occur at the same times whether or not you set the civilisation in motion, and your choice will not alter any other events.

Figure 2.19 represents the difference between the two worlds (which is analogous to the differences we saw in Figure 4.1 back in Chapter 2). This world of differences, $W_{\text{moving}} - W_{\text{stationary}}$, has negative values where the civilisation would have been left in place in $W_{\text{stationary}}$, and positive values at the positions it occupies while in motion in W_{moving} .

Figure 4.1: $W_{\text{moving}} - W_{\text{stationary}}$

If we apply SE2, it turns out that this world is worse than the world of zero value; and so W_{moving} is worse than $W_{\text{stationary}}$. To see why, expand from starting point $\mathbf{p}$; and (naively) use the distance metric d based on the coordinate representation in Figure 4.1. The cumulative sums we obtain are as plotted below. As you can see, they quickly fall below 0, and will remain there indefinitely as r approaches positive infinity. And so the area under the graph will approach $-\infty$.

Figure 4.2: Cumulative sum of $W_{\text{moving}} - W_{\text{stationary}}$, with expansions starting from $\mathbf{p}$

The same holds no matter which starting point we choose. So it seems that SE2 judges $W_{\text{stationary}}$ as better than W_{moving} . And this verdict is strange. The only change from the former world to the latter is setting (some of) its inhabitants in motion. They all still live and obtain moral value at the exact same times, but at different spatial positions. It is odd that this could make a

moral difference; that events in the life of a person (or civilisation) in motion can effectively be worth less than if they remained at rest.

But this case produces even greater strangeness if we recognise the Galilean principle of relativity: roughly that, under the laws of Newtonian mechanics, an observer cannot distinguish any particular velocity as being at rest, by any mechanical experiment they perform. So any observer travelling at constant velocity might as well be treated as at rest; the observer may well think they are. So, in our comparison of the civilisation at rest and the civilisation in motion, there is a symmetry between the two outcomes—in each, the resulting civilisation may consider themselves at rest, and that it's the inhabitants of the other outcome who would have been in motion.

Suppose the civilisation's inhabitants in W_{moving} consider themselves at rest. As far as they can tell, they would be. Then, from their perspective, the comparison we performed above would look like this.

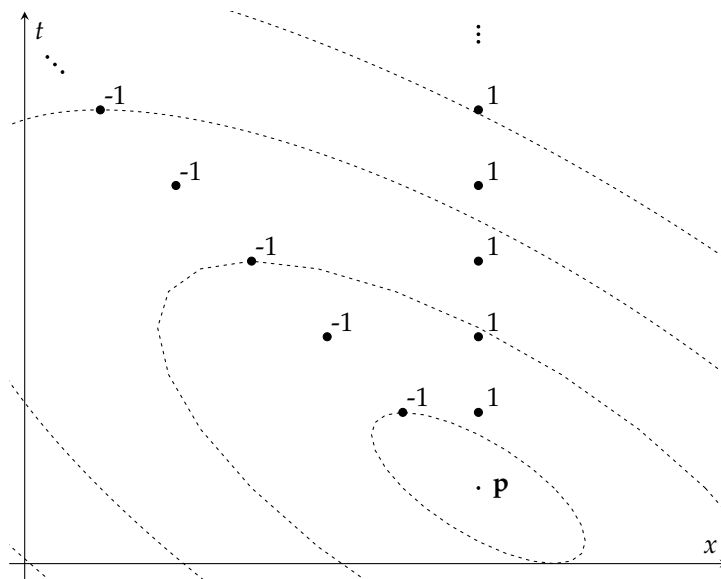


Figure 4.3: $W_{\text{moving}} - W_{\text{stationary}}$, with the regions centred on $\mathbf{p}$ as above, but plotted in a coordinate system under which the velocity of the inhabitants of W_{moving} is treated as at rest.

Those regions around $\mathbf{p}$ now look awfully gerrymandered. To the inhabitants of W_{moving} , they don't look like circles of fixed distance around $\mathbf{p}$; instead they are these strange skewed

ellipses. Were the inhabitants of W_{moving} to compare the worlds themselves, and treated their own velocity as the one at absolute rest and measured distance accordingly, they would draw those regions quite differently (as circles rather than ellipses in Figure 4.3). And, applying SE2 using those regions, they would reach the opposite verdict: that W_{moving} is the better outcome.

So we have a serious problem. The distance measure d , the regions of fixed distance around a given $\mathbf{p}$ and, it seems, the verdicts of SE2 all depend on which velocity counts as being at absolute rest. In Figure 4.3, I assumed that it was the velocity of the inhabitants of $W_{\text{stationary}}$. But the inhabitants themselves may not know that—there's no mechanical experiment they could perform that would distinguish their own velocity as any more 'at rest' than any other. Indeed, modern physics now tells us that there is no experiment at all which would distinguish one velocity over another (Einstein, 1905; Michelson and Morley, 1887)—there is no non-arbitrary way to say what counts as absolute rest. So we have no reason to think there is any such thing. And what does this mean for SE2? Depending on our interpretation of d , it means either: there simply *is no* distance metric d , and so SE2 gives no verdicts at all; or else d is relative to the perspective of the agent (or perhaps the observer), and so the verdicts of SE2 are relative too. But our moral judgements must be absolute, not relative to perspective, so the latter interpretation is implausible. (Even those who endorse moral relativism, who might allow moral judgements to be relative to the cultural setting they are made in, would be unlikely to allow judgements to be relative to the agent's *speed*!) And so SE2 gives no verdicts at all—we have no absolute distance measure, we cannot construct absolute regions of distance d around a point, and so we cannot give absolute judgements. It seems we must abandon SE2 and, along with it, any minimally aggregative moral theory.

4.2.2 Special relativity

Although this problem arises with just Galilean relativity, that isn't the full story. Our understanding of space and time did not stop with Galilean relativity and Newtonian mechanics. We now know that observers at different velocities disagree about much more than what counts as at rest; they also disagree about measurements of spatial distance, about measurements of time, and even about whether two given events are simultaneous. Their perspectives differ far more than they would under Galilean relativity (and, surprisingly, this greater difference will help later

in developing solutions).

To see such a disagreement in action, consider a pair of worlds analogous to before: in $W_{\text{stationary}}$, a civilisation stays 'at rest' and produces moral value at regular intervals; and, in W'_{moving} , that civilisation is instead set in motion. (On this diagram, its relative speed would be $\frac{4}{5}$ of the speed of light, but the same result arises even if the difference in velocity is minuscule.)

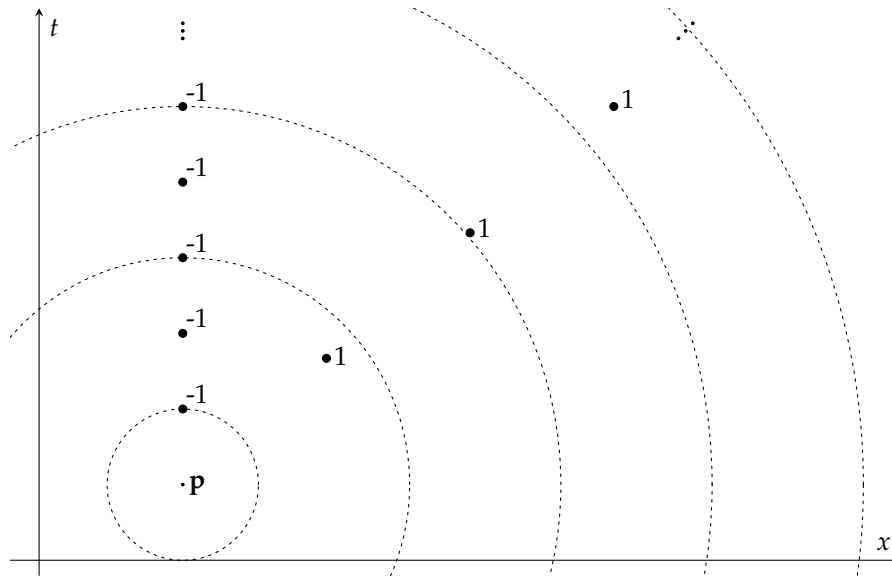


Figure 4.4: $W'_{\text{moving}} - W_{\text{stationary}}$

In both outcomes, the inhabitants of the civilisation produce one unit of value for each one unit of time, according to their own experience of time. But, as you can see in Figure 9, they will not agree that the inhabitants of the other outcome would produce value as quickly as they do. The above figure treats the inhabitants of $W_{\text{stationary}}$ as at rest and, from their perspective, those valuable events that would occur in W'_{moving} are spaced further apart in time; in the time it takes $W_{\text{stationary}}$ to produce three units of value, W'_{moving} appears to produce only one. In fact, the inhabitants of W'_{moving} would observe the exact reverse, that $W_{\text{stationary}}$ produces value more slowly!

This is due to *time dilation*, a phenomenon implied by Einstein's special theory of relativity (Cohn, 1904). If one observer watches t seconds tick by on their own wristwatch, any observer moving at speed v (as a fraction of the speed of light) relative to them will see that same ticking take t' seconds, where $t' = \frac{t}{\sqrt{1-v^2}}$.

Since no velocity is any more at rest than another, the reverse holds as well: the first observer will also see the ‘moving’ observer’s wristwatch tick more slowly. So of course the inhabitants of W'_{moving} and $W_{\text{stationary}}$ will disagree about which outcome produces value more quickly. And, so too, they will disagree about which events in their own outcome would be simultaneous with which events in the other: an inhabitant of $W_{\text{stationary}}$ will observe the first valuable event in $W_{\text{stationary}}$ occurring before the first such event would occur in W'_{moving} ; an inhabitant of W'_{moving} will observe the reverse. Like duration, there is no such thing as absolute simultaneity.

Suppose we try to apply SE2 again. If we were to (naively) apply the distance measure d as defined earlier, based on the perspective of $W_{\text{stationary}}$ ’s inhabitants, we would obtain the regions around $\mathbf{p}$ plotted in Figure 4.5 above: a sequence of nice concentric circles. But if we switched to the perspective of W'_{moving} , those regions would appear as below—with an even greater skew than last time.

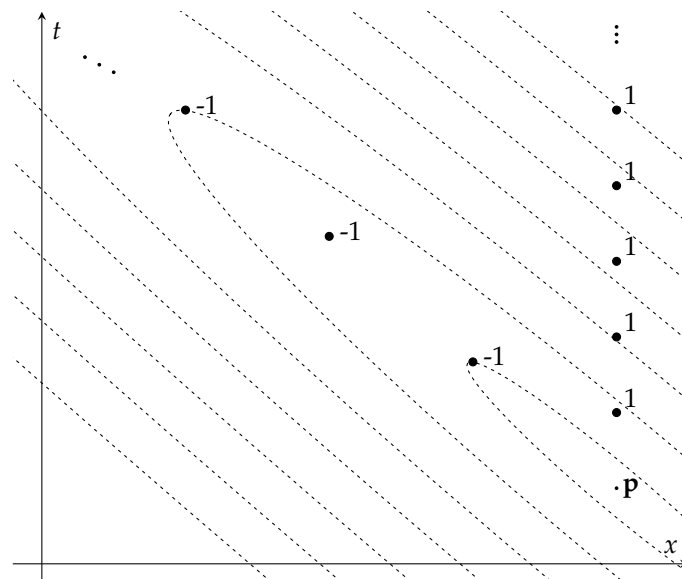


Figure 4.5: $W'_{\text{moving}} - W_{\text{stationary}}$, with regions centred on $\mathbf{p}$ as above, but plotted in a coordinate system under which the velocity of the inhabitants of W'_{moving} is treated as at rest.

As before, those regions around $\mathbf{p}$ now look awfully arbitrary. And again, if the inhabitants of W'_{moving} were to draw the regions of fixed distance d from their own perspective, they would do so very differently. This is because the distance d they would measure between any two (non-

identical) points $\mathbf{x}$ and $\mathbf{p}$ will be different. The measure d , taken in this manner, is not absolute; it is relative to the velocity at which the measurement is taken.

But the deeper problem is this: the definition $d^2 = \Delta x^2 + \Delta t^2$ makes no mention of perspective or velocity; instead, it was implicitly assumed that there were some such absolute quantities Δt and Δx . But there aren't. Any measurements of time or spatial distance will be relative to the velocity of the observer measuring them, by special relativity. So we have no measure d with which to construct regions of fixed distance, we have no such regions, and so we have no verdicts at all from SE2. The principle is silent in *all* cases.

But we might modify the definition of d , and of SE2, ever so slightly to still deliver some judgements. Here is one way we might do that. Define the measures Δt and Δx as quantities relative to the velocity of the agent making the moral comparison. Then d will be relative to their velocity too. And, so too, SE2's verdict will be relative—in the case above, an agent travelling at the velocity of W'_{moving} 's inhabitants would judge W'_{moving} as better than $W_{\text{stationary}}$, and an agent travelling at the velocity of $W_{\text{stationary}}$'s inhabitants would say judge $W_{\text{stationary}}$ as better. And both are right, since SE2's verdicts are relative to the velocity of the judge. But this is absurd. An agent in motion could judge the outcomes one way, while an agent at rest could judge the outcomes in a contradictory way, even when the outcomes are identical in each case? Or, alternatively, an agent could change which verdict is correct in an upcoming decision by simply moving faster? Moral 'relativism' of this kind is implausible, especially when we are seeking to compare outcomes *impartially*, as is a basic motivation of aggregative moral theories. And this relativism is all the more implausible given that these verdicts can change based on *arbitrarily small* differences in velocity—the same phenomena arises in the case above when the civilisations in W'_{moving} and $W_{\text{stationary}}$ differ by *any* non-zero velocity. So this modification is a non-starter.

Here is an alternative one. Again, let the metric d be relative to the velocity of the agent comparing two outcomes. But, for SE2 to make a judgement, don't just require that all *starting points* $\mathbf{p}$ deliver the same verdict. Also require that all *measurements of* d deliver the same verdict. The agent must produce the same ranking of the worlds regardless of their velocity. Otherwise, let SE remain silent: the two worlds under consideration are simply incomparable. But this approach is implausible too. The example above was a fairly mundane one: you can either leave be or set (ever so slightly) in motion some infinite sequence of value. If a method of aggregation

cannot compare these two outcomes, it seems seriously inadequate. And SE2 cannot—it falls silent, even in cases as mundane as this. So, again, it seems we must abandon SE2.

4.3 Existing solutions (and their problems)

This relativistic problem may seem fatal to the expansionist approach. SE2, and its nearby modifications, are either universally silent or otherwise implausible. So must we abandon it entirely, and moral aggregation along with it? I hope not. Here are two other proposals of how we might salvage it, both from Arntzenius (2014). Unfortunately, each of these proposals faces further problems, and serious ones at that.

4.3.1 Double lightcones

Here is one suggested solution, described by Arntzenius (*ibid.*: 44) and credited to Cian Dorr.

In the definition of SE2, forget about any distance measure d . All we need is a sequence of expanding regions of spacetime, each neatly bounded and each containing all of the regions that came before it. And, to overcome the relativistic problem, we need that sequence of regions to not vary with the perspective of the observer.

Under special relativity, we can obtain such regions as follows. Select two spacetime points, one earlier in time than the other according to all observers—perhaps two points with the same spatial position, as illustrated below. For the earlier point, draw its future *lightcone*: the region of points you could reach if at that point you emitted particles travelling at or below the speed of light in any direction. And, for the later point, draw its past lightcone: the region of points from which you could emit such particles and have them reach the later point. The intersection of those two regions—their ‘double lightcone’—forms a diamond-shaped region with both endpoints as vertices. Repeat this process with additional pairs of points to obtain your sequence of expanding regions. One such sequence of regions is illustrated below, expanding around $\mathbf{p}$, applied to the same example of W'_{moving} versus $W_{\text{stationary}}$ from above.

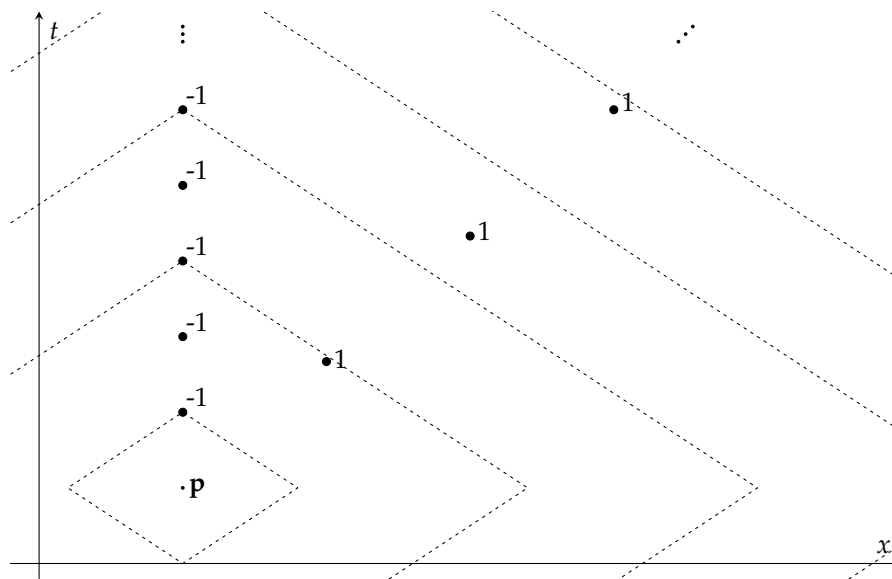


Figure 4.6: $W'_{\text{moving}} - W_{\text{stationary}}$ with a sequence of double lightcones around $\mathbf{p}$

We can apply a modified version of SE2 with this sequence of regions instead of the expanding circles of fixed radius. If we did, we would sum the -1 s faster than the $+1$ s, and so our cumulative sum would approach $-\infty$. We would be led to the verdict that W'_{moving} is worse than $W_{\text{stationary}}$. But is this verdict absolute, no matter the observer's velocity?

It is true that these regions would be constructed the same way by observers moving at any velocity, as long as those observers use the same pairs of endpoints to construct them. But the challenge lies in selecting those pairs of endpoints in the first place. To get a sense of why different observers or agents might disagree about which endpoints to use—and hence which double lightcones to use—consider that same sequence of regions from the perspective of an observer moving at the same velocity as the inhabitants of W'_{moving} .

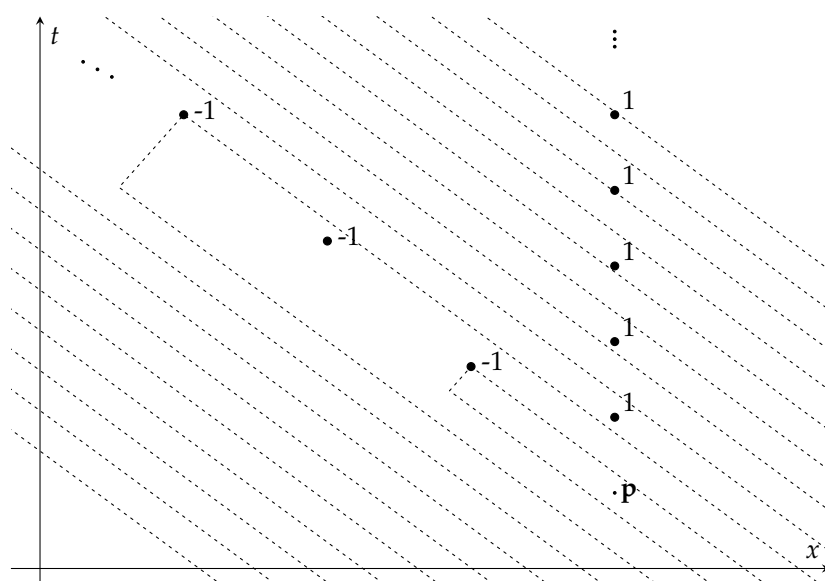


Figure 4.7: $W'_{\text{moving}} - W_{\text{stationary}}$ with the same sequence of double lightcones, viewed from the perspective of the inhabitants of W'_{moving}

Much like in the previous section, these regions now look awfully arbitrary. And we could easily construct very different ones. For instance, suppose we followed the same method as above and selected endpoints at the same spatial position as $\mathbf{p}$, but did so from the perspective of W'_{moving} . We would generate very different regions which, from the perspective of W'_{moving} , would have the same nice diamond shape as we saw in Figure 11 (and the same strange rectangles as in Figure 12 if we switched back to the previous perspective). But this would give us a different verdict: summing over such a sequence of regions we would sum the $+1$ s much faster than the -1 s; our cumulative sum would approach $+\infty$; and so we would conclude that W'_{moving} is better than $W_{\text{stationary}}$.

In short, there is no non-arbitrary way to select the endpoints from which we construct the sequence of double lightcones.³ Simply placing them at some particular position relative to your starting point $\mathbf{p}$ (e.g., at the same spatial position at an earlier/later time) will always result in different regions for agents travelling at different velocities. And different selections can deliver

³One way we might try to select those endpoints without arbitrariness is to set them at the positions of the valuable events in each world—at the $+1$ s and -1 s in the diagram. But setting each of those points as the later endpoint of a double lightcone will produce regions which overlap only partially and so do not form a sequence of expanding regions.

different verdicts, as demonstrated above. So this solution does not succeed; the problem remains.

4.3.2 The spacetime metric

Here is another suggested solution, also from Arntzenius (*ibid.*: 43).

In the definition of SE2, simply replace the distance measure d with a better measure of 4-dimensional distance: one that is invariant across changes in velocity. In the most basic spacetime of special relativity we have one such metric—the *spacetime metric* s —which is defined as follows.

$$s^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2$$

As before, Δy and Δz will be 0 in the examples below, so this simplifies to $s^2 = \Delta x^2 - \Delta t^2$.

But what does s represent, other than some abstract mathematical quantity? It actually measures the ‘proper time’, or ‘proper distance’, between any two points in spacetime. The *proper time* between a pair of points is the duration of time we’d record on a clock travelling from one to the other at constant velocity. (Note that this only exists if you *could* reach one from the other by travelling at less than the speed of light.) Meanwhile, the *proper distance* is the spatial distance that you’d measure between them if you were travelling at just the right velocity to see those points as simultaneous. (And this only exists if there is such a velocity—if neither lies in the other’s future lightcone.)

If we replace d with s in the definition of SE2, the regions we generate around $\mathbf{p}$ will look like those in Figure 4.8 below.⁴ Each of these hyperbolic regions corresponds to the set of all points within some constant proper time (or distance) of $\mathbf{p}$. Effectively, each region is the set of points that you could reach from $\mathbf{p}$ within some time $|s^2|$ (from your own perspective) by travelling at a constant speed, *or* that would look like they were within distance $|s^2|$ of $\mathbf{p}$ from some perspective. But recall that time passes more slowly for fast-moving observers, and lengths appear shorter. So, for two future points which seem to you to lie at the same time, the one *further* from you in space would actually be the *smaller* proper time from your position. Likewise, for two points which seem to you to lie at the same spatial position but at different times, the one further from you in time would be the smaller proper distance from you.

⁴Unlike d , s sometimes takes on imaginary values (whenever $\Delta t > \Delta x$). So $|s^2|$, or $\sqrt{|s^2|}$, actually serves as an appropriate replacement for d .

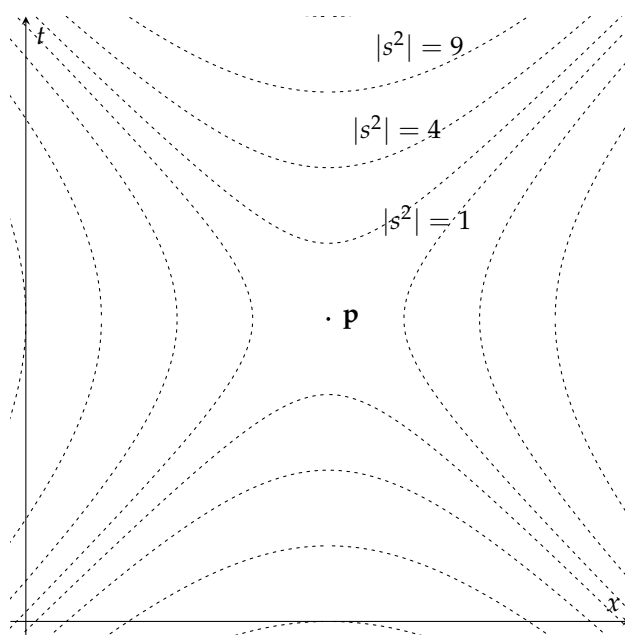
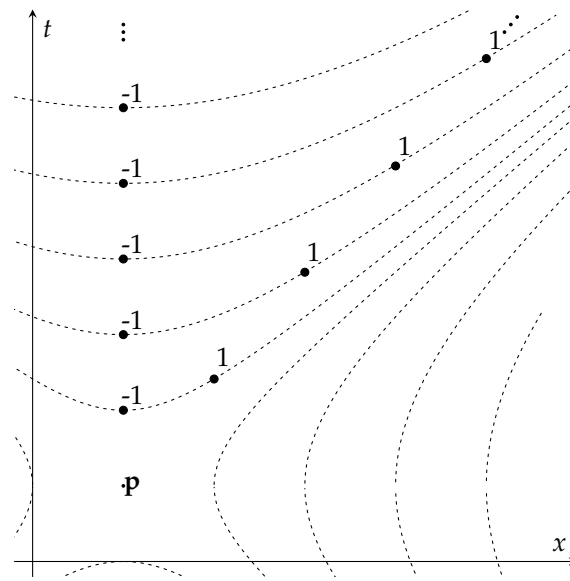


Figure 4.8: A sequence of regions each comprised of points within some fixed $|s^2|$ of $\mathbf{p}$.

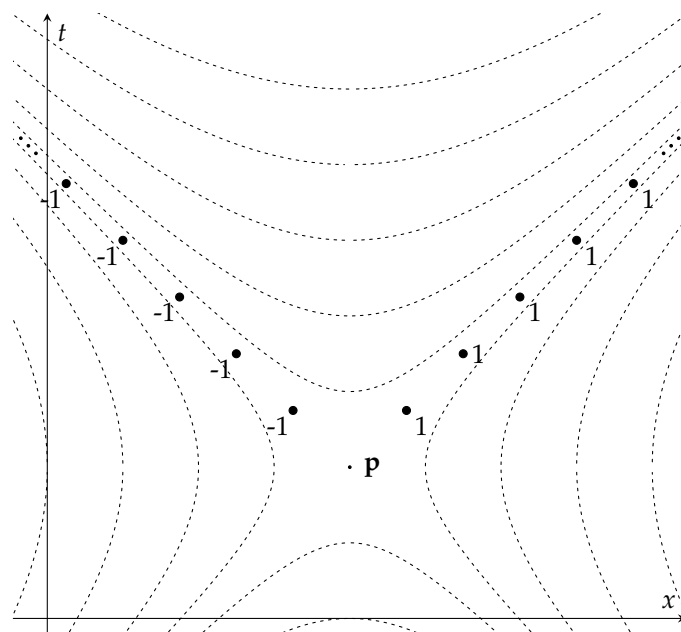
Crucially, the shape of each of these regions will not differ at all between the perspectives of observers at different velocities. So too, the value of s between any two points will not change. So any two observers would draw precisely the same regions no matter their velocity. They would draw specific points in different places, but they would still draw them within precisely the same regions as one another. So we have an expanding sequence of regions that everyone can agree on, and over which we can apply SE2 without controversy.

For instance, in the problem case of W'_{moving} versus $W_{\text{stationary}}$ from above, we would construct our regions as below. And as you can see, each -1 and its corresponding $+1$ lie on the boundary of the same region—they lie at the same distance from $\mathbf{p}$, according to the spacetime metric, and this would be the case even if we switched perspectives.

Figure 4.9: $W'_{\text{moving}} - W_{\text{stationary}}$

So, if we applied SE2 using these regions instead of the circles from earlier, each of those -1 s would cancel out the corresponding $+1$. Our cumulative sum would be 0 at all times (and, if we started from a point other than $\mathbf{p}$, it would be 0 for all sufficiently large regions). So we would judge W'_{moving} and $W_{\text{stationary}}$ as equally good. And recall that this comparison was between a world in which a sequence of valuable events was left stationary, and one in which those events were moving but still occurred at the same frequency (from the perspective of those involved in the events). As it turns out, we will obtain the same verdict for any such pair of worlds, no matter their difference in velocity. And that seems like the right verdict; changing the velocity of your descendants shouldn't make the world better or worse!

But, as you might guess, this supposed solution faces a serious problem of its own: any of those regions around $\mathbf{p}$ can contain *infinite* value (or *undefined* total value). For instance, consider a comparison of worlds W_{right} and W_{left} which differ according to the diagram below. W_{right} contains a sequence of valuable events extending off towards the right, while W_{left} contains a sequence extending off towards the left.

Figure 4.10: $W_{\text{right}} - W_{\text{left}}$

In $W_{\text{right}} - W_{\text{left}}$, every single one of those points with value $+1$ or -1 lies at $s = 0$ from $\mathbf{p}$. Suppose we try to take a cumulative sum, starting from $\mathbf{p}$ and summing value in order of its distance s from $\mathbf{p}$. That cumulative sum will break down immediately: we cannot assign a finite value on the region with distance 1 of $\mathbf{p}$, or even the region within distance 0.000001 of $\mathbf{p}$; the sum of value in each region doesn't converge to any value, nor does it diverge unconditionally to $+\infty$ or $-\infty$.

So this modified version of SE2 that uses the spacetime metric s brings us to a position similar to where we started: we wanted to compare worlds containing infinite value, but standard addition on the real numbers does not allow us to do so. For the same reason, we cannot compare worlds like these.⁵ So this proposal is inadequate.

⁵Why not just accept that worlds like this are exotic enough that it is fine if our theory cannot compare them? The key reason is that I suspect that every comparison we ever need to make, in practice, resembles this case. As argued in Wilkinson (n.d.), almost every significant moral decision we ever make will affect infinitely many future persons and an infinite quantity of future value. For some such decision between two acts, let W_1 and W_2 be the resulting worlds, and $W_1 - W_2$ the world composed of the differences between them. And suppose we take some region of spacetime which has an infinite volume lying in our future lightcone (our causal future). Within that region, we should expect $W_1 - W_2$ to contain infinitely much positive value and infinitely much negative value; the total sum of value in $W_1 - W_2$ within that region will be undefined. We face the same problem as in the case above. So I want a theory which deals with the case above because, without one, we likely cannot compare the actual outcomes of everyday decisions.

4.4 A better solution

That last solution was inadequate, but there is a promising solution in its near vicinity.

As Arntzenius diagnoses it, the problem with the last solution is that each of those regions we identified has infinite 4-dimensional volume.⁶ And thanks to this fact and their shape, we can arrange infinitely many valuable events within each of them, as we say in Figure 15 above.

To avoid this problem, it is sufficient to have a sequence of expanding regions each of which has only finite 4-dimensional volume (and which observers won't disagree about). Then each of those regions will contain only finite value, and we can proceed with our comparisons as before. But why is finite volume enough to guarantee finite value? I assume—quite plausibly, I think—that we can only ever fit finite moral value into a finite volume. Certainly on most standard accounts of what constitutes moral value, we cannot produce infinite value in merely finite space and time. For instance, a person takes up some finite spatial volume and, to undergo some extremely valuable experience, they require some duration of time over which to experience it; so the valuable event requires some non-zero four-dimensional volume. There are ways we can produce more and more value—we can make that person's valuable experience last longer, and/or we can add additional people having valuable experiences of their own—but to do so we must effectively use up a larger and larger volume of spacetime. You cannot scale up the amount of value to an infinite quantity without also scaling up the volume of spacetime involved to some

⁶There are two notions of volume we might use here. The first is the standard Euclidean notion, extended to four dimensions—a region with constant *spatial* volume V and duration in time t simply has 4-dimensional volume $V \times t$. The alternative is 'Riemannian volume', which is generated from the metric s . Fortunately, in flat spacetime, the Riemannian volume of any given region will simply be -1 times its Euclidean volume. So, if one is finite, so will be the other. And, interestingly, since the Riemannian volume is the same for all observers, so will be the Euclidean volume.

Sticking with the Euclidean volume, we can calculate the volume within one of those regions as follows. Take the region centred on $\mathbf{p}$ and containing all points within $|s^2| = 1$ of $\mathbf{p}$ —the innermost region in Figure 4.10 above. And, for simplicity, adopt a coordinate system such that $\mathbf{p} = (0,0,0,0)$ and that the boundary of the region intersects $(1,0,0,0)$ and $(0,0,0,1)$. Then note that the spatial volume of the region at any t will be $\frac{4\pi}{3}\Delta x^3$, where Δx is the total length of the intervals of possible values of $x > 0$ within the region at that time t . (Equivalently, Δx is the width of the region at t on the right of $\mathbf{p}$ on the graphs above.) We can then obtain the volume by taking the following integral.

$$\begin{aligned} \text{volume} &= \int_{-\infty}^{+\infty} \frac{4\pi}{3} \Delta x^3 dt \\ &= 2 \int_0^1 \frac{4\pi}{3} \sqrt{t^2 + 1}^3 dt + 2 \int_1^{+\infty} \frac{4\pi}{3} \left(\sqrt{t^2 + 1}^3 - \sqrt{t^2 - 1}^3 \right) dt \end{aligned}$$

And that second integral diverges unconditionally to $+\infty$. Likewise for any of the regions bounded by $|s^2| = k$; simply replace the 1s within the square roots with ks .

infinite volume; you cannot fit infinite value into finite volume.⁷

But, actually, there is a sufficient condition for making expansionism work which is even weaker than this. We don't need our regions to have finite volume; we just need them to have only a finite *portion* of their volume lying in the agent's future lightcone. Recall from earlier that a point's future lightcone is the region of all points you could reach by travelling from that point at or below the speed of light. To the best of our knowledge⁸, all of the events in the universe that you or I can affect lie in our future lightcone—to change events anywhere else, we would need to send a signal faster than the speed of light, or else back in time. So in practice, whenever an agent chooses between two actions, the outcomes of those actions cannot differ at any points outside the future lightcone of where and when the agent takes the action. That means that, at any point $\mathbf{x}$ outside the lightcone, the difference in value between the two outcomes will be $V_a(\mathbf{x}) - V_b(\mathbf{x}) = 0$. So, going back to our expanding regions, it won't actually matter if such a region contains an infinite volume, as long as it only has finite volume within the agent's future lightcone. This is enough for any of the pairs of outcomes we ever actually need to compare—those that arise from the alternative actions of a real-world agent.

With this knowledge in hand, we can solve the problems of the previous section. As before (and as depicted by the dotted lines below), we can use regions given by the spacetime metric, each consisting of all points within some fixed distance of $|s^2|$ from some central point $\mathbf{p}$. But suppose that $\mathbf{p}$ must be in the agent's past. Then each region around $\mathbf{p}$ will only have finite

⁷Here is one benign exception. Suppose a person lives for an infinite duration and, during their life, obtains infinite total value. If we treat their entire life as a single event and associate it with a single point in spacetime, then any region containing that point will contain infinite value, even if the region has only finite volume. And so there would be no sequence of expanding regions which could sensibly compare this outcome to another.

This possibility may seem problematic, but the problem is due to the simplicity of how I have so far modelled the world. We need not treat entire lives as single events; we could instead decompose them into smaller events. Nor do we need to associate events with discrete points in spacetime. We could instead distribute the value across the entire densely-packed regions of spacetime that they occupy. Admittedly, this would bring complications. For instance, the natural way to divide value evenly across infinitely many densely-packed points in a region would be via a *value density function*. To sum the value across a dense region, we would then need to take the (Lebesgue) integral of value density across the region, rather than simply taking a discrete sum. But we can overlook this complication here—for simplicity, I will stick with examples involving only finite values at discrete points.

⁸One of the stranger findings of quantum physics is that two entangled particles can be separated by significant distance but still behave in a strongly correlated manner. But, thankfully, this 'spooky action at a distance' (as Einstein called it) does not involve any transfer of information from one location to another, nor any transfer of force or physical work (see Eberhard and Ross, 1989). Rather, a measurement on one particle returns a random result and the same measurement on its entangled twin yields a correlated value. So we would not be able to causally affect events outside our future lightcone using quantum entanglement.

volume within the agent's future lightcone, as we needed.

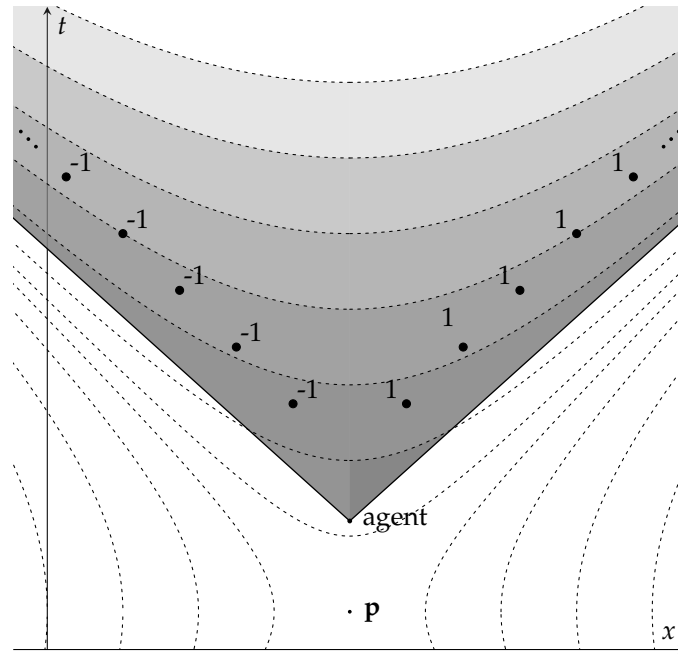


Figure 4.11: $W_{\text{right}} - W_{\text{left}}$

Suppose we used this sequence of regions to sum the value in $W_{\text{right}} - W_{\text{left}}$. The cumulative sum of value would remain at 0 at every stage—each region contains precisely as many 1s as -1 s, so they'll cancel one another out. Likewise if we start at any starting points in the agent's past lightcone and expand out far enough, the events in each sequence all cancel one another out. If only *every* allowable sequence of regions said this, we might be able to say that W_{right} and W_{left} are equally good.

And they can, if we restrict the allowable sequences of regions to those like this, that contain only finite value in each region. That brings us to my preferred modification of SE2, which I'll call *Relativistic Spatiotemporal Expansionism* (RSE).

RSE: Let W_a and W_b be worlds with the same spacetime points. For any starting point $\mathbf{p}$, let $\{r_1, r_2, r_3, \dots\}$ be the strictly increasing sequence of distances (according to the spacetime metric) between $\mathbf{p}$ and each $\mathbf{x}$ such that $V_a(\mathbf{x}) - V_b(\mathbf{x}) \neq 0$.

Then $W_a \succ W_b$ if there is *some* starting point $\mathbf{p}$ such that the following sum diverges unconditionally to $+\infty$, and *no* starting point $\mathbf{q}$ for which it is bounded both above

and below or diverges unconditionally to $-\infty$.

$$\lim_{r \rightarrow \infty} \sum_{i=1}^r (r_{i+1} - r_i) \left(\sum_{\{\mathbf{x} \mid |s(\mathbf{x}, \mathbf{p})^2| \leq r\}} (V_a(\mathbf{x}) - V_b(\mathbf{x})) \right)$$

And $W_a \simeq W_b$ if for there is *some* starting point $\mathbf{p}$ such that the sum is bounded both above and below, and *no* starting point $\mathbf{q}$ for which it diverges unconditionally to $+\infty$ or $-\infty$.

Although it seems complex, RSE has us follow almost exactly the same procedure as SE2: sum the value in W_a and W_b in order of its distance from $\mathbf{p}$; and, if W_a 's sum takes the lead and stays in the lead beyond some distance r' (or if the area under the graph of $\sum_{\{\mathbf{x} \mid |s(\mathbf{x}, \mathbf{p})^2| \leq r\}} (V_a(\mathbf{x}) - V_b(\mathbf{x}))$ is otherwise infinite), then W_a is better; and, if the sums are equal and stay equal beyond r' (or the area under the graph is otherwise bounded), then the worlds are equally good. The only differences are: we use the distance given by the spacetime metric (via $|s^2|$) to avoid problems with relativity; and, crucially, we don't need *all* starting points $\mathbf{p}$ to agree. To say that W_a is better, it's enough to have just one starting point $\mathbf{p}$ that gives W_a the winning sum, *as long as* there is no other starting point $\mathbf{q}$ that disagrees. Similarly, to say that the worlds are equally good, it's enough to have just one point $\mathbf{p}$ that says that the sums become equal, as long as there is no $\mathbf{q}$ to disagree.

In the case of W_{left} and W_{right} in Figure 4.11 above, we have a starting point $\mathbf{p}$ which generates a sequence of expanding regions over which the differences between worlds sum to 0 no matter how far we expand. And, as it turns out, there is no starting point that says that the sum is strictly positive or strictly negative (and finite)⁹ beyond some distance r' . So RSE judges W_{left} and W_{right} as equally good. And, intuitively, so they should be! One world involved sending off some sequence of value into space in one direction; the other involved sending it off in another, precisely mirrored, direction at the same speed; so neither world should be any better than the other.

So RSE delivers a judgement here, and the intuitively plausible one at that. And, since it uses a distance metric on which all observers can agree, it gives consistent verdicts no matter the

⁹There are some points we might choose that give an infinite value at every single stage of the sum: those precisely in line with either of the two sequences of value (but not both). This is why RSE specifies that starting points are only allowable if they give finite sums.

velocity we are travelling at while comparing outcomes. The problem is solved.

4.5 General relativity

The basic problem posed by relativity was that, once we considered the implications of the special theory of relativity, SE2 delivered verdicts very different from what we anticipated (or else no verdicts at all). So it is natural to ask at this point: what happens to this new aggregation method, RSE, if we consider the implications of the *general* theory of relativity?

Under special relativity alone, spacetime is *flat*: objects not subject to any force travel in straight lines, which respect some basic properties of Euclidean geometry; if those paths are parallel at some stage, they will never meet; if orthogonal, they will only ever meet once. But under general relativity, depending on the matter and radiation present, spacetime may be *curved*: those same objects may follow curved paths which are parallel (or orthogonal) at some stage but still meet eventually (or again). This curvature corresponds to what we might otherwise think of as a gravitational ‘field’—massive objects appear to accelerate towards one another. But, given the discrepancies we observe in measuring distance and time (e.g. Hafele and Keating, 1972), we know that this is not due simply to a field of force; it is due to curvature of spacetime itself.

In particular, in a curved spacetime, the metric s from will no longer apply universally. Over small enough regions, it will often still deliver measurements which approximate proper time and proper distance. But over larger regions it will not, and so different observers will measure different values of s . So we have a problem: when spacetime is curved, RSE as defined above will fail to deliver verdicts on which all observers will agree.

But this problem is easily solved. Every curved spacetime we encounter will have a close analogue of s , often simply called its metric, or g . This metric depends on the distribution of matter and radiation across spacetime. And so the measurements it produces between two points may depend on the *absolute* position of those points. But this is fitting, since the proper times (or proper distances) between those points depend on curvature, and so on absolute position, as well.

Here is one example of a metric for a curved spacetime. Suppose that spacetime contains just

one massive object¹⁰, which approximates the shape of spacetime close to the Earth or to our sun. This is called a *Schwarzschild spacetime*, and has the associated *Schwarzschild metric* g given by the following complicated-looking equation (Schwarzschild, 1916).

$$g = \frac{r}{r-k} dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 - \left(1 - \frac{k}{r}\right) dt^2$$

For readers who are curious, g is given in polar spherical coordinates, with r our distance from the centre of the object, θ and ϕ our polar coordinates on the sphere at that radius r , t the position in time, and k a positive constant determined by the object's mass. And we have g defined in terms of dr, dt , and so on. Why not $\Delta r, \Delta t$, and so on? Since the curvature varies from across spacetime, so will the equation for g from point to point. What this equation gives us, for each point $\mathbf{x}$, is the *rate of change* of the distance measure at $\mathbf{x}$. To actually obtain the distance measure *between* two points, we need to take the integral of g along a straight path between the two points, and then we need to choose the lowest value given by all of those paths—call it g' .

Equations aside, this is what that metric actually looks like (holding θ and ϕ fixed). The massive object is positioned at $r = 0$. And around the point $\mathbf{p}$ we have the regions of points within various fixed distances g' . There is also an agent placed slightly in the future of $\mathbf{p}$, just to show how their future lightcone intersects with those regions.

¹⁰By assumption, this body has no electric charge nor angular momentum, and the cosmological constant is 0.

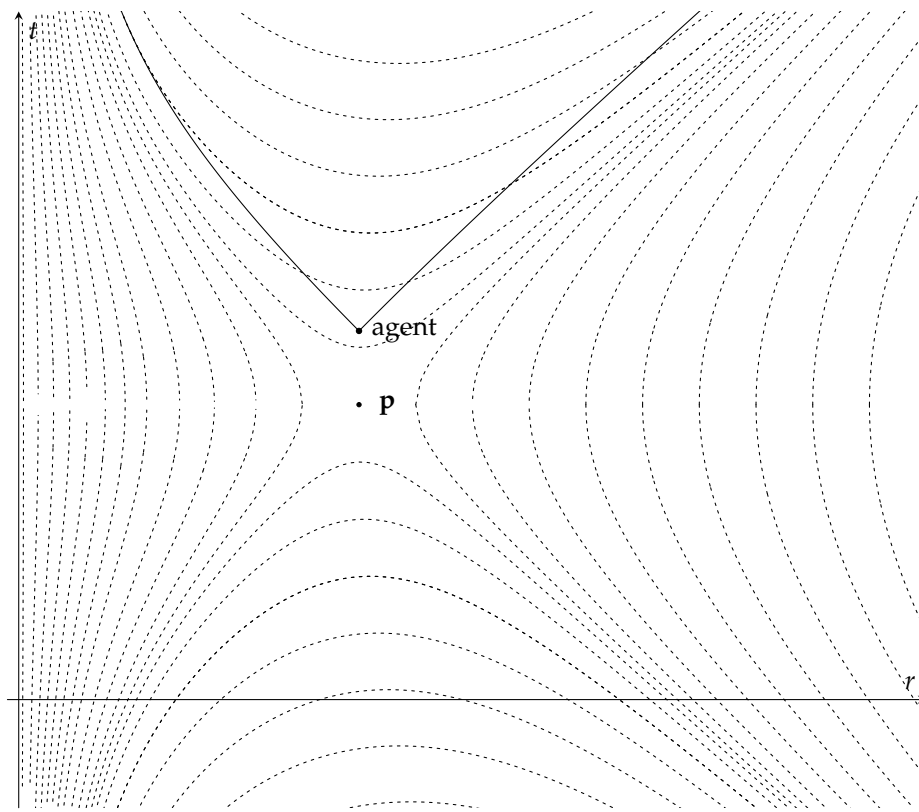


Figure 4.12: A sequence of regions each comprised of points within some fixed g' of $\mathbf{p}$.

Much as we did with other regions in flat spacetime, we can use this sequence of regions to apply RSE in curved spacetime. The sequence of expanding regions generated by distance measure g' is very similar to those generated by s , just skewed a bit by gravitational curvature—they bend towards $t = \pm\infty$ as r approaches 0 and we get closer to the massive body. And when we place a lightcone starting just a little later than the central point $\mathbf{p}$, as we can see here, only a finite portion of that lightcone will lie within each region (since the lightcone skews just as the regions do). So, with the right choice of $\mathbf{p}$, these regions will each only contain finite value, as we need them to. On top of that, these regions will be constructed the same way by all observers, regardless of their velocity, since g' is a distance measure independent of the observer's velocity (as it measures proper time and proper distance). And so, by simply switching our distance metric from $|s^2|$ to g' , it seems that we can easily modify RSE to work in curved spacetime.

But will this always work? Sure, we will always have a distance metric g' on which all observers can agree. But will we *always* be able to choose a point $\mathbf{p}$ such that only a finite volume

of each region will fall in the agent's future lightcone?

Yes, as it turns out. The Schwarzschild spacetime above approximates spacetime in the vicinity of just about any massive body. And a universe containing *many* such bodies would simply have the warping of spacetime that occurs around such a body (illustrated on the left side of Figure 4.17) happen at lots of different positions. But that warping didn't stop us from having only a finite volume of each region overlap with the agent's future lightcone; nor would having that same warping occur elsewhere stop us. And other realistic spacetimes appear not to stop us either. Whether we face a spacetime under which the universe is expanding or contracting (e.g., a Friedmann-Lemaître-Robertson-Walker spacetime), or in which massive bodies are rotating (as approximated by, e.g., Kerr spacetime) or have charge (as approximated by, e.g., a Reissner-Nordström spacetime), or even in which entropy is so great that curvature is approximately constant (e.g., a de Sitter or anti-de Sitter spacetime), we will still have only a finite overlap between those regions.¹¹ No matter which of the many plausible spacetimes our future might resemble, it seems that (a slightly modified form of) RSE will still be able to deliver comparisons.

Before I wrap up, there is a further problem that arises for aggregation under general relativity. Recall the method I suggested in Section 2 for identifying the 'same' (or counterpart) spacetime points across worlds: first we identify all of the events outside the agent's future lightcone with the same points in each outcome, since these events will be fixed no matter what the agent does; then we map each of the points *within* the future lightcone to whichever points have matching distances in space and time from all of those points outside the lightcone. But, from the last four sections, it is abundantly clear that this relation won't work—there are no absolute distances between points, in space and in time, that do not differ with the velocity of the observer. So we must reformulate the relation: instead, perhaps, we replace "distances in space and time" with "distances according to the spacetime metric s (or, in curved spacetime, g')"; we map each future point x to whichever of the points in other worlds have matching distance $|s(x, p)|$ from every point p outside the future lightcone (or, in curved spacetime, whichever points have matching distance g' between them). All observers can agree on these measures of distance, so all observers can agree on which points are the same (or counterparts) under this relation.

¹¹A full discussion of each of these is beyond the scope of this thesis, but I would refer interested readers to Carroll (2004) for technical descriptions of each of these forms of spacetime.

But this relation still has trouble with general relativity. Curved spacetime works very differently from Euclidean space, and differently even from flat spacetime—we may have a point x that is such and such distance (given by g') from each of several other points, but if we changed the curvature of spacetime then it might be that *multiple* points are such and such distance from each of those several other points, or it might be that *no* points are! So it is conceivable that the above counterpart / transworld identity relation will sometimes say that a point x in the agent's future lightcone in one world is the same as (or counterpart to) *multiple* points in another world, or *no* points. Thus, whenever we compare worlds in which the agent's action produces *just* the right changes in how matter and radiation are distributed, and so how spacetime is curved, our relation for saying which points are which breaks down.

This may seem problematic for RSE, but it isn't really. Which spacetime points are the same and which aren't enters into the definition of RSE in two places. The first: we need to be able to start our expansions at the same point p in each world. But we can do this even with our wonky relation—simply pick a point p outside the agent's future lightcone, which will always have a unique identity/counterpart across the worlds under comparison. The second: for each point x , we want to identify the same point (or its counterpart) across worlds to take the difference in value at that point, which we will then sum up over all such points within each region. But nothing much hangs on whether some value lies at x or at another point in the region; we really just need the (difference in) value across the whole region. So we can avoid any problem here by simply assigning $V_a(x) = 0$ (on whichever cardinal representation we are using for local values) whenever x is absent from world W_a ; and, when x has multiple identities/counterparts in W_a , simply let V_a be the sum of their values. This allows us to sum the differences in value across each region without needing to worry whether each point has a unique identity/counterpoint in every world. RSE works just fine despite this.

4.6 Summary

In this chapter, we saw a potentially devastating problem for position-dependent theories of infinite aggregation: thanks to special relativity, many such theories either cannot compare any outcomes at all, or else they must allow their comparisons to be dependent on the velocity of

whoever is doing the comparing.

To compare outcomes, under most such theories, we need standard Euclidean distance over four dimensions to be absolute and observer-independent. But it isn't. So such theories cannot make any judgements at all. Even on a charitable reading, by which we let the distance measure depend on the velocity of the observer, this approach leads to an absurd form of moral relativism—the correct judgement may differ if you just travel fast enough.

But as demonstrated above, this problem can be overcome. With careful consideration of the geometry of spacetime, we can modify spatiotemporal expansionism to deliver judgements, and plausible ones at that, even in a relativistic universe.

Problems with risk

5.1 Introduction

The previous three chapters have been focused on determining the correct method of comparing outcomes. But we need more than this if we are to provide practical moral judgements.

In practice, human agents like us are *never* certain of how our available actions would turn out. Particularly over an infinitely long future, we can never know with certainty what would happen at a given time in the distant future. From our limited perspective, each action is not associated with a single outcome but instead with a risky *lottery* over outcomes. So, even when we face decisions in which the only relevant considerations are those of moral betterness, making practical judgements requires more than just being able to judge outcomes as better than one another. We must also be able to judge *lotteries over outcomes* as better than one another.¹

For those who hold minimally aggregative views, the standard method for extending moral comparisons of outcomes to comparisons of lotteries is *expected value theory*: we evaluate the *expected* total value produced in each lottery, and one lottery is at least as good as another if and only if its expected value is at least as great (Jackson, 1991).² But, for infinite worlds, we often don't have finite total values over which we can take expectations. And by spatiotemporal ex-

¹Put differently, even though we may have a solution to the *axiological* question of how to compare worlds, we do not immediately have an answer to the *instrumental axiological* question of how to compare lotteries, or to the *subjective normative* questions of which risky actions are permissible and which are forbidden.

²Note that expected *value* is distinct from the frequently-used notion of expected *utility*, and expected value theory distinct from expected *utility* theory, a widely accepted view in normative decision theory more broadly. Under expected utility theory, *utility* is given by some (indeed, *any*) increasing function of value—perhaps a concave function, such that additional value contributes less and less additional utility. The utility of an outcome may even be bounded, such that arbitrarily large amounts of additional value contribute arbitrarily little additional utility. Where expected value theory says that a lottery is better the higher its expected *value*, expected utility theory says that it is better the higher its expected *utility*. Expected value theory can be considered a special case in which that utility function is linear.

pansionism (along with every other position-dependent view mentioned so far), we don't obtain any cardinal measure of worlds' value so, even then, we cannot construct expected values. So it may seem that such views cannot be extended to compare lotteries, and that they will be useless in practice.

In this chapter, I propose one method of extending such views to compare lotteries: what I will call the 'Expectations of Differences' approach. This method allows us to adhere to the spirit of expected value theory without explicitly assigning expected values to lotteries (the traditional, 'differences of expectations' approach).³ But, in doing so, we must satisfy certain crucial desiderata for any plausible ranking of lotteries. This is surprisingly difficult to do, as I demonstrate in §5.4. And the one solution proposed so far, by Bostrom (2011), Arntzenius (2014), and Meacham (2020), does not satisfy them all. Fortunately, the Expectations of Differences approach does satisfy them.

5.2 Comparing worlds: Additivity

As it happens, the method of comparing lotteries that I will propose below is not only compatible with spatiotemporal expansionism. It can be applied in conjunction with any plausible betterness relation over worlds. So even if you might disagree with the expansionist principles I introduced in earlier chapters, what follows should still be of interest.

Specifically, my proposal below will be compatible with any betterness relation over worlds that satisfies *Additivity*. What's Additivity? It's an extremely weak condition of betterness, which says the following. Suppose two worlds contain the same locations. Some locations might obtain more value in one world, some more in the other. Sum up all of those increases from one world to the other, minus all of the decreases. If that total sum of all differences is positive, then one world is better.

For instance, compare worlds W_1 and W_5 , which share the locations l_a, l_b , and so on. (It does not matter what we take as the relevant type of locations - they might be persons, for instance.)⁴

³The difference between the two approaches closely resembles a similar difference between expected value theory and 'relative expectation theory' (Colyvan, 2008; Colyvan and Hájek, 2016).

⁴Note that Additivity as I've put it may consist of any of a number of different principles—one for each type of location you might take as relevant. As I've put it, if a betterness relation satisfies Additivity for *any* type of location, then it satisfies Additivity.

The only differences between worlds are that location l_a does better in W_1 , and that locations l_b to l_f do better in W_5 . Perhaps this comparison represents a decision in which we can either rescue one person from death, or rescue five others (with all else equal).

	l_a	l_b	l_c	l_d	l_e	l_f	l_g	l_h	l_i	l_j	$\dots$
W_1 :	1	0	0	0	0	0	1	1	1	1	$\dots$
W_5 :	0	1	1	1	1	1	1	1	1	1	$\dots$

We cannot say much about the total value of either world, but we can say something about the subtotal over *any finite subset* of locations. For instance, take the set of all locations which differ between worlds: l_a through l_f . The subtotal over these is 1 for W_1 and 5 for W_5 . The worlds have equal values at all other locations, so let us ignore those. And, left with only those first few locations, the difference between W_5 and W_1 is $5 - 1 = 4$. Thus we might claim that W_5 is the better world.

This reasoning matches the rule of Additivity. Put precisely:

*Additivity*⁵: For any worlds W_1 and W_2 with the same locations $\mathcal{L}$, $W_1 \succcurlyeq W_2$ if

$$\sum_{l \in \mathcal{L}} (V_1(l) - V_2(l)) \geq 0$$

(either by converging unconditionally, or by diverging unconditionally to $+\infty$).

Note that Additivity is not biconditional—some worlds may be better than others even though Additivity doesn't hold. After all, Additivity by itself does not give us a complete betterness $\succcurlyeq$ relation. Nor does it give a $\succcurlyeq$ relation that compares even a significant fraction of the pairs of worlds our actions might produce in practice. It is just a minimal condition, and one that is satisfied by almost every proposal in the literature.

It is also well-justified—a $\succcurlyeq$ relation that violated it would be a strange betterness relation indeed. As (Lauwers and Vallentyne, 2004, p. 39) show, any transitive $\succcurlyeq$ which satisfies all three of the following highly plausible principles will also satisfy Additivity.

Pareto Over Any Locations: For any worlds W_1 and W_2 with the same locations $\mathcal{L}$, if

⁵This principle is presented and defended by Vallentyne and Kagan (1997, p. 11), Lauwers and Vallentyne (2004, p. 21), and Basu and Mitra (2007).

for all $l \in \mathcal{L}$, $V_1(l) \geq V_2(l)$, then $W_1 \succcurlyeq W_2$.

If, as well, some $l_i \in \mathcal{L}$ has $V_1(l_i) > V_2(l_i)$, then $W_1 \succ W_2$.

This Pareto principle is weaker than either the Pareto Over Persons or Pareto Over Positions principles introduced in Chapter 2. It will hold if *either* of those two principles are satisfied—Pareto need only hold for *some* type of location, not a specific type of location. And it is overwhelmingly plausible that Pareto holds for at least some type of location.

Separability of Value: If W_1 and W_2 contain the same locations and $W_1 \succcurlyeq W_2$ then, adding their corresponding local values, $W_1 + W \succcurlyeq W_2 + W$ for all $W \in \mathcal{W}$ with the same locations.⁶

This principle, also called *Translation Scale Invariance* in the literature, ensures that betterness is sensitive *only* to differences in local value. If between one pair of worlds there is the same pattern of differences as between another pair of worlds, we must rank both pairs the same way. It does not matter what local values we start with in a world; all that matters is what we add or take away; if a certain combination of additions and removals is an improvement to a world, it would count as an improvement to any world. This is a key distinguishing feature of aggregative theories in the finite context (such as any which endorse totalism or total prioritarianism). And, moving to the infinite context, Separability of Value seems a necessary condition for staying true to the spirit of aggregation.

And the final condition we need, *Finite Sum*, requires only that $\succcurlyeq$ remains consistent with what our judgements would be if the universe could contain only finite total value. I take this as a crucial requirement for extensional adequacy.

Finite Sum: If there is a finite total sum of local values in both W_1 and W_2 , and the sum in W_1 is at least as great as that in W_2 , then $W_1 \succcurlyeq W_2$.

Each of these principles (and their conjunction) is hard to deny. And so Additivity is hard to deny. Given this, it is unsurprising that every one of these principles—and hence Additivity too—is satisfied by each of the plausible proposals in the literature (e.g. Vallentyne and Kagan, 1997; Bostrom, 2011; Arntzenius, 2014; Jonsson and Voorneveld, 2018, and my proposal in Chapter 2).

⁶Define addition of worlds as follows. For all worlds W_1 and W_2 with the same locations $\mathcal{L}$, the world $W_3 = W_1 + W_2$ is given by $V_3(l) = V_1(l) + V_2(l)$ for all $l \in \mathcal{L}$. Later, I'll define subtraction similarly.

Given this broad agreement, I will assume in what follows that Additivity holds for $\succsim$. But I still remain agnostic as to how we should strengthen it, so my conclusions will still be relevant for all of those stronger views.

5.3 Comparing lotteries: Local expectations

We don't just want to compare worlds; we want to compare *lotteries* over worlds.

Formally, a lottery L is represented as a probability measure on $\mathcal{W}$ —it maps all sets of possible worlds to probabilities in the interval $[0, 1]$, while obeying the standard probability axioms. The set of all such probability measures on $\mathcal{W}$ (the set of all possible worlds) is denoted by $\mathcal{P}$. We can also define the *domain* of each lottery L_i by $\mathcal{W}_i = \{W \in \mathcal{W} | L_i(\{W\}) > 0\}$. For two lotteries L_i and L_j , the union of their domains $\mathcal{W}_i \cup \mathcal{W}_j$ can be abbreviated to $\mathcal{W}_{(i,j)}$.

I will make a few other abbreviations to keep the notation in check. $L(\{W\})$ will be abbreviated to $L(W)$ when the input to L is a set $\{W\}$ with only one element. And $L_i(\succsim W)$ will be used as an abbreviation for the probability that L_i gives to the set of all worlds in $\mathcal{W}_i$ that are *at least as good as* W (or, equivalently, $\{W' \in \mathcal{W}_i | W' \succsim W\}$). And, last, when a lottery has only one world W in its domain, mapped to probability 1, I'll use W to denote both world and lottery.

To compare such lotteries, we need another 'at least as good as' relation: a binary relation $\succsim_L$, this time on $\mathcal{P}$. As before, strict betterness ($\succ_L$) and equality ($\simeq_L$) are represented as the asymmetric and symmetric components, respectively. And again, as basic desiderata, I will assume that $\succsim_L$ must be reflexive and transitive. It must also be consistent with $\succsim$: if $W_1 \succsim W_2$ then $W_1 \succsim_L W_2$ too.

With this notation in hand, how might we compare lotteries in a plausible way? In the remainder of this section, I'll describe one proposal, made by both Arntzenius (2014) and Bostrom (2011) and also endorsed by Meacham (2020).

Consider the following two lotteries: W_1 , which delivers world W_1 for sure; and L , which brings even odds of W_2 or W_3 , as specified below.

	l_a	l_b	l_c	l_d	l_e	l_f	l_g	l_h	l_i	l_j	l_k	$\dots$
W_1 :	1	0	1	0	1	0	1	0	1	0	1	$\dots$

$$L \left\{ \begin{array}{l|cccccccccccccc} L(W) & & l_a & l_b & l_c & l_d & l_e & l_f & l_g & l_h & l_i & l_j & l_k & \cdots \\ 1/2 & W_2 : & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \cdots \\ 1/2 & W_3 : & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \end{array} \right.$$

Note that the labelling of locations here may be arbitrary: it need not represent any special structure in our locations (e.g., their position in time). So we can say that W_1 contains infinitely many locations with local value 1, and infinitely many with value 0, but we cannot say anything else about which of those sets of locations is larger.

If Additivity holds, then $W_2 \succ W_1 \succ W_3$. But it says nothing about W_1 versus the risky bet L . Using only Additivity, we cannot straightforwardly construct expected values: the differences between W_1 and each of the others have infinite sums, so there's no clear way to assign a real value to each option.

But here is one way we might still compare the two. In lottery L , each location l has its own prospects: probability $\frac{1}{2}$ of 2, and probability $\frac{1}{2}$ of 0. So we can list the expected *local* value for each l under lottery L .

$$\begin{array}{cccccccccccccc} l_a & l_b & l_c & l_d & l_e & l_f & l_g & l_h & l_i & l_j & l_k & \cdots \\ \mathbb{E}_L(V(l_i)) : & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \end{array}$$

This defines a new object with the same structure as a world—an 'expected world', with the same set of locations and with a value function given by the *expected* local values under the lottery. Equivalently, $\mathbb{E}W_L = \langle \mathcal{L}, \mathbb{E}_L(V) \rangle$.

This 'expected world' can easily be compared to W_1 , via Additivity (indeed, by Pareto alone). And it's strictly better. So we might say that $L \succ_L W_1$.

In effect, this matches the proposals made by Arntzenius⁷, Bostrom⁸, and Meacham. To an approximation, they all endorse *Local Expectations*. (See notes for how they differ.)

Local Expectations: For any lotteries L_1 and L_2 such that all $W \in \mathcal{W}_{(1,2)}$ contain the same locations $\mathcal{L}$, $L_1 \succ_L L_2$ if $\mathbb{E}W_{L_1} \succ \mathbb{E}W_{L_2}$.

If both Local Expectations and Additivity hold, then we can make the comparison between the above lotteries W_1 and L , as demonstrated. And, in general, we then have the result that $L_1 \succeq_L L_2$ if the following sum is (unconditionally) greater than or equal to 0 (or diverges unconditionally to $+\infty$).

$$\sum_{l \in \mathcal{L}} (\mathbb{E}_{L_1}(V(l)) - \mathbb{E}_{L_2}(V(l)))$$

This seems promising: it appears we can sidestep the whole problem of taking expectations over infinite totals. We just need to lower our expectations, down to the local level.

⁷Arntzenius (2014, p. 55-6) proposes the following.

Weak Location Criterion: For lotteries A and B , $A \succ B$ iff $\sum_{l \in \mathcal{L}} (\mathbb{E}_A(V(l)) - \mathbb{E}_B(V(l)))$ "...is absolutely convergent and > 0 , where we are summing over all (epistemically possible) [locations]..." $l \in \mathcal{L}$.

This is *almost* equivalent to the conjunction of Local Expectations and Additivity. Except: 1) it defines only a strict betterness relation $\succ_L$, so does not imply that $L_1 \simeq_L L_2$ for any lotteries; 2) it remains silent if the sum diverges unconditionally to $+\infty$, even in cases of certainty in which Additivity gives a verdict. And 3) it allows worlds in both lotteries to contain different locations (e.g., the same people) summing value instead over all epistemically possible locations, which requires that we assign some value to the non-existent lives. I don't want to take a stand on such cases here, and my modifications to (1) and (2) should be uncontroversial.

⁸Bostrom's (2011:20-4) proposal is more complicated. He describes (but doesn't explicitly endorse) an approach by which we represent the total value of each world with a *hyperreal number*: a vector of (countably) infinite length which consists of the cumulative sums of local values, summed in some common order. In the example above, if we chose to sum in the order l_a, l_b, l_c , etc, then we would represent the total value of W_1 by the hyperreal $(1, 1, 2, 2, 3, 3, \dots)$ and of W_2 by $(2, 4, 6, 8, 10, \dots)$. And we can say that W_2 is better: if one hyperreal has entries larger than another at 'sufficiently many' positions, then it is the larger number. W_2 has larger entries than W_1 at *all* positions, so it has the larger total.

The standard of 'sufficiently many' can vary here but one very minimal condition is that, if the entries of one hyperreal are greater than another in all but finitely many positions, then that the former is the larger hyperreal. In effect, this condition is equivalent to Additivity. (There are stronger conditions we might apply too, but those can be chosen to make the hyperreal approach equivalent to any plausible aggregation rule we want—see Pivato (2014).)

We can also sum hyperreals and multiply them by real probabilities just as we can with vectors, and so they can give us expected values in roughly the old-fashioned way. L would then have expected value $(1, 2, 3, 4, 5, 6, \dots)$, which is identical to the hyperreal total value of $\mathbb{E}W_L$ from above. Since this is larger than the total value of W_1 , we could say that $L \succ_L W_1$. In effect, this is equivalent to applying Local Expectations as above, but under a different formalism.

5.4 The problem

But there's a problem with Local Expectations. Consider the example of *Egregious Energy*.

Example: Egregious Energy

A new energy source has been discovered, and you must decide whether humanity takes advantage of it.

If we do use it, once we get it working we will obtain enormous amounts of energy and many lives will be improved. But there are a few downsides. One is that the fuel needed is limited—we will only reap its benefits for a short time. Another is that it may take a while to get it working but, the longer it takes, the longer that cornucopia of energy will last. But the greatest problem is pollution—this energy source produces a novel form of pollution which will badly harm human health. That pollution and its effects will decrease over time, but we will never be able to eradicate it. If we use this source of energy, that pollution will continue to harm human well-being for the entire future of humanity, which I'll assume will be infinitely long. In short, using this energy source will produce some finite benefit, but also cause an infinite total amount of harm. (The relevant probabilities and values are given below.)

And, for simplicity, exactly the same persons (and person-time-slices, and other possible types of locations) will exist at exactly the same physical positions whether or not we adopt this new energy source.

Since the same persons exist at the same physical positions either way, we can treat the resulting worlds as having the same locations. Then we can represent your options as W_0 and the lottery L . If you forego the energy source, you produce W_0 , a world with constant cardinal value 0 at all locations, simply representing the baseline of what would have happened otherwise.⁹ And if you choose to have humanity adopt it, you produce L , a lottery over infinitely many worlds

⁹Note that local values of 0 here do not imply that the lives in question are right on the boundary of not worth living. They might be extremely valuable lives. But these local values are here represented cardinally—the numbers only capture the relative size of the differences between them. So these same representations of W_0 and L could describe worlds in which everyone has a blissful life and suffers only a slight reduction in quality of life due to the pollution. Or they could describe worlds in which everyone suffers terribly and that pollution makes life even worse.

$\{W_1, W_2, \dots, W_j, \dots\}$ (each with probability $\frac{1}{2^j}$). You're uncertain of how long it takes the energy source to start working, during which time everyone just obtains the same value 0 as in W_0 . Once it's working, some number of people obtain some greater value, represented by 2. And then, once the fuel runs out, every person obtains less value than the baseline, represented by some negative number. Note that, in every single one of these worlds in L , the total value diverges unconditionally to $-\infty$. So they're all worse than W_0 —*infinitely* worse, in fact.¹⁰

		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	$\cdots$	
	$W_0:$	0	0	0	0	0	0	0	0	$\cdots$	
L	$L(W)$		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	$\cdots$
	$1/2$	$W_1:$	2	$-1/2$	$-1/2$	$-1/4$	$-1/4$	$-1/4$	$-1/4$	$-1/8$	$\cdots$
	$1/4$	$W_2:$	0	2	2	$-1/4$	$-1/4$	$-1/4$	$-1/4$	$-1/8$	$\cdots$
	$1/8$	$W_3:$	0	0	0	2	2	2	2	$-1/8$	$\cdots$
	$1/16$	$W_4:$	0	0	0	0	0	0	0	2	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

We can use Local Expectations to compare L to W_0 . As above, we take L 's expected value for each location. We obtain $\mathbb{E}W_L$ as below.

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	$\dots$
$\mathbb{E}W_L :$	1	$1/4$	$1/4$	$1/16$	$1/16$	$1/16$	$1/16$	$1/64$	$\dots$

The expected local values are all greater than 0, so Additivity (or just Pareto alone) implies that $\mathbb{E}W_L \succ W_0$. And, together with Local Expectations, that implies that $L \succ_L W_0$. It is allegedly better to adopt the energy source, even though that is guaranteed to leave infinitely many people worse off.

But this is implausible. *Every single world* in L is worse than W_0 (by Additivity). So L guaran-

¹⁰In each world W_i , the local value at l_j is given by:

$$V_i(l_j) = \begin{cases} 0 & \text{for } i < 2^j \\ 2 & \text{for } 2^i \leq j < 2^{i+1} \\ -\frac{1}{2^k} & \text{for } 2^k \leq j < 2^{k+1}, \forall k > i \end{cases}$$

Note also that the sequence of locations $(l_1, l_2, \dots)$ may be chronological, but that need not have any moral significance.

tees us a worse outcome. Yet Local Expectations still says that it is the better choice. This means that Local Expectations clashes with *Guaranteed Betterness*.

Guaranteed Betterness: If *every* world in $\mathcal{W}_1$ (the domain of L_1) is better than *every* world in $\mathcal{W}_2$, then $L_1 \succ_L L_2$.

This conflict can be stated more formally as Theorem 1.

Theorem 1: For any reflexive, transitive relation $\succ_L$ on $\mathcal{P}$, if $\succ_L$ satisfies Local Expectations then it cannot satisfy both *Guaranteed Betterness* and Additivity.

To prove this would be straightforward. We would need only compare W_0 and L from Egregious Energy. *Any* relation $\succ_L$ on the set of all lotteries which satisfies both *Guaranteed Betterness* and Additivity must give the verdict that W_0 is better. And any such relation which satisfies Local Expectations and Additivity (or just Pareto) must disagree: it will say L is better. So such relations can *only* satisfy *Guaranteed Betterness* and Additivity if they do not satisfy Local Expectations.

Is this so bad? Yes. Although Local Expectations seems plausible, *Guaranteed Betterness* seems *undeniable*. Why? One of the most basic claims of decision theory is: if we must decide between two outcomes of which one is strictly better, and there is no risk involved, *of course* we ought to choose the better one. Likewise for minimally aggregative views in ethics—in the ethical cases we’re interested in, we ought to choose the better outcome. Intuitively, this shouldn’t change if we add a little uncertainty to the outcomes of each, but little enough that *every* outcome of one option L_2 is still worse than every outcome of the other, L_1 . No matter the result of L_1 or of L_2 , no matter the state of the world, no matter if we swap those states around, no matter how lucky or unlucky we are, we are *certain* that L_2 will turn out *worse*. For goodness’ sake, we had better not choose it!

Further, the motivation of expected value reasoning (at least in contexts in which it is appropriate to use) is to help us obtain what we ultimately care about. As Schoenfield (2014, p. 268) forcefully points out in another context, we ultimately care about obtaining outcomes that are valuable, not outcomes that are merely *expectedly valuable*. We should only be interested “...in *expected* value insofar as it helps [us] obtain what is actually important: value. ... Any theory of expected value that makes demands that don’t make sense given our concern with *value* can’t do what expected value theory is meant to do.” (*ibid.*) With the goal of promoting value, a demand

that we abandon Guaranteed Betterness make no sense all.

In my view, this goal of promoting value also places a stronger requirement on our rules for comparing lotteries. Suppose that one lottery L_1 has at least as great a probability of turning out at least as good as some world W as some other lottery L_2 . And suppose that this holds for *all* worlds W —for *any* world, L_1 has as high or higher probability of resulting in W or something even better. If our goal is to promote value, surely L_1 must be at least as good a lottery as L_2 . So says *Stochastic Dominance*.

Stochastic Dominance: Let L_1 or L_2 be any lotteries on $\mathcal{W}$ with $\mathcal{W}_{(1,2)}$ totally ordered

by $\succsim$. If $L_1(\succsim W) \geq L_2(\succsim W)$ for all $W \in \mathcal{W}_{(1,2)}$, then $L_1 \succsim_L L_2$.

If, as well, $L_1(\succsim W) > L_2(\succsim W)$ for some $W \in \mathcal{W}$, then $L_1 \succ_L L_2$.

This is a strengthening of Guaranteed Betterness. But it is not a radical one. When our decision theory deals with finite payoffs, both principles are consistent with, but weaker than, standard expected value theory—for instance, they do not rule out risk aversion (see Buchak, 2013). Accept expected value theory and you must accept Stochastic Dominance. But, while expected value theory is somewhat controversial in the existing literature, Stochastic Dominance and its kin are not. Only in rare circumstances have philosophers proposed normative decision theories that violate this form of Stochastic Dominance.¹¹

But Local Expectations violates Stochastic Dominance, along with its weaker siblings. So we must reject it. But how then can we compare lotteries over infinite worlds?

5.5 My proposal: Expectations of Differences

The challenge is to provide a rule for comparing lotteries over infinite worlds which avoids the above problems, and which satisfies Stochastic Dominance. This challenge is greater in some cases than in others. I'll start by giving a weak version of my rule which can handle easy cases. Then, I'll give a stronger version which can handle harder cases too.

¹¹Examples include Seidenfeld et al. (2009); Smith (2014) and Lauwers and Vallentyne (2016). Schoenfeld (2014) rejects a similar, but distinct, form of Stochastic Dominance which gives verdicts even when outcomes in the domains aren't totally ordered by the betterness relation. She raises no objection to weaker formulations like mine.

5.5.1 Version 1

We start with a class of easy cases: comparing lotteries in which each pair of worlds differ by at most a finite sum of local differences. L_1 and L_2 are two such lotteries.

$$L_1 \left\{ \begin{array}{l|l} L_1(W) & l_a \quad l_b \quad l_c \quad l_d \quad l_e \quad l_f \quad l_g \quad l_h \quad \dots \\ 1/2 & W_0 : 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad \dots \\ 1/2 & W_6 : 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad \dots \end{array} \right.$$

$$L_2 \left\{ \begin{array}{l|l} L_2(W) & l_a \quad l_b \quad l_c \quad l_d \quad l_e \quad l_f \quad l_g \quad l_h \quad \dots \\ 1/2 & W_2 : 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad \dots \\ 1/2 & W_3 : 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad \dots \end{array} \right.$$

Here's how we might proceed. First, with just Additivity, we can rank the worlds $W_6 \succ W_3 \succ W_2 \succ W_0$.

Next, we might pick a world at random—say, W_0 —and represent these worlds by their sum of differences with that baseline world. For W_0 , that's 0. For W_6 , it's 6. For W_2 , it's 2. And for W_3 , it's 3. So we have a nice finite, cardinal 'total' to represent each world. And that gives us all we need to represent each lottery with an expected 'total' value.

$$L_1(W_0) \cdot 0 + L_1(W_6) \cdot 6 = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 6 = 3$$

$$L_2(W_2) \cdot 2 + L_2(W_3) \cdot 3 = \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 3 = 2\frac{1}{2}$$

Calculated this way, the expected 'total' of L_2 is less than that of L_1 , so we might claim that L_1 is the better lottery. Putting this approach more precisely, we have *Expectations of Differences 1*.

Expectations of Differences 1 (ED1): Let L_1 and L_2 be any lotteries on $\mathcal{W}$ for which all worlds in $\mathcal{W}_{(1,2)}$ contain the same locations $\mathcal{L}$. If there exists a world $W_* \in \mathcal{W}_{(1,2)}$ such that the sum

$$\sum_{W_i \in \mathcal{W}_1} L_1(W_i) \left(\sum_{l \in \mathcal{L}} (V_i(l) - V_*(l)) \right)$$

is at least as great as the corresponding sum for L_2 , then $L_1 \succsim_L L_2$.

ED1 resembles the way we calculate finite expectations, but with crucial tweaks. In the finite

setting, we take the expectation of the *total* value in each world and see how they differ. Here, instead, we represent the value of each world by the sum of all its differences from some ‘baseline’ world W_* . And then, for each lottery, we take the expectation of that sum.

The standard approach compares expected totals to see which is larger—effectively, it relies on the *differences of expectations*. My approach is to instead use the *expectations of differences*. If we wanted, we could do this in finite cases. We would reach the same verdict as we get with expected total values—in finite cases, the approaches are equivalent. But in infinite cases like this, they come apart, and one of the two clearly does better.

But this approach won’t always succeed. For some pairs of lotteries, the expectation of differences will not exist for *any* W_* . Then my rule often falls silent. But it won’t *always* fall silent in such cases. Consider a further class of easy cases: when one of those sums diverges unconditionally to (positive or negative) infinity. As long as *only* one of them does, then we can still say which lottery is better. For instance, we can do so in the case of Egregious Energy. Recall W_0 and L .

		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	$\dots$
	$W_0 :$	0	0	0	0	0	0	0	0	$\dots$
$L \left\{ \begin{array}{l} L(W) \\ 1/2 \\ 1/4 \\ 1/8 \\ 1/16 \\ \vdots \end{array} \right.$		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	$\dots$
	$W_1 :$	2	$-1/2$	$-1/2$	$-1/4$	$-1/4$	$-1/4$	$-1/4$	$-1/8$	$\dots$
	$W_2 :$	0	2	2	$-1/4$	$-1/4$	$-1/4$	$-1/4$	$-1/8$	$\dots$
	$W_3 :$	0	0	0	2	2	2	2	$-1/8$	$\dots$
	$W_4 :$	0	0	0	0	0	0	0	2	$\dots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Unlike the Arntzenius-Bostrom-Meacham approach, Expectations of Differences implies that W_0 is better than L . How? First, let our ‘baseline world’ be W_0 . Then the sum of differences between W_0 and W_0 is simply 0. Meanwhile, between each world in the domain of L and W_0 , the

sum diverges unconditionally to $-\infty$.¹²

We have unconditional divergence for one lottery (L) but not the other (W_0), so we face the sort of situation mentioned above. And ED1 can still say, thankfully, that $W_0 \succcurlyeq L$. After all, $-\infty$ is a lot less than 0. So we seem to be doing better than the Arntzenius-Bostrom-Meacham approach.

ED1 has other arguments in its favour too. Intuitively, it seems a plausible and natural way to judge lotteries. It's a rough analogue of Additivity for this new setting of comparing lotteries. In fact, in cases of certainty, it *implies* Additivity. And, like Additivity, the most compelling reason to accept it is that it is entailed by some highly plausible conditions. (All proofs appear in an appendix at the end of this chapter.)

Theorem 2: For any reflexive, transitive relation $\succcurlyeq_L$ on $\mathcal{P}$, if $\succcurlyeq_L$ satisfies Stochastic Dominance, Finite Expectations, and Separability of Value (for Lotteries) then it satisfies ED1.

Stochastic Dominance will be familiar from the previous section. Recall that it is consistent not just with standard expected value theory but also with risk aversion. To impose risk neutrality, we can use *Finite Expectations*. This is the analogue of Finite Sum (from Section 3) for comparing lotteries. It requires only that $\succcurlyeq_L$ remains consistent with what our judgements would be if our lotteries contained only finite expected total value. If $\succcurlyeq_L$ doesn't satisfy this, then it's not an adequate extension of finite expected value theory.

Finite Expectations: Let L_1 and L_2 be any lotteries such that the expected total sum of local values converges unconditionally to finite values k_1 and k_2 , respectively.
 $L_1 \succcurlyeq_L L_2$ if and only if $k_1 \geq k_2$.

Then there's the highly plausible *Separability of Value (for Lotteries)*. As did Separability of

¹²The calculation is:

$$\begin{aligned}
 & \sum_{W_i \in \mathcal{W}} L(W_i) \left(\sum_{l \in \mathcal{L}} (V_i(l) - V_0(l)) \right) \\
 &= \sum_{W_i \in \mathcal{W}} \frac{1}{2^i} \left(\sum_{l \in \mathcal{L}} (V_i(l) - 0) \right) \\
 &= \sum_{W_i \in \mathcal{W}} \frac{1}{2^i} \left(0 + 0 + \dots + 2 \cdot 2^{i-1} + \sum_{k=1}^{\infty} -1 \right) \\
 &= -\infty
 \end{aligned}$$

Value above for comparisons of worlds, this says that we can take any two lotteries and add whatever string of local values we want to (all of the worlds in the domain of) them, and the resulting lotteries are ranked the same way. This implies that comparisons of lotteries are *only* sensitive to differences in the local values of the worlds in their domains, and that local values are additively separable. This is a direct analogue of Separability of Value for the lottery context, so much so that I'll abbreviate it to 'Separability of Value' in what follows.

Separability of Value (for Lotteries): For any lotteries L_1 and L_2 and world $W' \in \mathcal{W}$ with the same locations as all worlds in $\mathcal{W}_{(1,2)}$, let lottery L'_1 on $\mathcal{W}$ be defined as $L'_1(W) = L_1(W - W')$ for all $W \in \mathcal{W}_1 + W'$, and similarly for L'_2 . If $L_1 \succ_L L_2$ then $L'_1 \succ_L L'_2$.¹³

I find Stochastic Dominance, Finite Expectations, and Separability of Value, as well as their conjunction, hard to deny. So ED1 is on firm ground as a minimal principle for comparing lotteries, just as Additivity is for comparing worlds. And it seems to me to be on strictly *firmer* ground than Local Expectations, which implies Finite Expectations and Separability of Value but violates all three of the dominance principles we saw above.

5.5.2 Version 2

But that first principle doesn't get us far. Consider L_1 and L_2 .

$$L_1 \left\{ \begin{array}{l|l} L_1(W) & l_a \quad l_b \quad l_c \quad l_d \quad l_e \quad \dots \\ 0.1 & W_4 : 4 \quad 4 \quad 4 \quad 4 \quad 4 \quad \dots \\ 0.9 & W_1 : 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad \dots \end{array} \right.$$

$$L_2 \left\{ \begin{array}{l|l} L_2(W) & l_a \quad l_b \quad l_c \quad l_d \quad l_e \quad \dots \\ 0.2 & W_{2,1} : 2 \quad 1 \quad 2 \quad 1 \quad 2 \quad \dots \\ 0.8 & W_0 : 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad \dots \end{array} \right.$$

Take any pair of those worlds and sum their local differences; the result is infinite. If we apply ED1 here, no matter which baseline-world we pick, the expected sums are undefined. Using ED1 alone, we cannot compare these lotteries. Yet, intuitively, we can. So we need a stronger rule.

¹³This also implies the converse since, for all W' , the condition also applies to $-W'$.

Before I present that rule, consider how little we need to compare two lotteries over finite payoffs. We can make judgements without knowing the probabilities in each lottery—we just need to know the *differences* in probability of each outcome between one lottery and the other. And we can make judgements without knowing the values of the payoffs—we just need to know the (scale of the) differences between them. We can even make judgements often without knowing the *precise* differences in probabilities and values—upper and lower bounds may be enough.

Here are two lotteries over finite payoffs which we can compare with incomplete information.

L'_1 : value $v + u + 1$ with probability p ; value v with probability $1 - p$.

L'_2 : value $v + 1$ with probability $2p$; value 0 with probability $1 - 2p$.

Here, v and u are some real numbers such that $v > 0$ and $u > 2$. And p is some probability in $(0, \frac{1}{2})$. We cannot assign precise values to any of them, perhaps because those values are vague or indeterminate. But we can still judge which lottery is better. Take their expected values: L'_1 has expectation $v + p(u + 1)$, which is greater than $v + 3p$; and L'_2 has expectation $2p(v + 1)$, which is less than $v + 2p$. So L'_1 is better. Imprecision is no problem here.

We have similar information when comparing the above lotteries L_1 and L_2 over infinite worlds—it turns out that they are structurally equivalent to the finite lotteries L'_1 and L'_2 . So we can take a fairly similar approach.

First, take the difference in probabilities of obtaining a world at least as good as each W_i between L_1 and L_2 , given by $\Delta p_i = L_1(\succcurlyeq W_i) - L_2(\succcurlyeq W_i)$. Where this is positive, L_1 has a greater probability of delivering W_i or better; where it's negative, L_2 has the greater probability.

$$\begin{aligned}\Delta p_4 &= L_1(\succcurlyeq W_4) - L_2(\succcurlyeq W_4) = 0.1 \\ \Delta p_{2,1} &= L_1(\succcurlyeq W_{2,1}) - L_2(\succcurlyeq W_{2,1}) = -0.1 \\ \Delta p_1 &= L_1(\succcurlyeq W_1) - L_2(\succcurlyeq W_1) = 0.8 \\ \Delta p_0 &= L_1(\succcurlyeq W_0) - L_2(\succcurlyeq W_0) = 0\end{aligned}$$

Second, take the difference between each world and the next best world in the domain. If our worlds had finite total values, we would represent those differences with finite values. But, with infinite worlds, we must represent differences as worlds themselves—worlds given by the differences in local values. For each world W_i which has some distinct next best world W_j in

$\mathcal{W}_{(1,2)}$, the difference between them is given by $D_i = \langle \mathcal{L}, V_i - V_j \rangle$, with local values $V_i(l) - V_j(l)$ at every location. (Note that if there is no world in $\mathcal{W}_{(1,2)}$ that is strictly worse than W_i , then D_i is undefined.)

	l_a	l_b	l_c	l_d	l_e	$\dots$
$D_4 :$	2	3	2	3	2	$\dots$
$D_{2,1} :$	1	0	1	0	1	$\dots$
$D_1 :$	1	1	1	1	1	$\dots$

Third, compare how large those differences are relative to one another. Again, this would be straightforward with real, finite differences, but less so when the differences are infinite worlds. Fortunately, we *can* compare their size. To start with, we can use the same old $\succ$ relation that we would use to compare the size of worlds. By Additivity (or just Pareto), we'll have $D_4 \succ D_1 \succ D_{(2,1)}$.

And, further, we can judge their relative size by comparing them to scalar multiples of one another.¹⁴ A scalar multiple can be defined as $k \cdot W = \langle \mathcal{L}, k \times V_i \rangle$ for any real k , with local values each k times that in W . Then we have:

$$2 \cdot D_1 \prec D_4 \prec 3 \cdot D_1$$

$$0 \cdot D_1 \prec D_{2,1} \prec 1 \cdot D_1$$

So we can say that D_4 is more than twice, but less than three times, as great as D_1 . And $D_{2,1}$ is less than D_1 , but greater than 0 times D_1 . And that's all we need. The information we have here is analogous to what we had for L'_1 and L'_2 above, and the lotteries L_1 and L_2 are analogous to them as well. D_1 is analogous to v , D_4 analogous to u , and $D_{2,1}$ analogous to the value 1. We know that $D_4 \succ 2 \cdot D_{2,1}$, much like we knew that $u > 2$ before. And so, if these lotteries can be compared in an analogous way, we can say $L_1 \succ_L L_2$.

And we can make judgements much more generally, even for lotteries that do not resemble L'_1 and L'_2 specifically. To see why, note that, if we were dealing with *any* lotteries L_1 and L_2 over

¹⁴Since the local values in the original worlds W_i and W_j are represented on a common interval scale, the local values in D_i can be represented on a *ratio scale*—they have an absolute zero, 0. So it makes sense to compare their absolute size via scalar multiplication, and likewise for the difference-worlds at large.

finitely-valued outcomes $w \in \mathcal{W}_{(1,2)}$, we would say that $L_1 \succsim_L L_2$ if and only if L_1 had at least as great an expected value as L_2 :

$$\mathbb{E}(L_1) - \mathbb{E}(L_2) = \sum_{w \in \mathcal{W}_{(1,2)}} \left((L_1(w) - L_2(w)) \times V(w) \right) \geq 0$$

And this equation rearranges to the following, where the outcomes $(w_1, w_2, w_3, \dots)$ are ordered from least to most value. Here, we use the probability of each lottery turning out at least as good as each w_i , rather than the probability of w_i alone.

$$\mathbb{E}(L_1) - \mathbb{E}(L_2) = \sum_{w_i \in \mathcal{W}} \left((L_1(\succsim w_i) - L_2(\succsim w_i)) (V(w_i) - V(w_{i-1})) \right) \geq 0$$

That difference in probabilities between L_1 and L_2 is just Δp_i from above. And the difference in values is equivalent to D_i . If we can represent the D_i s with some real values k_i which measure their relative size, then we can use this equation as is. But we don't even need their precise relative size. We can get by with k_i s which capture an upper or lower bound on the relative size of each D_i .

In effect, we can apply *Expectations of Differences 2*, which goes like this.

Expectations of Differences 2 (ED2): Let L_1 and L_2 be any lotteries on $\mathcal{W}$ such that all $W \in \mathcal{W}_{(1,2)}$ have the same locations.

If, for some $W_* \in \mathcal{W}_{(1,2)}$, and for all $W_i \in \mathcal{W}_{(1,2)}$ such that $\Delta p_i \neq 0$, there exists $k_i \in \mathbb{R}$ such that

$$k_i \Delta p_i \cdot D_* \preccurlyeq \Delta p_i \cdot D_i \quad \text{and} \quad \sum_{W_i \in \mathcal{W}_{(1,2)}} k_i \Delta p_i \geq 0$$

then $L_1 \succsim_L L_2$.

Here, W_* is just some world in the domain, and D_* the difference between it and the next best world, against which the size of each D_i will be compared.

The first equation constrains the k_i that we can use as a stand-in for each D_i . It represents an upper or lower bound on how many times greater D_i is than D_* . To say that $L_1 \succsim_L L_2$, we need k_i to give an *upper* bound on the size of D_i when W_i (or better) is *more* likely under L_1 —

when $\Delta p_i > 0$. And we need k_i to give a *lower* bound when it's more likely under L_2 —when $\Delta p_i < 0$. That's why Δp_i appears on both sides of that equation—it allows the inequality to switch direction when $\Delta p_i < 0$.

Once we have those k_i s to represent the size of the differences, the second equation runs the same probability-weighted sum we saw above. Like above, when it's positive, L_1 is better. And that's ED2.

But should we accept ED2? I think so, and not just based on that analogy with imprecise values. For one, it implies all of the (very plausible) judgements of ED1, as long as $\succsim$ obeys Additivity.

Theorem 3: For any reflexive, transitive relation $\succsim_L$ on $\mathcal{P}$, if $\succsim_L$ satisfies Expectations of Differences 2 and Additivity, then it satisfies Expectations of Differences 1.

Given the compatibility of ED2 with ED1, we already know that this version of Expectations of Differences will deal with cases like Egregious Energy better than rival views. But it's also stronger than the previous version, as demonstrated in the example above.

Also in its favour, the rule is implied by the conjunction of several highly plausible principles.

Theorem 4: For any reflexive, transitive relation $\succsim_L$ on $\mathcal{P}$, if $\succsim_L$ satisfies Stochastic Dominance, Separability of Value (for Lotteries), Independence, and Extrapolated Expectations, then it satisfies Expectations of Differences 2.

You'll recognise Stochastic Dominance and Separability of Value from above. Then we have two newcomers.

First, *Independence* is the same principle that is often used to axiomatise expected utility theory. Suppose we evaluate L_1 is at least as good as L_2 . Independence says that we could mix each lottery with some third lottery L_3 , whatever it might be, and the resulting mixed lotteries would be ranked the same way. Equivalently, take a mixed lottery which gives you probability p of running lottery L_1 as normal, and a probability $1 - p$ of running L_3 instead; Independence says that that mixed lottery is still at least as good as the mixed lottery in which L_2 replaces L_1 . More formally, it can be stated as follows.

Independence: Define lotteries $L_{1\vee 3}, L_{2\vee 3}$ on $\mathcal{W}$ by

$$L_{1\vee 3}(W) = p \times L_1(W) + (1 - p) \times L_3(W) \quad \text{for all } W \in \mathcal{W}$$

$$\text{and } L_{2\vee 3}(W) = p \times L_2(W) + (1 - p) \times L_3(W) \quad \text{for all } W \in \mathcal{W}$$

For any lotteries L_1, L_2 , and L_3 on $\mathcal{W}$ and for any $p \in [0, 1]$, $L_{1\vee 3} \succsim_L L_{2\vee 3}$ if and only if $L_1 \succsim L_2$.

We then have *Extrapolated Expectations*, which should also be uncontroversial for those who accept expected value theory in the finite context. It merely says that, for any world $W' \succ \mathbf{0}$ (where $\mathbf{0}$ is the world with value 0 at every location), a lottery with probability p of W' (and $\mathbf{0}$ otherwise) is precisely as good as $p \cdot W'$.

Extrapolated Expectations: For any world $W' \in \mathcal{W}$ such that $W' \succ \mathbf{0}$ and any $p \in (0, 1]$ the following lottery $L \simeq_L p \cdot W'$.

$$L(W) = \left\{ \begin{array}{ll} p & \text{for } W = W' \\ 1 - p & \text{for } \mathbf{0} \\ 0 & \text{for } W \notin \{W', \mathbf{0}\} \end{array} \right\}$$

If we faced a similar lottery over finite payoffs, w and 0, we'd assign it expected value $p \cdot w$ without hesitation. Extrapolated Expectations requires that we treat infinite worlds in the same way. And it's on firm footing of its own—if our $\succsim$ for comparing worlds obeys Additivity, as any plausible $\succsim$ does, then Extrapolated Expectations follows straightforwardly from Finite Expectations and Stochastic Dominance. It is also worth noting that it follows from Local Expectations—if one is tempted to adopt Local Expectations instead of ED2, Extrapolated Expectations must hold either way.

And if you accept Extrapolated Expectations, Independence, Stochastic Dominance, and Separability of Value, then you must also accept Expectations of Differences 2, by Theorem 4.

5.5.3 Weaknesses of Expectations of Differences

You might notice in the definition of Expectations of Differences 2 that the antecedent of the conditional is awfully strong. The rule, then, is awfully weak. It can only provide judgements when:

- 1) *Every* world in the domain of either lottery contains precisely the same locations;
- 2) *Every* pair of such worlds is comparable;
- 3) *Many* pairs of such worlds have a difference which is also comparable to many of the other differences; and
- 4) Those differences are *also* comparable with certain scalar multiples of other differences (depending on the probabilities in the lotteries).

To demonstrate my view's weakness, here is a pair of lotteries similar to those I used earlier to demonstrate Local Expectations in action, but with *slightly* different probabilities. (Here, ϵ is some small positive number.)

$$\begin{array}{ccccccccccccccc}
 & l_a & l_b & l_c & l_d & l_e & l_f & l_g & l_h & l_i & l_j & l_k & \dots \\
 W_1 : & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & \dots
 \end{array}$$

$$L \left\{ \begin{array}{l|ccccccccccccccc}
 L(W) & l_a & l_b & l_c & l_d & l_e & l_f & l_g & l_h & l_i & l_j & l_k & \dots \\
 1/2 - \epsilon & W_2 : & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \dots \\
 1/2 + \epsilon & W_0 : & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots
 \end{array} \right.$$

What does ED2 say? Well, here are our differences-below, and a relevant scalar multiple.

$$\begin{array}{ccccccc}
 & l_a & l_b & l_c & l_d & l_e & \dots \\
 D(W_2, W_1) : & 1 & 2 & 1 & 2 & 1 & \dots \\
 D(W_1, W_0) = W_1 : & 1 & 0 & 1 & 0 & 1 & \dots \\
 (1 + \epsilon') \cdot D(W_1, W_0) : & 1 + \epsilon' & 0 & 1 + \epsilon' & 0 & 1 + \epsilon' & \dots
 \end{array}$$

By Additivity alone, we cannot compare $D(W_2, W_1)$ to the scalar multiple $(1 + \epsilon') \cdot D(W_1, W_0)$, since the former does better at locations $l_b, l_d, l_f, \dots$ and the latter does better at the rest. The sum of

their local differences is undefined. And given the probabilities here, there is no way to generate k_i s that satisfy both of the equations needed for ED2.

Thus, ED2 and Additivity together say nothing about how we should compare W_1 to L . And this is *even though* every pair of worlds at play is comparable and, on top of that, so are the differences between each of them. Even worse, we only made the smallest of changes to the lottery—all we did was give a mere $+\epsilon$ of probability mass to one outcome, and now we cannot say a thing!

My first response to this silence is simply: *tu quoque*. Suppose we adopt the rival view of Arntzenius, Bostrom, and Meacham, and we swap out ED2 for Local Expectations. Then we still cannot compare these two lotteries. So I am doing no worse here. But that is little comfort.

My second response is that this is no shortcoming of Expectations of Differences; it is a shortcoming of Additivity. Additivity is a very weak constraint on betterness. I have used it so far *because* it is so weak—this weakness makes it uncontroversial and, indeed, all of the plausible stronger principles in the literature are consistent with it. We can treat the conjunction of ED2 and Additivity similarly: it is just a weak and (hopefully) uncontroversial condition for comparing lotteries. So it is fitting that it makes no judgement in this case. The correct judgement is not obvious, so it should remain silent.

My third, closely related, response is that actually ED2 allows us to do even better. It can be combined with almost any betterness relation we choose—the definition above makes no reference to Additivity, but instead to some unspecified $\succsim$ relation. We might adopt the $\succsim$ relation from Vallentyne and Kagan (1997, p. 19), Jonsson and Voorneveld (2018), or whatever else we like. I suggest that we adopt RSE from the previous chapter. (Or, in cases where relativity presents no problem, we can just adopt SE2 for simplicity). As we saw in Chapter 3, SE2 (and so RSE as well) has probability arbitrarily close to 1 of being able to compare any worlds that would result from each of two actions we might ever need to choose between. And as a result, as I show in the next section, SE2 (or RSE) can compare the lotteries associated with any two actions we might ever need to choose between.

And even without a betterness relation stronger than Additivity, ED2 does just fine in the examples in the previous sections, including Egregious Energy. So this approach is not so weak after all.

5.6 Lotteries over chaotically different worlds

Recall from Chapter 3 the sorts of outcomes we actually face in practice, such as W_1 and W_2 below. These outcomes may differ in an entirely predictable way at some locations: for instance, the agent may be certain (or near enough) that additional value v arises at location l_a in W_2 , as below. But the agent will be uncertain of the value that arises at infinitely many other locations, with that uncertainty independent of what the agent chooses and, for at least infinitely many such locations, independent of what happens at the other such locations. We can represent the actual outcomes of the agent's available options as below. Each X_i and Y_i is a random variable drawn from the same distribution as the corresponding variable in the other outcome, and such that their differences $X_i - Y_i$ are distributed symmetrically about 0 and at least partially independently.

$$\begin{array}{rcccccc}
 & x_a & x_b & x_c & x_d & x_e & \cdots \\
 W_1 : & 0 & X_b & X_c & X_d & X_e & \cdots \\
 W_2 : & v & Y_b & Y_c & Y_d & Y_e & \cdots
 \end{array}$$

This would be extremely complicated to represent in the form of lotteries L_1 and L_2 , to list each possible world, or to calculate each and every Δp_i and D_i . But we don't need to.

Take any *particular* outcome that could result from L_2 , such as $W'_2 = (v, Y_b, Y_c, Y_d, Y_e, \dots)$. This outcome will be strictly better than the corresponding outcome $W'_1 = (0, Y_b, Y_c, Y_d, Y_e, \dots)$ that could result from L_1 (by SE2, or simply by Pareto Over Positions). And those two outcomes have precisely the same probability in their respective lotteries. So too, the probability that L_2 turns out *at least as good as* W'_2 is precisely as great as the probability that L_1 turns out at least as well as W'_1 . And the probability that L_1 turns out at least that good but *worse* than W'_2 will be greater than 0. So we have a strictly *higher* probability of getting an outcome at least as good as W'_2 under L_2 than we do under L_1 , and that holds for *all* such possible outcomes of L_2 .

This means that every Δp_i will be greater than zero. So ED2 straightforwardly implies that that L_2 is a better lottery than L_1 . (Even without the full machinery of ED2, Stochastic Dominance by itself would imply that L_2 is better.) And so we can conclude that, in cases of two lotteries that each contain chaotic effects but have identical probability distributions over those chaotic effects, the lottery that has better non-chaotic, predictable effects can be judged the better lottery overall.

In the case from Chapter 3, we can indeed say that it is better to save the child than it is to let them walk into traffic, even though we are clueless about exactly how either action would turn out. With a combination of SE2 and ED2, we can indeed obtain plausible practical judgements in cases like this.¹⁵

5.7 Summary

In earlier chapters, the discussion focused on how to compare outcomes with infinite moral value. But you may have worried that such discussion would be of little help in comparing the options available to us in practice. After all, when agents like us must make real-world decisions, we are almost never certain of how our choices will turn out. So it is not enough to know how to compare outcomes; we need to be able to compare *lotteries* over outcomes. Being able to do the former won't help us much at all unless we have a method of extending comparisons of outcomes to comparisons of lotteries.

In this chapter, I hope to have addressed this worry. We can indeed extend comparisons to infinitely-valued outcomes to make comparisons of infinitely-valued lotteries. And we can do so without bringing on nasty implications such as violations of Guaranteed Betterness (as are implied by each of the previous proposals). We now have one way of doing so: the Expectations of Differences approach.

For this approach to deliver plausible judgements, we need our betterness relation over infinitely-valued worlds to be extremely strong. It is fortunate then we have such a strong relation from the preceding chapters—that given by spatiotemporal expansionism. Combining that relation with the Expectations of Differences principle given above, we can compare a great many lotteries. We can even compare the lotteries that agents face when their available actions produce widespread chaotic effects. And it's a good thing that we can because, as we saw in Chapter 3, we often face decisions between such lotteries!

¹⁵Note that, with these principles, the situation is much like that of the finite context. We can compare actions in terms of instrumental betterness, according to the lotteries they produce. But we cannot compare them in terms of objective betterness, according to the outcomes they *actually* produce, at least not with any confidence. But this is precisely what we can say in similar cases (and even much simpler cases—see Jackson 1991) in the finite context (Greaves, 2016), so this situation seems satisfactory.

5.8 Proofs

Theorem 2: For any reflexive, transitive relation $\succsim_L$ on $\mathcal{P}$, if $\succsim_L$ satisfies Stochastic Dominance, Finite Expectations, and Separability of Value (for Lotteries) then it satisfies ED1.

Proof: Define the binary relation $\succsim_{ED1}$ on $\mathcal{P}$ by:

$L_1 \succsim_{ED1} L_2$ if and only if there is some $W_* \in \mathcal{W}$ such that

$$\sum_{W_j \in \mathcal{W}_1} L_1(W_j) \left(\sum_{l \in \mathcal{L}} (V_j(l) - V_*(l)) \right) \geq \sum_{W_j \in \mathcal{W}_2} L_2(W_j) \left(\sum_{l \in \mathcal{L}} (V_j(l) - V_*(l)) \right) \quad (i)$$

This is the relation given by ED1 alone. To prove Theorem 2, it suffices to show that, for any L_1 and L_2 such that $L_1 \succsim_{ED1} L_2$, it must also hold that $L_1 \succsim_L L_2$ if $\succsim_L$ abides by Stochastic Dominance, Finite Expectations, and Separability of Value (for Lotteries).

Let $L_1, L_2 \in \mathcal{P}$ be some such lotteries. (We already know that such lotteries exist—see Section 6.) Then there exists some $W_* \in \mathcal{W}$ such that (i) holds. So either 1) both sides of the inequality converge unconditionally to some real values, with the LHS greater than or equal to the RHS, or 2) the LHS diverges unconditionally to $+\infty$, the RHS diverges unconditionally to $-\infty$, or both.

1) If both sides of (i) converge:

Define two additional lotteries L_1^* and L_2^* by $L_1^*(W - W_*) = L_1(W)$ and $L_2^*(W - W_*) = L_2(W)$ for all $W \in \mathcal{W}_{(1,2)}$ (and $L_1^*(W), L_2^*(W) = 0$ otherwise). Since both sides of (i) converge unconditionally to real values, the expected total sums of local value in L_1^* and L_2^* will be finite, and the expected total sum for L_1^* will be greater than or equal to that for L_2^* . So $L_1^* \succsim_L L_2^*$, by Finite Expectations. Then, by Separability of Value (for Lotteries), $L_1 \succsim_L L_2$, as required.

2) If either or both sides of (i) diverge:

The LHS diverges to $+\infty$, or the RHS diverges to $-\infty$, or both. If both, then define L_1^* as above and $L_2^*(W_0) = 1$. If not both, then define L_1^*, L_2^* as above.

If the LHS diverges to $+\infty$, then there is some lottery L_1^{**} with finite expected sum such that $L_1^* \succsim_L L_1^{**}$ by Stochastic Dominance and $L_1^{**} \succsim_L L_2^*$ by Finite Expectations. We can obtain L_1^{**} from L_1^* by replacing any worlds in its domain that have total sums of local value greater than some chosen large finite bound with some other worlds with totals below that bound.

If the LHS does not diverge but the RHS does diverge to $-\infty$, then there is some lottery L_2^{**}

such that $L_2^* \preceq_L L_2^{**}$ by Stochastic Dominance and $L_2^{**} \preceq_L L_1^*$ by Finite Expectations.

Either way, $L_1^* \succ_L L_2^*$. Separability of Value (for Lotteries) then implies that $L_1 \succ_L L_2$. $\square$

Theorem 3: For any reflexive, transitive relation $\succ_L$ on $\mathcal{P}$, if $\succ_L$ satisfies ED2 and Additivity, then it satisfies ED1.

Proof:

Define a binary relation $\succ_{ED2}$ on $\mathcal{P}$ by:

$L_1 \succ_{ED2} L_2$ if and only if 1) for all worlds $W_a, W_b \in \mathcal{W}_{(1,2)}$, $\mathcal{L}_a = \mathcal{L}_b$, and 2) there is some $W_* \in \mathcal{W}_{(1,2)}$ such that, for each $W_i \in \mathcal{W}_{(1,2)}$, there is some $k_i \in \mathbb{R}$ such that

$$k_i \Delta p_i \cdot D_* \preceq \Delta p_i \cdot D_i \quad \text{and} \quad \sum_{W_i \in \mathcal{W}_{(1,2)}} k_i \Delta p_i \geq 0$$

This is the relation given by ED2 alone. Define $\succ_{ED1}$ as above, and let $L_1, L_2 \in \mathcal{P}$ be any lotteries such that $L_1 \succ_{ED1} L_2$. And assume that $\succ$ satisfies Additivity. Then, to prove Theorem 3, it suffices to show that $L_1 \succ_{ED2} L_2$.

Since $L_1 \succ_{ED1} L_2$, for some $W' \in \mathcal{W}_{(1,2)}$,

$$\begin{aligned} \sum_{W_j \in \mathcal{W}_1} L_1(W_j) \left(\sum_{l \in \mathcal{L}} (V_j(l) - V'(l)) \right) &\geq \sum_{W_j \in \mathcal{W}_2} L_2(W_j) \left(\sum_{l \in \mathcal{L}} (V_j(l) - V'(l)) \right) \\ \Rightarrow \sum_{W_j \in \mathcal{W}_{(1,2)}} (L_1(W_j) - L_2(W_j)) \left(\sum_{l \in \mathcal{L}} (V_j(l) - V'(l)) \right) &\geq 0 \quad (i) \end{aligned}$$

Assume that the signs of $L_1(W_j) - L_2(W_j)$ and $\sum_{l \in \mathcal{L}} (V_j(l) - V'(l))$ differ for some $W_j \in \mathcal{W}_{(1,2)}$, and hence also that the signs are the same for some other W_i . (ii) (If not, then we immediately have $L_1 \succ_{ED2} L_2$.)

Given (i), either 1) for at least one W_j , $\sum_{l \in \mathcal{L}} (V_j(l) - V'(l))$ diverges unconditionally to $+\infty$, or 2) all of those sums are finite.

1) For at least one W_j , $\sum_{l \in \mathcal{L}} (V_j(l) - V'(l))$ diverges unconditionally to $+\infty$:

There will be at least one D_i with an infinite sum of local values and $\Delta p_i > 0$, but none with $\Delta p_i > 0$. Given (ii), we also have at least one D_* with negative Δp_* . Given that the sum diverges unconditionally, that D_* must only have a finite total sum of local values. For all finite-sum D_i , set $k_i = 0$. For all infinite-sum D_i , set $k_i = 1$. Then we have $k_i \Delta p_i \cdot D_* \preceq \Delta p_i \cdot D_i$ for all D_i .

And $\sum_{W_i \in \mathcal{W}_{(1,2)}} k_i \Delta p_i \geq 0$, since it will simply be the sum of the (at least one) infinite-sum D_i with positive Δp_i . Both conditions are thus satisfied, and we have $L_1 \succ_{ED1} L_2$, as required.

2) For all W_j , $\sum_{l \in \mathcal{L}} (V_j(l) - V'(l))$ is finite:

Then each D_i will also have a finite total sum of local value, $S_i = \sum_{l \in \mathcal{L}} D_i(l)$.

Let $D_* = D'$. Then let each $k_i = \frac{S_i}{S_*}$. Since all S_i are finite and all $D_i \succcurlyeq 0$, Additivity says that $k_i \Delta p_i \cdot D_* \simeq \Delta p_i D_i$ if and only if $k_i \Delta p_i \times S_* = \Delta p_i S_i$. Since $k_i = \frac{S_i}{S_*}$, this holds for all D_i , as required.

We now seek the second condition. First, note that $P(W) \cdot V + (1 - P(W)) \cdot V' = V' + P(W) \cdot (V - V')$. By iterating that rearrangement, we can obtain the following from (i).

$$\begin{aligned}
 & \sum_{W_i \in \mathcal{W}_{(1,2)}} \left(P_{L_1}(W_i \text{ or better}) - P_{L_2}(W_i \text{ or better}) \right) \left(\sum_{l \in \mathcal{L}} (V_i(l) - V_j(l)) \right) \geq 0 \\
 & \quad \left(\text{where } W_j = \max\{W \in \mathcal{W}_{(1,2)} \mid W \prec W_i\} \right) \\
 & \Leftrightarrow \sum_{W_i \in \mathcal{W}_{(1,2)}} \Delta p_i \sum_{l \in \mathcal{L}} D_i(l) \geq 0 \\
 & \Leftrightarrow \sum_{W_i \in \mathcal{W}_{(1,2)}} \Delta p_i S_i \geq 0 \\
 & \Leftrightarrow \frac{1}{S_*} \sum_{W_i \in \mathcal{W}_{(1,2)}} \Delta p_i S_i \geq 0 \\
 & \quad (\text{since } S_* > 0) \\
 & \Leftrightarrow \sum_{W_i \in \mathcal{W}_{(1,2)}} \Delta p_i k_i \geq 0 \quad \square
 \end{aligned}$$

To prove Theorem 4 below, it will help to first establish Lemma 1. If Independence and Extrapolated Expectations hold, then this lemma effectively implies that we can evaluate any lottery in which every outcome is some scalar multiple of some world $W \succ 0$. And we evaluate it as equally good as the probability-weighted sum of the scalar multiples of W . Effectively, the lottery is valued at its expected value, as some multiple of W .

Lemma 1: Let $L \in \mathcal{P}$ be any finitely-supported lottery such that, for some $W \in \mathcal{W}$, $L(W_j) > 0$ if and only if W_j is such that $W_j = k_j \cdot W$ for some positive, real k_j . If $\succ_L$ is a reflexive, transitive relation on $\mathcal{P}$ which satisfies Independence and Extrapolated Expectations, then:

$$L \simeq_L \left(\sum_{W_j \in \mathcal{W}} L(W_j) \times k_j \right) \cdot W$$

Proof:

Let $k_{\max}$ be the greatest such k_j (which exists and is unique, since L is finitely-supported). For each k_j , $0 < \frac{k_j}{k_{\max}} < 1$.

Let L_j be the lottery such that $L_j(k_{\max} \cdot W) = \frac{k_j}{k_{\max}}$ and $L_j(\mathbf{0}) = 1 - \frac{k_j}{k_{\max}}$. (So the remaining probability mass goes to $\mathbf{0}$.) Since $k_j \cdot W = \frac{k_j}{k_{\max}} \cdot (k_{\max} \cdot W)$, Extrapolated Expectations implies that:

$$k_j \cdot W \simeq_L L_j \quad (\text{i})$$

By Independence, we can replace each world $k_j \cdot W$ in L with the lottery L_j . In other words, by (i) and Independence, $L \simeq_L L'$, where:

$$L'(W') = \begin{cases} \sum_{k_j \in \mathbb{R}} \left(L(k_j \cdot W) \times \frac{k_j}{k_{\max}} \right) & \text{for } W' = k_{\max} \cdot W \\ 1 - L'(k_{\max} \cdot W) & \end{cases}$$

By Extrapolated Expectations, $L' \simeq_L \left(\sum_{k_j \in \mathbb{R}} \left(L(k_j \cdot W) \times \frac{k_j}{k_{\max}} \right) \right) \cdot (k_{\max} \cdot W)$.

$\therefore L \simeq_L \left(\sum_{W_j \in \mathcal{W}} (L(W_j) \times k_j) \right) \cdot W$, as required. $\square$

Theorem 4: For any reflexive, transitive relation $\succsim_L$ on $\mathcal{P}$, if $\succsim_L$ satisfies Stochastic Dominance, Separability of Value, Independence, and Extrapolated Expectations, then it satisfies ED2.

Proof:

Define $\succsim_{\text{ED2}}$ as above, and let L_1, L_2 be any lotteries such that $L_1 \succsim_{\text{ED2}} L_2$. To prove Theorem 4, it suffices to show that $L_1 \succsim_L L_2$.

If $\Delta p_i \geq 0$ for all $W_i \in \mathcal{W}_{(1,2)}$, then $L_1 \succsim_L L_2$, by Stochastic Dominance, as required.

If not, there are some $W_i, W_j \in \mathcal{W}_{(1,2)}$ such that $\Delta p_i < 0$ and $\Delta p_j > 0$. Assume from here on that such W_i and W_j exist.

From the definition of $\succsim_{\text{ED2}}$, there is some $W_* \in \mathcal{W}_{(1,2)}$ such that, for each $W_i \in \mathcal{W}_{(1,2)}$ there is $k_i \in \mathbb{R}$ such that:

$$k_i \Delta p_i \cdot D_* \preceq \Delta p_i \cdot D_i \quad (\text{i})$$

$$\text{and } \sum_{W_i \in \mathcal{W}_{(1,2)}} k_i \Delta p_i \geq 0 \text{ (ii)}$$

For some such k_i s and W_* , define lotteries $L_1^D, L_2^D \in \mathcal{P}$ by

$$L_1^D \left(\left(\sum_{\{W_j \in \mathcal{W}_{(1,2)} | W_j \preceq W_i\}} k_j \right) \cdot D_* \right) = L_1(W_i) \quad \text{for all } W_i \in \mathcal{W}_1$$

and likewise for L_2^D , mutatis mutandis.

And $D_* \succ \mathbf{0}$ since it is the difference between some world and a worse world, by Separability of Value. So Lemma 1 applies to both L_1^D and L_2^D .

$$\therefore L_1^D \simeq_L \left(\sum_{W_i \in \mathcal{W}_1} L_1(W_i) \left(\sum_{\{W_j \in \mathcal{W}_{(1,2)} | W_j \preceq W_i\}} k_j \right) \right) \cdot D_* \quad (\text{by Lemma 1})$$

and likewise for L_2^D (mutatis mutandis).

So $L_1^D \succcurlyeq_L L_2^D$ iff that scalar multiple of D_* is as good or better than the corresponding world for L_2^D . Since $D_* \succ \mathbf{0}$, that holds if and only if:

$$\sum_{W_i \in \mathcal{W}_1} \left(L_1(W_i) \left(\sum_{\{W_j \in \mathcal{W}_{(1,2)} | W_j \preceq W_i\}} k_j \right) \right) \geq \sum_{W_i \in \mathcal{W}_2} \left(L_2(W_i) \left(\sum_{\{W_j \in \mathcal{W}_{(1,2)} | W_j \preceq W_i\}} k_j \right) \right)$$

$$\Leftrightarrow k_{\min} + \sum_{W_i \in \mathcal{W}_{(1,2)}} (L_1(\succcurlyeq W_i) \times k_i) \geq k_{\min} + \sum_{W_i \in \mathcal{W}_{(1,2)}} (L_2(\succcurlyeq W_i) \times k_i)$$

$$\text{for } k_{\min} = \min\{k_i | W_i \in \mathcal{W}_{(1,2)}\}$$

$$\Leftrightarrow \sum_{W_i \in \mathcal{W}_{(1,2)}} k_i \Delta p_i \geq 0, \text{ which is given by (ii).}$$

Therefore, $L_1^D \succcurlyeq_L L_2^D$.

For some such k_i s and W_* as above, define lotteries L_1^*, L_2^* on $\mathcal{W}$ by

$$L_1^*(W - W_*) = L_1(W) \quad \text{for all } W \in \mathcal{W}_1$$

and likewise for L_2^* (mutatis mutandis).

These lotteries resemble L_1 and L_2 ; they have the same probabilities, but W_* is subtracted from each outcome. As a result, each outcome $W - W_*$ can be represented as a sum of differences

$$W - W_* = \sum_{\{W_j \in \mathcal{W}_{(1,2)} | W_* \preccurlyeq W_j \preccurlyeq W\}} D_j$$

$$(\text{or, for } W \prec W_*, W - W_* = - \sum_{\{W_j \in \mathcal{W}_{(1,2)} | W \preccurlyeq W_j \preccurlyeq W_*\}} D_j).$$

But, by (i), whenever $\Delta p_i > 0$,

$$W_i - W_* = \sum_{\{W_j \in \mathcal{W}_{(1,2)} | W_* \preccurlyeq W_j \preccurlyeq W_i\}} D_j \succcurlyeq \left(\sum_{\{W_j \in \mathcal{W}_{(1,2)} | W_j \preccurlyeq W_i\}} k_j \right) \cdot D_*.$$

And, whenever $\Delta p_i < 0$, the inequality is reversed.

Therefore, by Stochastic Dominance and (i), $L_1^* \succcurlyeq_L L_1^D$ and $L_2^* \preccurlyeq_L L_2^D$. And, since $L_1^D \succcurlyeq_L L_2^D$, transitivity implies that $L_1^* \succcurlyeq_L L_2^*$.

By Separability of Value, $L_1 \succcurlyeq_L L_2$ if and only if $L_1^* \succcurlyeq_L L_2^*$. Therefore, $L_1 \succcurlyeq_L L_2$, as required.

□

The question of fanaticism

6.1 Introduction

Where do we now stand? We have a coherent, and fairly strong, method for comparing outcomes with infinite value. Likewise, we have such a method for comparing lotteries over such outcomes.

But there is one key question about infinitely valuable outcomes that has plagued philosophers all the way back to Pascal (1670) and which none of the preceding discussion has directly addressed. Specifically, should we prefer any probability (no matter how *tiny*) of infinite gain over a *sure* finite gain (no matter how great)?

To illustrate this worry, consider *Dyson's Wager* (a moral analogue of the wager that interested Pascal).

Dyson's Wager

You have \$2,000 to use for charitable purposes. You can donate it to either of two charities.

The first charity distributes bednets in low-income countries in which malaria is endemic.¹ With an additional \$2,000 in their budget this year, they would prevent one additional death from malaria. You are certain of this.

The second charity does speculative research into how to do computations using 'positronium'—a form of matter which will be ubiquitous in the far future of our

¹I have in mind the Against Malaria Foundation. As of 2019, the charity evaluator GiveWell estimated that the Against Malaria Foundation prevents the death of an additional child under the age of 5 for, on average, every US\$3,710 donated (GiveWell 2020). Including other health benefits, a total benefit *roughly* equivalent to that is produced for, on average, every US\$1,690 donated. Of course, in reality, a donor can never be *certain* that their donation will result in an additional life saved. This assumption of certainty is for the sake of simplicity.

universe. If our universe has the right structure (which it probably does not), then in the distant future we may be able to use positronium to instantiate all of the operations of human minds living blissful lives, and thereby allow morally valuable life to survive indefinitely long into the future.²³ From your perspective as a good epistemic agent, there is some tiny, non-zero probability that, with (and only with) your donation, this research would discover a method for stable positronium computation and would be used to bring *infinitely many* blissful lives into existence. Assume that there would be no morally relevant difference between the outcomes except whether that child dies of malaria and whether those infinitely many lives in the far future come to exist or not. (There are, for instance, no chaotic effects here.)

Which is the better option here: saving a life with certainty, or generating a tiny probability of bringing about infinitely many future lives?

My proposals in the preceding chapters entail that the better lottery is the one that results from donating to the speculative research. And, in general, they say that it is better to have a slim probability of producing a world with infinite additional value than it is to have a merely finite gain (even with certainty). Combine spatiotemporal expansionism with the ED2 principle from the previous chapter⁴ and you reach this verdict.

But this verdict is not unique to my proposal. We can reach much the same verdict on the standard views of the finite setting. Suppose that, if the positronium research does not succeed, the universe will otherwise contain only finite total value. Any minimally aggregative view will say that this is vastly worse than producing those infinitely many valuable lives. Then we might take the standard theory of moral decision-making under risk—expected value theory—which says that one option is better than another if and only if it has the greater probability-weighted

²³Dyson (1981) was the first to suggest positronium as a medium for computation and information storage. This follows Dyson (1979), wherein it is argued that an infinite duration of computation could be performed with finite energy if the computation hibernates intermittently, and if the universe has a particular structure. Tipler (1986) suggests an alternative method which may work if the universe has a different structure. Sandberg (n.d.) argues that both Dyson and Tipler's proposals are unlikely to work, as our universe appears to match neither structure. Nonetheless, it is still epistemically *possible* that the universe has the right structure for Dyson's proposal. And possibility is sufficient for my purposes.

³Would such artificially-instantiated lives hold the same moral value as lives led by flesh-and-blood humans? I assume that they would, if properly implemented. See Chalmers (2010) for arguments supporting this view. And note that, for the purposes of the example, all that's really needed is that it is epistemically *possible* that the lives of such simulations hold similar moral value.

⁴In fact, the weaker ED1 principle implies this too.

sum of value (i.e., the greater *expected value*).⁵ Given that the value of producing infinitely many lives is positively infinite, the expected value of funding the speculative research is positively infinite too, while the expected value of saving the child from malaria is merely finite. So the combination of these views will say that it is better to fund the research and let the child die.

And this verdict is thought to be counterintuitive by many. All the more counterintuitive is that it can still be better to donate to the speculative research *no matter how low* the probability is (short of being 0), since there are so many blissful lives at stake. For instance, the odds of your donation actually making the research succeed could be 1 in 10^{100} . (10^{100} is greater than the number of atoms in the observable universe). The chance that the research yields nothing at all would be 99.99... percent, with another 96 nines after that. And yet my proposal (and expected value theory) says that it is better to take the bet, despite it being *almost guaranteed* that it will actually turn out worse than the alternative; despite the fact that you will almost certainly have let a person die for no actual benefit. Surely not, says my own intuition. On top of that, suppose that \$2,000 spent on preventing malaria would save more than one life. Suppose it would save a billion lives, or any enormous finite value. My proposal would say that it's still better to fund the speculative research—it says that it would be better to sacrifice those billion or more lives for a minuscule chance at the infinitely many blissful lives (and likewise if the number of blissful lives were finite but still sufficiently many). But endorsing that verdict, regardless of how low the probability of success and how high the cost, seems *fanatical*. Likewise, even without infinite value at stake, it would also seem fanatical to judge a lottery with sufficiently tiny probability of arbitrarily high *finite* value as better than getting some modest value with certainty.

But my proposal could be modified to avoid such fanatical verdicts, just as expected value theory has previously been modified to avoid them. Many have proposed that expected value theory be modified to ignore outcomes with small enough probabilities (e.g. le Rond d'Alembert, 1984; de Buffon, 1777; Monton, 2019; Smith, 2014), as is often done in applied contexts such as

⁵Recall from above that expected *value* theory is distinct from expected *utility* theory. Under the latter, *utility* is given by some (indeed, *any*) increasing function of value—perhaps a concave function, such that additional value contributes less and less additional utility. The utility of an outcome may even be bounded, such that arbitrarily large amounts of additional value contribute arbitrarily little additional utility. And, if the utility function is bounded, then the expected utilities of lotteries will be bounded as well. As a result, expected utility theory can avoid the fanatical verdict described here. But, if it does, it faces the objections raised in Sections 6.4, 6.5, and 6.6. Where relevant, I will indicate in notes how the argument applies to expected utility theory.

law and risk management, and we could do the same to ED2.⁶ Others have proposed that we evaluate outcomes not by their expected values but by their expected *utilities* (given by some increasing function of an outcome's value), and that the correct utility function is bounded above so as to keep the expectation of utility bounded as well (e.g. Arrow, 1971, p. 64). Again, we could do the same to ED2.⁷ Or we could simply abandon ED2 in favour of a weaker principle like Stochastic Dominance, as Tarsney (n.d.) does for expected value theory.

Should we abandon ED2 in favour of one of these alternatives? Likewise, in the finite context, should we abandon expected value theory in favour of one of its rivals? No, we shouldn't, as I argue in this chapter. As it turns out, there are compelling arguments in favour of accepting fanatical verdicts such as those above.

I will focus on a particular claim that is sufficient to imply fanatical verdicts—namely, *Fanaticism*.⁸

Fanaticism: For any tiny (finite) probability $\epsilon > 0$, and for any finite value v , there is some finite V that is large enough that L_{risky} is better than L_{safe} (no matter which scale those cardinal values are represented on).⁹

L_{risky} : value V with probability ϵ ; value 0 otherwise

L_{safe} : value v with probability 1

The lotteries L_{risky} and L_{safe} closely resemble the options in Dyson's Wager: one gives a slim chance of a potentially astronomical value V ; the other a certainty of some modest value v . But, in a deviation from the previous four chapters, there is no need for any of the outcomes here to

⁶In the sum that appeared in the definition, simply replace all sufficiently small probabilities Δp_i with 0.

⁷Simply modify the first equation in the definition to discount k_i somewhat so it's bounded above.

⁸This use of the term 'fanaticism' seems to originate with Bostrom (2011) and Beckstead (2013, ch. 6) (however Beckstead uses the term 'Fanaticism' for a similar claim specific to infinite values and instead uses 'Recklessness' for a claim more akin to my version of Fanaticism). My formulation is slightly stronger than each of theirs but also, unlike theirs, applicable even if infinite total value cannot exist.

⁹Note that this parenthetical makes Fanaticism quite a strong claim. As defined here, it requires that the ranking of L_{risky} above L_{safe} holds not only when the number of lives in each outcome is proportional to V , v , and 0. It must hold whenever outcomes can be *cardinally represented* with those values. Recall that cardinal representations of value are unique only up to positive affine transformations—two outcomes represented by 0 and v on one scale could instead be represented by $0 \times a + b$ and $v \times a + b$ (for any positive a and real b). Conversely, an outcome that contains many happy lives might still be represented cardinally with value 0. So Fanaticism doesn't apply only to risky lotteries in which some possible outcome contains zero valuable lives, or zero value on net. It also applies to lotteries that can be represented as L_{risky} and L_{safe} even though every one of their outcomes contains enormous numbers of blissful lives, or enormous amounts of suffering, as long the differences in value between those outcomes are proportional to 0, v , and V .

have infinite total value— V and v are explicitly finite numbers.

But whether Fanaticism of this form holds true will still have important ramifications in the infinite context. Suppose that just the value V in L_{risky} were replaced with a positive infinite value, much like it is in Dyson's Wager. The resulting lottery would be strictly better than L_{risky} by Stochastic Dominance (see §6.3). And, by Fanaticism, it would also be better than any such L_{safe} . So we would have a fanatical verdict. And this verdict would carry over to the infinite context. Simply add the same values to the same infinitely many locations in all three outcomes; as long as lotteries over infinite outcomes are compared on the basis of only their differences (or, equivalently, if the principle of Separability of Value (for Lotteries) from Chapter 5 holds), then the lottery with a chance of infinite gain will still be better.

Given this, I will streamline the discussion in this chapter by focusing on the claim of Fanaticism and on lotteries over finitely-valued outcomes. If Fanaticism is true, we are still led to fanatical verdicts in the infinite setting. And, conversely, to avoid fanatical judgements in the infinite setting we must deny Fanaticism. But Fanaticism also has force outside the infinite setting, and so should be of broader interest too.

So, is Fanaticism true? You might think not. It may seem deeply counterintuitive, just as the fanatical verdict in Dyson's Wager was, and so you might think it easy to reject. Many decision theorists have done so (e.g. le Rond d'Alembert, 1984; de Buffon, 1777; Arrow, 1971; Monton, 2019; Smith, 2014; Tarsney, n.d.), while few have been willing to explicitly accept fanatical verdicts.¹⁰ But, as I argue in this chapter, there are compelling arguments in favour of Fanaticism. In particular, there are compelling arguments that do not rely on expected value theory, and so we have good reason to accept it even if we reject that particular theory.

The chapter proceeds as follows. §6.2 addresses some common motivations for rejecting Fanaticism. §6.3 briefly introduces the formal framework for what follows. And §6.4 through §6.6 present arguments in favour of Fanaticism, each premised on weaker claims than expected value theory, and each (by my reckoning) more compelling than the last. The first is a basic continuum argument. The second is that, to deny Fanaticism, we must accept either what I'll call

¹⁰Examples include Pascal (1670) himself, Parfit (1984, §3.27), and Hájek (2014). And even they most often endorse such verdicts only because they are a consequence of expected value theory, not because they see good independent justification for them. I suspect that even many diehard adherents of expected value theory are uncomfortable with the fanatical verdicts supplied by their theory.

‘scale dependence’ or an absurd sensitivity to arbitrarily small differences in probability. And the final nails in the coffin are an updated version of Parfit’s classic Egyptology Objection and what I’ll call the *Indology Objection*, by which those who deny Fanaticism must make judgements which appear deeply irrational. Together, these arguments show that, if we reject Fanaticism, we face disturbing implications. Better to just accept it. And if a theory implies Fanaticism, we should consider that a point in favour of the theory rather than a point against.

6.2 The case against fanaticism

You might think that Fanaticism is simply absurd, and nothing more needs saying about it. I have often heard this response. So, before presenting arguments in favour of Fanaticism, let us address a sample of the most common and compelling arguments against it: the argument from intuition; the argument from impossibility; the argument from tolerance; and the argument from mortality.¹¹ Although each argument seems initially plausible, I don’t think any of them turn out to be compelling. And even if you find them more compelling than I do, they certainly do not rule *decisively* against Fanaticism; the arguments in favour are still worth hearing.

6.2.1 The argument from intuition

The one objection to Fanaticism that I suspect is the most persuasive of all, despite its extreme simplicity, is this: Fanaticism is just so frightfully counterintuitive that it must be false.¹² It seems so very implausible that I should give up a certainty of a good payoff for a *tiny* chance of something better, no matter how tiny the chance. Rather, intuition says that an event with such a tiny chance can never tip the scales against the sure payoff.

But the psychology literature tells us that our initial intuitions about matters of probability are often misguided. We frequently commit the Conjunction Fallacy (Tversky and Kahneman, 1983), the Gambler’s Fallacy (Chen et al., 2016), the Hot Hand Fallacy (Gilovich et al., 1985), and the

¹¹There are others which I won’t address here, both for brevity and because I find them even less compelling. See de Buffon (1777), Monton (2019), Smith (2014), Schwitzgebel (2017), and Monton (2019) for other arguments which can be applied against Fanaticism. See also Hájek (2014), Isaacs (2016), Monton (2019), and Hájek (n.d.) for some compelling responses.

¹²In effect, this is a core point of Bostrom’s (2009) objection. Meanwhile, Bostrom (2011), Beckstead (2013), and Tarsney (n.d.) all treat Fanaticism as implausible for, it seems, the same reason.

Base Rate Fallacy (Kahneman and Tversky, 1982). Perhaps worst of all, we are prone to simply ignoring facts of probability—many of us choose based entirely on how we anticipate things will *actually* turn out, with no regard for risks of loss or chances of gain (Baron et al., 1993; Gurmankin and Baron, 2005). Even when we do account for probability, we often radically overestimate some probabilities due to availability bias (Tversky and Kahneman, 1974), and underestimate others out of indefensible optimism (Hanoch et al., 2019).

And our intuitions are no better with decision-making in the face of low-probability events. For instance, jurors are just as likely to convict a defendant based primarily on fingerprint evidence if that evidence has probability 1 in 100 of being a false positive as if it were 1 in 1,000, or even 1 in 1 million (Garrett et al., 2018). In another context, when presented with a medical operation which posed a 1% chance of permanent harm, many respondents considered it no worse than an operation with no risk at all (Gurmankin and Baron, 2005). And in yet another context, subjects were unwilling to pay any money at all to insure against a 1% chance of catastrophic loss (McClelland et al., 1993). So it seems that many of us, in many contexts, treat events with low probabilities as having probability 0, even when their true probability is as high as 1% (which is not so low). Upon reflection, this is clearly foolish. We may have the raw intuition that we may (or should) ignore all outcomes with probability 1% but it does not hold up to scrutiny, and so is likely mistaken.

Given how widespread these mistakes are, we should give little weight to our intuitions about what we should do in cases of low probability, including those which lead us to recoil from fanatical verdicts—with a little more scrutiny, those intuitions may appear foolish too.

6.2.2 The argument from impossibility

Here is another objection to Fanaticism. We might confidently claim that: ‘any event with a sufficiently low probability *will not happen*’. This claim is known as *Cournot’s Principle*, and was endorsed by many leading figures of probability theory, including not just Cournot (1843) but also Markov (1900), Levy (1925), Kolmogorov (1950), and Borel (1962). If they’re right that such events simply *won’t* happen, then it may seem fine to ignore them in our decision-making.

Why might we believe Cournot’s Principle? One reason might be that we should believe to be false any proposition which has probability sufficiently close to 0 (but not necessarily equal to

0) of being true. So, for any sufficiently unlikely event, we should believe that it won't happen. But this isn't reason enough. A low probability of an event may be enough to *believe* that it will not happen, but belief—at least in a binary form like this—is not enough for decision-making. We can believe a proposition without being entirely confident in it. And it would be foolish to ignore all possibilities which we believe won't occur, but nonetheless may well occur.

A second reason to accept it is this. Suppose there are only finitely many distinct ways that the full history of the universe could turn out. You might think that this is the case if all events can be fully specified by the properties of finitely many discrete particles at finitely many discrete times, and if all such properties admit discrete values (as do length and duration, in multiples of the Planck length and Planck time). Then, since there are only *finitely* many possible histories of the universe, in any given lottery there can be only finitely many outcomes with chance greater than zero. So there must be some minimum probability ϵ . And indeed no event with probability less than this will ever happen. There are no such events!

But this justification is unpersuasive. For one, for it to hold, every single physical property must admit discrete values, rather than continuous values. Perhaps that is true, but it quite plausibly isn't. For two, the universe would need to contain only finitely many particles, finitely many distinct times, and finitely many spatial positions. For three, not only must the universe be finite in all of these respects, but it must also be guaranteed to not to exceed some specific finite bound in each respect. It might not only be certain that the universe has, say, *some* finite number of particles, *de dicto*, but it may not be certain that it has some *specific* finite number of particles, *de re*. There could still be infinitely many possible numbers of particles just as there are infinitely many integers, even though each of them is finite. And that would give us infinitely many possible outcomes, thereby quashing the justification above. And finally, for three, Fanaticism wouldn't be ruled out even if we granted all of these required assumptions. After all, it's a claim about what we should do *if* we faced such a decision. The mere fact that we do or do not face such decisions in practice does not imply anything about what we should do if we did. A fully general normative theory—one that applies to a broader range of possible outcomes and probabilities—could well still be fanatical.

But no matter its justification, we can reject Cournot's Principle and the objection from impossibility. The normatively relevant notion of probability, I should think, is the agent's subjective

degree of belief (their *credence*) in the outcome.¹³ And there are events to which we should assign arbitrarily low credence, but which have occurred or will occur. For instance, consider the outcome that the Sydney Opera House is presently the exact distance that it actually is (*de re*) from the Golden Gate Bridge.¹⁴ As an agent myself, I am wildly uncertain of what that distance is. Perhaps I know that it must be an integer multiple of the Planck length, based on my understanding of physics. But I am uncertain even of exactly how long a Planck length is. (I admit I had to look it up.) You should be too, since there is no way we can measure the Planck length down to an arbitrarily precise value (nor is there any way we can be sure that the textbook we consult is accurate). An agent like you or me, even if we are epistemically rational, should place arbitrarily low credence on the exact distance between the two monuments being within an arbitrarily small interval of length. But there is some such precise length between the two sites, which lies within some arbitrarily small interval to which we've given credence arbitrarily close to zero. And yet that outcome has happened! So Cournot's Principle is demonstrably false.

6.2.3 The argument from tolerance

Here is another argument against Fanaticism, originally from Smith (2014, p. 471-2) and discussed at length by Monton (2019, p. 9-10).

Decision theories that reject Fanaticism aren't the only ones that say that we can ignore some outcomes. Every (remotely plausible) decision theory says that we can at least sometimes treat an outcome as having probability 0 in a given lottery: at the very least, when that outcome *does* have probability 0.

But decision-making is a practical activity, as Smith points out; it is done by agents with limited time and computational resources. Such agents cannot assign arbitrarily precise probabilities to every possible outcome in every lottery, so we must not require them to. Any norms of decision-making must be *tolerant*: they must allow agents at least a small margin of error in the conditions required for the norm to apply (at least for conditions like probability that have continuous values). And that includes the norm that we may permissibly ignore events with

¹³Alternatively, it might be the *evidential probability* of the outcome, which carries the same implications below.

¹⁴To avoid some of the possible vagueness here, we might replace 'the Sydney Opera House' with 'the centre of gravity of the Sydney Opera House', and similarly for the Golden Gate Bridge.

probability 0—we must allow some imprecision in that probability being 0, up to some margin of error ϵ . So, if an outcome has probability less than some ϵ in a given lottery, we must permit an agent to ignore it too.

An immediate response is that this conclusion, interpreted in a particular way, does not rule out Fanaticism. The value of ϵ —the non-zero probability below which outcomes may permissibly be ignored—may not be absolute. It might vary by lottery, by decision scenario, or by context. For instance, whenever we are comparing lotteries over only finitely many outcomes, ϵ might be set lower than the least probable outcome. This is still compatible with Fanaticism, and Smith appears to favour this version.¹⁵ But what if we interpret the conclusion the other way? Then there is some ϵ such that, in all lotteries, agents are permitted to ignore outcomes with probabilities less than ϵ . This does rule out Fanaticism, and is explicitly endorsed by de Buffon (1777).

Whichever interpretation we adopt, Smith's argument faces various objections (see Hájek, 2014; Isaacs, 2016; Monton, 2019, p. 10). Here are two of my own preferred objections to the Buffon interpretation.

By Smith's assumption, a theory of normative decision-making must be tolerant. And, to be tolerant, each norm it endorses must require only some non-arbitrary level of precision (in the probabilities involved). And if we accept this assumption of tolerance, our theory will also endorse the norm that, when evaluating a lottery, we may permissibly ignore outcomes with probabilities below ϵ . But tolerance requires that we tolerate imprecision in this norm as well. How much imprecision must we tolerate? I see no reason—at least, no reason that is not ad hoc—why we must be any more precise with this norm than with the norm of ignoring outcomes of probability 0. So a tolerant theory will also permit us to ignore outcomes with probabilities below 2ϵ .¹⁶ And that too will constitute a norm, so we can extend this to outcomes with probabilities below 3ϵ , 4ϵ , and so on all the way up to 1. We may permissibly ignore all outcomes! But this is absurd. So the assumption of tolerance must be false. Precision in probabilities must be required for at least some norms, and it is not clear why this would not include the norm that we may permissibly ignore outcomes with probability 0. After all, the difference between a probability of

¹⁵It also still rules out the problems with which Smith is more concerned: avoiding counterintuitive judgements in the St Petersburg and Pasadena games.

¹⁶Note that Smith's own proposal does not imply this. But then it is not a tolerant theory after all.

0 and one greater than 0 seems categorical, unlike most probability differences.

Along similar lines, take any norm that says that some particular lottery L_a is better than some L_b —perhaps L_a is some moderately risky lottery with probability 0.9 of value 0 and probability 0.1 of a better outcome, and L_b is just some small positive value with certainty. If our decision theory is tolerant, then the norm that L_a is better than L_b must not require precision. We must also accept that we are permitted to choose a lottery with the same outcomes as L_a but with slightly lower probability ($0.1 - \epsilon$) of success. And, as above, we will also be permitted to choose lotteries with *even lower* probability ($0.1 - 2\epsilon$, $0.1 - 3\epsilon$, and so on). We will be permitted to choose these riskier lotteries *even though* they do not have outcomes any more valuable than in L_a . In a way, this result is even more extreme than what Fanaticism tells us—we are sometimes permitted to choose some risky lotteries that even many fanatical theories (such as expected value theory) would forbid. For anyone who finds Fanaticism unacceptable, this too should be hard to accept. By either argument, tolerance of this sort leads to unacceptable conclusions. We should reject it, even if that removes our justification for neglecting small probabilities and thereby rejecting Fanaticism.

6.2.4 The argument from mortality (or YOLO)

A further, more compelling, argument against Fanaticism comes from Monton (2019, p. 14-6).

Before that, a bit of context: a common argument in favour of expected value theory is that maximising expected value is a good policy in the long run (Feller, 1968). Take any decision between some lotteries L_a and L_a , and suppose you faced that decision not once but many, many times in succession, and that the lotteries were run independently each time. Suppose you choose L_a every time. Then, as long as the decision is repeated sufficiently many times, the values you *actually* end up obtaining each time will average to the expected value of L_a , with probability extremely close to 1. Thanks to the weak law of large numbers, the same applies to any lottery, no matter how small the probabilities involved. So it seems a good idea for an agent to adopt the policy of maximising expected value, even when it gives fanatical verdicts—in the (very) long run, the agent will actually do better than under many of the alternatives.

Monton's 'YOLO' argument stands in contrast to the classical long run argument for expected value theory. In practice, agents like us *only live once*, and for a relatively short time. We don't

have arbitrarily many decisions over which to average our gains and losses. And, if we care about how our lives *actually* go, then maximising expected value is not such a good policy. Especially when expected value theory recommends fanatical verdicts, there just won't be enough time in a human life to repeat the decision sufficiently many times to have much chance of getting the astronomical value. A policy that would almost certainly turn out much better is to reject the more fanatical verdicts of expected value theory.

On the face of it, this does seem like good reason to reject fanatical verdicts when making *prudential* decisions. Specifically, when an agent compares lotteries based on how much they improve the agent's own life, it seems quite appropriate for them to care about how their life *actually* goes.

But I am interested in Fanaticism in the *moral* context. And there the argument is less compelling. How well the world as a whole goes is not determined by just a few decisions by a single agent, but instead by countless different agents making separate small-scale moral decisions—indeed, perhaps infinitely many agents. In this setting, having all of those agents maximise expected value seems quite a good policy—as by the long run argument, it would produce greater actual value than many alternative policies. And that includes maximising expected value even when the verdict seems fanatical. Repeated enough times, even fanatical choices will pay off eventually.

Of course, these comments are not a decisive objection to Monton's argument applied in the moral context. It still seems at least somewhat plausible that an agent ought to be concerned with how much moral value they actually produce, rather than whether they adhere to a sensible policy for the entire group of agents. But I hope it also seems at least remotely plausible that Monton's conclusion *doesn't* hold in the moral context, and that Fanaticism may be true. As it was for the previous two arguments, Monton's argument against Fanaticism is not quite rock-solid. Overall, the case against it is wobbly enough to seriously consider whether Fanaticism might be true.

6.3 Background assumptions

In a departure from previous chapters, here I will be considering outcomes with finite total values rather than those containing infinite total value spread across infinitely many locations. This allows for some simplifications in the formalism.

As before, we have some set of epistemically possible outcomes, but we can restrict that set to only those outcomes with finite total value. Denote that set by $\mathcal{O}$. As before, there is an ‘at least as good as’ relation $\succsim_{\mathcal{O}}$ on the set of outcomes, which will be reflexive, transitive, and complete. And this relation can be represented by a cardinal value function $V : \mathcal{O} \rightarrow \mathbb{R}$ which outputs the total value of the outcome. And as a *cardinal* value function, V is unique only up to affine transformations—for any V_1 or V_2 we might use here, $V_2(O) = a \times V_1(O) + b$ for some positive a and real b .

But, as in the previous chapter, what we are really interested in are *lotteries*. As before, I’ll assume that the relevant aspects of a lottery can be fully described by a (real-valued) probability measure L on a set of outcomes.¹⁷ The set of all lotteries on $\mathcal{O}$ is denoted by $\mathcal{L}$.

As before, I’ll use several abbreviations. The probability of a single outcome, $L(\{O\})$, can be shortened to $L(O)$. When a lottery L gives a certainty of an outcome O (when $L(O) = 1$), L can be written as O . To represent a lottery with the same probabilities as L , but outcomes which have precisely k times the value (on the same cardinal representation), I’ll write $k \cdot L$.¹⁸ And I’ll write $L_a + L_b$ to represent the lottery obtained from adding together the values of the outcomes of lotteries L_a and L_b , run independently.¹⁹

Fanatical verdicts, and their rejection, consist in *comparisons* of lotteries. So we need another ‘at least as good relation’. Let $\succsim$ be a binary relation on $\mathcal{L}$. Strict betterness ($\succ$) and equality ($\sim$) are defined as the asymmetric and symmetric components, respectively.

¹⁷As before, this probability can represent either the agent’s subjective degree of confidence or the evidential probability of each outcome arising. Almost all of what follows can be read in terms of either. With minor changes, what follows could also be read in terms of objective chance or whatever notion of probability you consider relevant for decision-making.

¹⁸Formally, define $k \cdot L$ as a probability measure on $\mathcal{O}$ such that, for all $O \in \mathcal{O}$, $k \cdot L(O_k) = L(O)$ for some O_k that has k times the value of O (i.e., $V(O_k) = k \times V(O)$).

¹⁹This can be made more precise. Let V_i denote the random variable corresponding to a given lottery L_i , which outputs the total value of the outcome—equivalently, the probability that V_i takes on a value in the interval $[a, b]$ is given by $L_i(\{O | V(O) \in [a, b]\})$. Then, for any $L_a, L_b \in \mathcal{L}$, define $L_a + L_b$ as some lottery on $\mathcal{O}$ —there may be several—which corresponds to the random variable $(V_a + V_b)$. Equivalently, $L_a + L_b$ is some lottery such that the probability $L_a + L_b(\{O | V(O) \in [c, d]\})$ is always equal to the probability that $(V_a + V_b)$ takes on a value in the interval $[c, d]$.

I won't assume that $\succsim$ is *complete* on $\mathcal{L}$, but I will assume that it is reflexive and transitive. As before, both of these properties are overwhelmingly plausible.

Note also that I will assume throughout that Stochastic Dominance holds. Recall from above that it states the following (where $\mathcal{O}_{\succsim O'}$ denotes the set of outcomes in $\mathcal{O}$ that are at least as good as O').

Stochastic Dominance: For any $L_a, L_b \in \mathcal{L}$, if $L_a(\mathcal{O}_{\succsim O'}) \geq L_b(\mathcal{O}_{\succsim O'})$ for all $O' \in \mathcal{O}$, then $L_a \succsim L_b$.

If, as well, $L_a(\mathcal{O}_{\succsim O'}) > L_b(\mathcal{O}_{\succsim O'})$ for some $O' \in \mathcal{O}$, then $L_a \succ L_b$.

In a setting like this, Stochastic Dominance is fairly uncontroversial among decision theorists, including some who reject expected value theory. As far as I know, no serious proposal has been made in normative decision theory which violates Stochastic Dominance (at least in this setting).²⁰

And Stochastic Dominance ties together the fates of Fanaticism and the fanatical verdict in Dyson's Wager. Recall the lotteries L_{safe} and L_{risky} from the definition of Fanaticism. In L_{safe} , we can set $v = 1$. And define L_{infinite} as follows.

L_{infinite} : an outcome containing infinitely many blissful lives with probability ϵ ; value 0 otherwise

L_{risky} : value V with probability ϵ ; value 0 otherwise

L_{safe} : value 1 with probability 1

L_{infinite} and L_{safe} should look familiar—these match the lotteries you must choose between in Dyson's Wager, if we represent the outcome in which one life is saved with value 1 and the outcome in which no lives are saved with value 0. By Fanaticism, L_{risky} is better than L_{safe} , at least for large enough finite V . And Stochastic Dominance implies that L_{infinite} is better than L_{risky} .²¹

²⁰Schoenfield (2014) rejects a similar principle, but only for lotteries over outcomes for which we don't have a complete ordering. With our complete $\succsim_{\mathcal{O}}$ relation on $\mathcal{O}$, even Schoenfield would endorse Stochastic Dominance here.

²¹The reasoning behind this is as follows. We represented the value of one additional life being saved with a 1 and additional lives saved with a 0. Any finite positive V will hence correspond to V additional lives of equal value being saved (or produced). This implies that the outcome in which infinitely many blissful lives are produced cannot be represented on the same scale, and indeed it is better than any outcome that can be. But, had that outcome been representable on the same scale with some finite number, this wouldn't hold.

6.4 A continuum argument

You might argue for Fanaticism in the following way: expected value theory is true; expected value theory implies Fanaticism; therefore, Fanaticism is true. And likewise for verdicts which seem fanatical, like that which the expected value theorist must accept in Dyson's Wager. Such verdicts are rarely defended on any grounds other than 'That's what expected value theory says'. But we can do better than this. Fanaticism is far weaker than expected value theory in all its strength (and weaker even than expected *utility* theory), so it should be easier to justify.

In this and the next two sections, I'll give four arguments for Fanaticism but not for expected value theory. Here is the first, which originates in Beckstead (2013, p. 149-47) and reappears in Beckstead and Thomas (n.d.,) p. 4-6).

Take the lottery L_0 , which might represent a stranger's life being saved.

L_0 : value 1 with probability 1

And take another lottery L_1 by which vastly more strangers are saved, with a very small probability of failure—perhaps 10^{10} lives saved with probability 0.999999 of success.

L_1 : value 10^{10} with probability 0.999999; value 0 otherwise

Intuitively, L_1 seems better. Accepting a *slightly* lower probability of success for a *vastly* greater payoff seems a good trade. But then consider L_2 , which has a slightly lower probability of success but, if successful, results in many more lives saved.

L_2 : value 10^{10^2} with probability 0.999999²; value 0 otherwise

This seems better than L_1 , or at least it seems that there is some number of lives high enough that it would be better. And so we could continue, with L_3, L_4 , and so on until some L_n , such that 0.999999^n is less than ϵ , for any arbitrarily small ϵ you want. Intuition suggests that *vastly* increasing the payoff can compensate for a *slightly* lower probability; that each lottery in the sequence is better than the one that precedes it. So the final lottery in the sequence must be better than the first. But the final lottery has a probability less than ϵ of any positive payoff at all. So we have Fanaticism.²²

²²Even for readers who reject aggregative theories of betterness, much the same argument can be made in terms of averageism, pure egalitarianism, maximin, maximax, and narrow person-affecting views. For instance, if we assume averageism, we can start with L_0 , by which the *average* value of all lives will be 1

As I see it, this argument rests on two intuitively plausible principles: the first, the transitivity of $\succ$; the second, *Minimal Tradeoffs*.

Minimal Tradeoffs: There is some real $r < 1$ such that, for any real value v and any probability p , there is some real r^* such that, on any cardinal representation, the lottery represented by L_a is better than that represented by L_b (as defined below).

L_a : value v with probability p ; 0 otherwise

L_b : value $r^* \times v$ with probability $r \times p$; 0 otherwise

This principle has intuitive force. For any given lottery with the form of L_a , *surely* there is any lottery with only that ever-so-slightly lower probability of success that is better. Intuition suggests that we can always make at least some tradeoff between probability and value: we can always compensate for a, perhaps, 0.000001% lower probability of success with a vastly greater payoff.

But you might be unconvinced by the continuum argument. You may find it far more intuitively plausible that Fanaticism is false than that Minimal Tradeoffs is true. And, other than clashing with intuition, I see nothing obviously problematic about denying Minimal Tradeoffs. Perhaps we should just reject it; perhaps there is at least one threshold of probability p' between p and 0 (which might be unknown or vague) at which we have a discontinuity, such that *no* value with probability below p' is better than *any* value with probability above p' . One way to do this is by adopting expected utility theory with a bounded utility function which, if the bound is low enough, would imply that the former lottery is better than the latter. But, whichever approach we use to reach that conclusion, there is some point at which we can no longer trade off probability against value, no matter how great the value. And what of it? It seems a little counterintuitive, but perhaps less so than Fanaticism.

If you accept Minimal Tradeoffs then you must accept Fanaticism. And, conversely, to reject Fanaticism you must also reject Minimal Tradeoffs and the continuum argument it generates. But you might think that a small price to pay.

with near certainty, and 0 otherwise. L_1 can be the lottery which gives probability 0.99999 of an outcome in which the average value of all lives will be 10^{10} , or 0 otherwise. We can rapidly scale up the average value of lives just as we scaled up the total values above, while gradually scaling down the probability of non-zero value. Eventually, we reach a lottery L_n with probability ϵ of some astronomical average value, or 0 otherwise, but transitivity says that it must be better than L_0 .

6.5 A dilemma for the unfanatical

The next argument for Fanaticism comes in the form of a nasty dilemma that you face if you deny it. In brief, if you deny Fanaticism then you must accept at least one of: abandoning *Scale Independence*, as defined below; and allowing comparisons of lotteries to be *absurdly sensitive* to tiny changes (even more so than if you accepted Fanaticism).

*Scale Independence*²³: For any lotteries L_a and L_b , if $L_a \succcurlyeq L_b$, then $k \cdot L_a \succcurlyeq k \cdot L_b$ for any positive, real k .

I find Scale Independence highly plausible.²⁴ Take any pair of lotteries, perhaps a pair that can be cardinally represented by L_1 and L_2 .²⁵

L_1 : value 1 with probability 1

L_2 : value 2 with probability 0.9; value 0 otherwise

I'll remain agnostic on which is better; $\succcurlyeq$ might recommend either (or neither). But suppose we *scaled up* each lottery by 2, to obtain $2 \cdot L_1$ and $2 \cdot L_2$ —lotteries with the same probabilities as above, but with the value of each outcome doubled (on the same cardinal scale). If we ranked L_1 as better then, intuitively, that ranking shouldn't change as we double the values. After all, the scaled-up lotteries have the same structure as the originals. Whatever general principles lead us to say whether L_1 is better than L_2 should plausibly also lead us to say the same about $k \cdot L_1$ and $k \cdot L_2$, for any positive k you like.

It would be rational to deviate from Scale Independence if the values we were dealing with were dollars, or apples, or some other commodity that is merely *instrumentally* valuable—it would make sense for your first dollar, or apple, to be far more valuable than your hundredth. But when we are dealing with moral value itself, it is incoherent to say that additional units of value are worth less and less—by definition, adding one unit of value is always an improvement of one

²³Ramsey (1964, p. 182-3) endorses a similar principle, which he simply calls "...a certain measure of consistency...".

²⁴One way to violate it is to adopt expected utility theory with a bounded utility function (see Rabin, 2000). I find this good reason to reject such a form of expected utility theory. On the other hand, there are also theories that avoid Fanaticism without a bounded utility function and hence scale dependence, e.g., the de Buffon (1777) and Smith (2014) methods of ignoring outcomes with sufficiently low probabilities, and a similar proposal from Buchak (2013, pp. 73-4).

²⁵For readers unsympathetic to totalism, these payoffs could just as easily represent: the average values of lives in each outcome; the priority-weighted total value in each outcome; the total value in the lives of all persons who would exist regardless of the agent's choice; and so on.

unit of value. And under any aggregative theory of value, at least some version of which I'm assuming here to be true, units of value can correspond to tangible objects such as (identical) human lives. By the lights of aggregative theories and by at least my own intuition, an additional (identical) human life is always worth the same amount no matter how many lives already exist—and of course it always contributes precisely the same amount to the total value of the outcome. So, in keeping with intuition, it should not matter in the slightest whether we are comparing the lotteries L_1 and L_2 above or instead some scaled-up multiples $k \cdot L_1$ and $k \cdot L_2$. In the latter pair, we have done the equivalent of adding $k - 1$ additional copies to the contents of each outcome (perhaps $k - 1$ additional lives to L_1 and $2k - 2$ additional lives to that outcome in L_2). But each additional copy should be no more or less valuable than the first; they should contribute the same to our evaluation of the lottery.²⁶ So, I would claim, we should rank the resulting lotteries just the same way as we ranked them without those $k - 1$ copies added in.

A further reason to accept Scale Independence here is that most aggregative theories that have been proposed in the finite context (including total utilitarianism) provide only that value may be represented cardinally, not that it can be represented on a ratio or unit scale.²⁷ They typically do not provide any absolute zero of value, nor any unique scale on which to represent it. A pair of outcomes represented with values 1 and 2 on one scale can just as easily be represented with any values k and $2k$ (for positive k) on other scales. When evaluating those outcomes aggregatively, we cannot say that any such representation is more valid than any other. Suppose we took lotteries L_1 and L_2 above; their outcomes could just as easily have been represented on another scale such that they had values k , $2k$, and 0—the same values as $k \cdot L_1$ and $k \cdot L_2$ had on the previous scale. On that new scale, we of course still say that L_1 is better than L_2 if and only if we said so on the original scale, else we would be inconsistent. But why not judge $k \cdot L_1$ and $k \cdot L_2$ similarly? For us to be able not to, our evaluations of lotteries must depend on more than just the probabilities of outcomes and cardinal values that our aggregative theory assigns

²⁶A similar story can be told for averageism and other theories of betterness, although a little more awkwardly. On plausible versions of averageism, an additional period of bliss given to everyone contributes the same amount to the average value of all lives no matter how many such periods of bliss have already been experienced. We can copy the experiences each person has in much the same way we can copy the number of additional lives in the totalist scenario. It seems plausible then that we should rank lotteries with $k - 1$ copies of each experience added into outcome just the same way as we rank them without those copies.

²⁷This holds not only for aggregative theories but also for averageism, maximin, maximax, and narrow person-affecting views.

them; they must also depend on *which* cardinal representation is being used when those values are assigned. In effect, they depend on richer details of the value of outcomes than is provided by the cardinal representation given by the aggregative theory. And this makes for a troubling shift in how we evaluate lotteries—doing so is no longer a matter of using our axiology to assign values to outcomes and then using our theory of instrumental rationality to turn those values and each outcome’s probabilities into an evaluation. Instead, our evaluation is sensitive to facts about the values of outcomes that even our theory of value is not sensitive to. That is awfully strange and, I think, absurd. We should be able to apply our theory of rationality to the verdicts of our axiology without ‘double-dipping’ into facts about the value of outcomes. And so we should reject the suggestion that $k \cdot L_1$ and $k \cdot L_2$ not be compared in a similar way to L_1 and L_2 ; we should accept Scale Independence.²⁸

For those who reject Fanaticism, the dilemma they face is between rejecting Scale Independence as well and accepting an absurd level of sensitivity to tiny changes in lotteries. I’ll illustrate below what I mean by ‘sensitivity to tiny changes’ but, for now, be assured that it’s just as implausible as violating Scale Independence.

The argument for the dilemma goes like this. First, for Fanaticism to be false, there must be some probability $\epsilon > 0$ and some cardinal value v such that some lottery L_{risky} as defined below is no better than L_{safe} , no matter how big V is.

L_{risky} : value V with probability ϵ ; 0 otherwise

L_{safe} : value v with probability 1

But what about values below v ? How would getting some lower value with certainty compare to an ϵ -probability of arbitrarily enormous V ? There are only two distinct possibilities (at least if we maintain Stochastic Dominance and transitivity). The first is: there is *some* smaller $v' > 0$ for which L_{risky} is better than a sure outcome with value of v' , for sufficiently large V . The second is: for *any* such v' , L_{risky} *isn't* better than any sure outcome with value v' , no matter how large V is.²⁹ As we’ll see, each possibility impales us on a respective horn of the dilemma.

²⁸An alternative to endorsing Scale Independence is to adopt an aggregative theory that assigns more than just a cardinal value to outcomes—perhaps a version that values outcomes on a ratio or unit scale. That version would be far less parsimonious, and I cannot think of any compelling independent justification for it. It seems at least a considerable cost to adopt such a version to avoid endorsing Scale Independence and the implication of Fanaticism.

²⁹This difference is analogous to the difference between weak and strong superiority in Arrhenius and

Assume that the first possibility holds. L_{risky} is better than a sure outcome with value v' , for some $0 < v' < v$ and some large V . Still, L_{risky} won't be better than a sure outcome with value v —i.e., L_{safe} —or any sure outcome with even greater value (by Stochastic Dominance). But somewhere below v this changes—for some positive v' , the same doesn't hold.

Since that v' is a positive real number, we know that there will be some real k such that $k \times v' \geq v$. And that means that a sure outcome with value $k \times v'$ is no worse than L_1 , no matter how great V is. But here is the problem: a certainty of $k \times v'$ is the scaled-up counterpart of v' ; and the counterpart of L_{risky} , which is $k \cdot L_{\text{risky}}$, is just L_{risky} with V replaced by another value which is k times higher. That still has the same form as L_{risky} . So it cannot be better than the sure outcome, not if we reject Fanaticism. But then our verdict for v' versus L_{risky} doesn't match our verdict for $k \times v'$ versus $k \cdot L_{\text{risky}}$. So we violate Scale Independence.

But there was a second possibility: that there is no such $v' > 0$; that L_{risky} is no better than a certainty of *any* $v' > 0$. If so, we can avoid the problem of scale dependence: there's no inconsistency between the judgements for v and judgements for any 'scaled-down' counterpart v' , since *all* such v' compare the same way to lotteries like L_{risky} .³⁰ But we then face another serious problem.

Take the probabilities ϵ and 1. We can give a sequence of increasing probabilities in between ϵ and 1, spaced evenly apart and as finely as we want: $\epsilon < p_1 < p_2 < \dots < p_n < 1$. By assumption, *no* amount of value with probability ϵ (and value 0 otherwise) is better than *any* (positive) amount of value with probability 1. But then there must be some pair of successive p_i, p_{i+1} such that no amount of value at p_i is better than any amount of value with probability p_{i+1} . If there were no such pair, we could give a sequence of better and better, and less and less likely, lotteries much like that in the previous section.

Crucially, that p_i and p_{i+1} can be *arbitrarily* close together, since the same result holds no matter how finely spaced the sequence was. No amount of value at p_i is better than *any* amount of value at p_{i+1} . We could have some *astronomical* value at p_i , and an imperceptibly small amount of value at p_{i+1} , and the latter would still be better. So we have a radical discontinuity in how we value lotteries.

Rabinowicz (2015).

³⁰The comparison won't be the same for values 0 and below, but that's okay. Such values cannot be scaled up to v ; multiply 0 or any negative number by any real positive value you want and you still won't reach v .

This will make our judgements of betterness *absurdly sensitive* to tiny changes in probability. To illustrate that sensitivity, consider the following four lotteries. Here, $\epsilon' > 0$ is some arbitrarily small number. And, again, p_i and p_{i+1} are some tiny probabilities that are arbitrarily close together.

L_0 : value 0 with probability 1

L_1 : value $10^{10^{10}}$ with probability p_i ; 0 otherwise

L_2 : value ϵ' with probability p_{i+1} ; 0 otherwise

L_3 : value $10^{10^{10}}$ with probability p_{i+1} ; 0 otherwise

By the above, if we reject Fanaticism and maintain Scale Independence, then we are forced to rank these lotteries as follows: $L_0 \prec L_1 \prec L_3$ and $L_0 \prec L_2 \prec L_3$, but $L_1 \not\prec L_2$. And that's a peculiar ranking. L_1 and L_3 are almost indistinguishable; their probabilities may differ by an arbitrarily tiny amount. Likewise for L_0 and L_2 , except it's their payoffs that differ slightly. For each pair, we must say that one is better than the other, but does not seem *much* better—in a sense, the better lottery is only a trivial improvement over the other, whether by a slight increase in payoff or in probability. But, despite appearances, we must accept that L_3 is *much* better than L_1 , as well as that L_2 is much better than L_0 . How so? Between L_1 and L_3 comes L_2 . The lottery L_2 has an *astronomically* smaller payoff than L_3 and so (in a sense) is *vastly* worse than L_3 , and yet we cannot say that it is so bad so as to be worse than L_1 . Effectively, we must accept that, starting from L_3 , it is no worse to lose almost all of the potential value via L_2 than it is to lose that sliver of probability via L_1 . And similarly for L_0 and L_2 : between them must come L_1 . L_1 has astronomically greater potential payoff than L_0 (and much higher probability of positive payoff too), yet we cannot say it's better than L_2 . Effectively, we cannot say it is any better to gain that enormous value with probability p_i than it is to gain an ever-so-slightly more probable shot at tiny value ϵ . So those lotteries are not just made worse by those tiny changes in probability or payoffs—they are made *far* worse (in an intuitive sense). That is the level of sensitivity we must accept if we reject Fanaticism and maintain Scale Independence. But such extreme sensitivity in our evaluations of lotteries seems absurd. Evaluations of lotteries should not change quite so wildly—so discontinuously—with arbitrarily tiny changes in probability or payoff.³¹ So I do not

³¹One might object here, along lines similar to Smith (see §6.2.2), that accepting Fanaticism results in

think this sensitivity—this horn of the dilemma—is any less absurd than what we faced earlier.

I should briefly note that, beyond its intuitive implausibility, this sensitivity will likely lead to practical difficulties. The probabilities that human agents have access to are subjective degrees of belief and evidential probabilities. In practice, we are often unable to pin down either sort of probability with arbitrary precision, at least not with our merely finite capacity for calculation.³² But we'll need to be arbitrarily precise when it comes to whether an outcome has probability p_i or the slightly greater p_{i+1} to be able to compare lotteries, at least by this horn of the dilemma. If not, in cases like that immediately above, we would have no idea whether a lottery is as good as L_3 or as bad as L_1 . We must either require (unrealistically) that agents be arbitrarily precise in their probabilities³³, or else accept that in practice our decision theory will be unable to tell agents that (at least some of) these lotteries are better than others.

This is a dilemma we must face if we reject Fanaticism: we must accept either scale dependence, by which our judgements of lotteries can vary without any structural reasons for doing so, or that those judgements are sensitive to imperceptibly small differences in either probability and value. Both options seem absurd to me, indeed more absurd than Fanaticism itself.

similar sensitivity—for outcomes of sufficient value, varying their probability between 0 and some tiny ϵ in some lottery can radically change the instrumental value of the lottery. But I think this objection has less intuitive force than the above. Changing an outcome's probability from 0 to something other than 0 is a categorical change, while changing it from p_i to p_{i+1} is a mere difference in degree. It seems to me less worrying for the former to be able to radically change a lottery's value than for the latter to do so. But, even if you find both forms of sensitivity troubling, the discussion above at least demonstrates that rejecting Fanaticism (at least by this horn of the dilemma) brings little respite from problems of absurd sensitivity or intolerance.

³²For discussion of why we often cannot require epistemic agents to settle on precise probabilities, see Schoenfield (2012) and Joyce (2010). For discussion of how to evaluate lotteries with imprecise probabilities, to which much of the discussion in this chapter can be translated, see Seidenfeld (2004), Huntley, Hable & Troffaes (2014), Bradley & Steele (2015), and Bradley (2015). On any of the proposals therein, agents with access only to imprecise probabilities will face serious practical problems if rankings are sensitive to the degree described above. On so-called 'liberal' proposals (see Joyce 2010: 314), agents will be indifferent among L_1 , L_2 , and L_3 . And on 'conservative' or 'maximal' proposals (see *ibid*; Bradley & Steele 2015), agents must accept that the rankings of those lotteries are indeterminate. And similar indifference or indeterminacy will often arise when agents decide among lotteries with only small differences in probabilities.

³³One might object that accepting Fanaticism will also sometimes require agents to pin down probabilities with great precision, whenever outcomes have probabilities very close to 0. But it will never require *arbitrary* precision. When a lottery features an outcome with a given value, there will generally still be some non-zero probability below which that outcome can be ignored. (The only exception is when that lottery and the others in the agent's option set are then equally good.) But the sensitivity identified above will require agents to distinguish between probabilities p_i and p_{i+1} , no matter how arbitrarily small the difference between them may be.

6.6 Egyptology and Indology

But it gets worse. There are even greater costs you must pay to reject Fanaticism: you face either one or both of the following objections, each of which is intuitively absurd.

6.6.1 The Egyptology Objection

Hailing from population ethics, the *Egyptology Objection* is a classic argument against various axiologies, including averageism, egalitarianism, and maximin.³⁴

As a brief refresher, it goes like this. Suppose you are making a moral decision here and now in the 21st Century, and the available outcomes differ *only* by some small-scale changes in the very near future. Your actions will not change what happens in distant galaxies or what happened in, say, ancient Egypt. Then, intuitively, what you ought to do should depend only on the events altered by your choice. It should not depend on events in distant galaxies or in ancient Egypt, at least not if your actions do not change them. Intuitively, it seems absurd that those unaltered, remote events make any difference to your evaluation.³⁵

But, under some theories of value—thankfully not aggregative theories—what you ought to do in such cases *is* dependent on unaltered, remote events. Take a (standard, welfarist) averageist view. Here and now, should you bring an additional person into existence? It depends—will that additional life raise or lower the average value of lives that ever exist? Sometimes yes, sometimes no, depending on what actually happened in ancient Egypt, in distant galaxies, and everywhere else. Averageism will thus imply that what happened in ancient Egypt can affect whether it is better to now bring an additional person into existence or not. And, further, to be confident of which is the better outcome, you may need to do some serious research into Egyptology (along with plenty of other historical studies). But this is implausible. By intuition, we can ignore events that are unaffected by our actions and took place millennia ago, as we do under aggregative theories like totalism, critical-level views, prioritarianism, and aggregative person-affecting views (and any other axiology which admits an additively separable representation). To me, this is one of the main appeals of those views.

³⁴It appears earliest in McMahan (1981, p. 115) but is often attributed to Parfit (1984, p. 420).

³⁵Such dependence brings practical problems too. If what you ought to do is dependent on such events—not just in the remote future but also the remote past, of which we are often clueless—then we have little guidance for what we ought to do (*cf* (Mogensen, 2021)). This seems a major disadvantage for a moral view.

But even if we accept an aggregative theory of value, we may still face much the same objection in practice.³⁶ Some methods of comparing lotteries give rise to an updated Egyptology Objection, including all of those that deny Fanaticism. That's right: to deny Fanaticism, you must not only endure the costs detailed in the previous two sections, you must also fall prey to a version of the Egyptology Objection. (As we'll see below, you may also face another even more serious objection.)

You will most easily fall prey to it if you reject *Background Independence*. (And some proposals do.)³⁷

Background Independence: For any lotteries L_a and L_b and any outcome O , if $L_a \succcurlyeq L_b$, then $L_a + O \succcurlyeq L_b + O$.

Recall that the sum of two lotteries $L + O$ is simply the lottery you get if you run each lottery independently and sum up the values of their outcomes. But, here, O is a very simple lottery—it is just the outcome O with certainty. That outcome has some cardinal value; call it b . The lottery $L_a + O$ is simply the lottery L_a with value b added to the value of each outcome. Background Independence implies that we can take any two lotteries and add any constant value b to each of their possible outcomes, and this won't change their ranking.

Suppose for now that Background Independence is false. Then there is some pair of lotteries, L_a and L_b , such that: L_a is at least as good as L_b ; and, if we added a certain constant b to the value of every one of their outcomes, that would change their ranking. Since we're assuming an aggregative theory, the value of each outcome in each lottery must represent the total aggregate of value that would result, including all valuable events across all of space and time. That includes the events that occurred in ancient Egypt. And we can assume that the same events occurred in ancient Egypt in every outcome of either L_a or L_b .³⁸ With those events included, L_a is at least as good as L_b .

³⁶Even on other axiologies, including averageism, we face similar (albeit slightly less compelling) objections. See Footnote 35.

³⁷One such proposal is expected utility theory with a utility function that is concave and/or bounded (e.g. Arrow, 1971). As Beckstead and Thomas (n.d., p. 15-6) point out, this results in comparisons of lotteries being strangely dependent on events that are unaltered in every outcome and indeed some irrelevant to the comparison.

³⁸If there is some pair of lotteries L_a or L_b that violates Background Independence, then there is also some pair that violates it and have the exact same events occur in ancient Egypt. To obtain such a pair, simply take L_a and L_b and replace each of their outcomes with an outcome of equal total value but an identical history of ancient Egypt.

But what if we consider a different hypothetical pair of lotteries, identical to L_a and L_b *except* that the events of ancient Egypt were different? Would the ranking be any different, had events gone differently back then? Yes, it would, if Background Independence is false and if those events differed drastically enough. They might differ in such a way that they increased in value by b . Then the total value of every outcome would also be increased by b . And, given that L_a and L_b were the lotteries that violated Background Independence, we know that this changes their ranking. So our modified version of L_a is no longer at least as good as the modified version of L_b . Thus, if we deny Background Independence, we face a rather severe version of the Egyptology Objection: when choosing between two lotteries, which is better can vary depending on what events occurred in ancient Egypt, *even if* the choice doesn't change those events at all (*cf* Beckstead and Thomas n.d.).³⁹

To avoid this implication, we must at a minimum accept Background Independence. I think we ought to accept it regardless—I find it highly plausible for much the same reasons as Scale Independence was. In any case, I'll assume for the remainder of this section that Background Independence holds.

But even if we accept Background Independence, if we still reject *Fanaticism* then it turns out that a version of the Egyptology Objection still arises, albeit a less severe one.

To see why, first recall the lotteries from the definition of Fanaticism: L_{risky} and L_{safe} . If we reject Fanaticism then, for some tiny enough probability $\epsilon > 0$ and some cardinal value v , some lottery L_{risky} (as defined below) is no better than L_{safe} , no matter how big V is.

L_{risky} : value V with probability ϵ ; 0 otherwise

L_{safe} : value v with probability 1

³⁹This new version of the Egyptology Objection, along with the other objections described in the rest of this section, can also be applied under non-aggregative theories of betterness as well. It will just look a little different.

Suppose that (standard, welfarist) averageism is correct. Then to add b to value of an outcome would be to increase the average value obtained by all persons in that outcome. And to add b to the value of every outcome in a lottery L would be to increase the average value in every outcome by b . We might imagine doing this by delivering a gift to every inhabitant of the outcomes of L , with that gift producing the same boost in value, b , for each recipient. Even by the lights of averageism it seems that, if L_a is at least as good as L_b , then delivering that gift to everyone shouldn't change the ranking—a modified L_a with additional value b should still be at least as good as L_b with additional value b . And so we will have an averageist analogue of the Egyptology Objection presented above. And this objection will still be worrying for averageists, even if it is not quite as devastating as its analogue is for aggregationists.

According to Background Independence, it must hold that the lottery L_{risky} with any constant b added to every outcome's value is also no better than L_{safe} with that same b added to every one of its outcomes' values. If we face a decision between L_{risky} and L_{safe} , it would not matter if events in ancient Egypt were vastly different—we'd compare the lotteries the same way with or without that additional value b .

But consider a further pair of lotteries: $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$. These are obtained by modifying L_{risky} and L_{safe} such that, in all outcomes, the events of ancient Egypt are different but you are *uncertain* of exactly how different they are. We are no longer dealing with a fixed value b of which you are certain. We are still adding the same value to all of them, but you are uncertain of what value it is. And that uncertainty is described by the lottery B . To make these lotteries more concrete, perhaps you must decide between two lotteries in which the value of all present and future events match the lotteries L_{risky} and L_{safe} , but you are uncertain of the value of past events such as those in ancient Egypt. As a devout aggregationist, you must evaluate outcomes and lotteries based on the total aggregate of value across all of space and time. And even if you know that those past events will turn out the same way no matter what you do, your uncertainty over exactly how they would turn out is a part of your uncertainty over the total value of the outcome. So you must compare $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$; you cannot simply compare L_{risky} and L_{safe} .

As it turns out, there are some possible lotteries B that lead us to say that $L_{\text{risky}} + B$ is strictly better than $L_{\text{safe}} + B$, even though L_{risky} isn't any better than L_{safe} . And this judgement, that $L_{\text{risky}} + B$ is better, is implied by even the extremely weak principle of Stochastic Dominance introduced above.⁴⁰

As a brief refresher, recall that Stochastic Dominance simply states that: if a lottery L_a gives at least as high a probability as L_b of resulting in an outcome which is at least as good as O , for every possible outcome O , then L_a is at least as good a lottery; and if L_a also gives a *strictly greater* probability of turning out at least as good as some O , then it is *strictly better*.

⁴⁰We can reach a similar result, and thus a similar *reductio*, with most non-aggregative theories of betterness. Following from Footnote 35, we can adopt averageism and, instead of improving/worsening events in ancient Egypt, distributing an identical and morally valuable gift to each person in the world. You might be uncertain of the exact value of the gift—your uncertainty of its value might correspond to B . But still it seems that whether you distribute that gift is irrelevant to whether the risky lottery is better or worse than the safe one.

Stochastic dominance is easy to spot graphically. To illustrate, consider the cumulative probability graphs of the following two lotteries, L_1 and L_2 .

L_1 : value 1 with probability $\frac{1}{2}$; value 0 otherwise

L_2 : value 2 with probability $\frac{1}{3}$; value 1 with probability $\frac{1}{3}$; value 0 otherwise

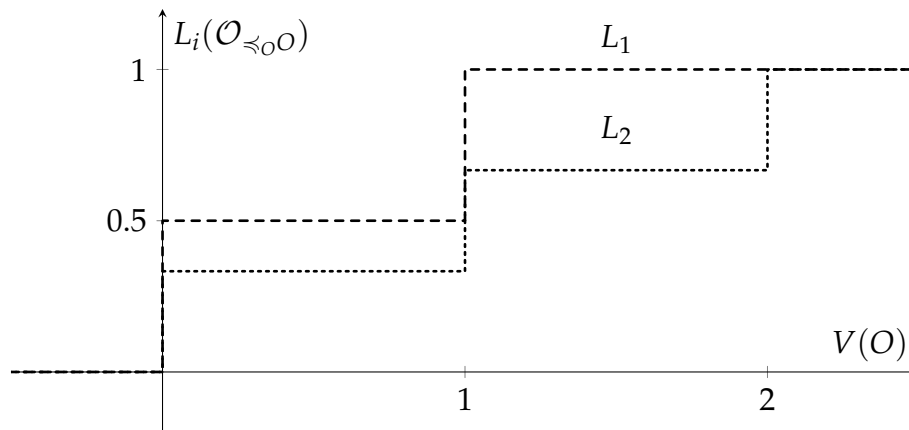


Figure 6.1: Cumulative probability graphs for L_1 and L_2 ; here and below, $\mathcal{O}_{\preceq_O O}$ denotes the set of outcomes in $\mathcal{O}$ as which O is at least as good.

On a graph like this, one can easily see when Stochastic Dominance says that $L_2 \succcurlyeq L_1$. Cumulative probability, on the vertical axis, is just the probability that the lottery produces an outcome *no better* than an outcome with some particular value. Meanwhile, Stochastic Dominance says that one lottery is at least as good as another if its probability of producing an outcome *as good or better* is just as high for all outcomes—or, equivalently, if the probability of an outcome *as good or worse* is at least as low. So Stochastic Dominance will say that $L_2 \succcurlyeq L_1$ if and only if L_2 's cumulative probability is always as low or lower than that of L_1 , as it is on this graph. And, here, L_2 often has *strictly* lower cumulative probability. So Stochastic Dominance says it is strictly better than L_1 .⁴¹

But accepting Stochastic Dominance doesn't rule out denying Fanaticism. L_{risky} doesn't

⁴¹This relationship between Stochastic Dominance and cumulative probability would break down if some outcomes were incomparable to others. There would then be a difference between $O_a \succcurlyeq_O O_b$ and the negation of $O_a \prec_O O_b$. But, fortunately, aggregative theories give us a total preorder of $\mathcal{O}$, so we can sidestep this complexity.

stochastically dominate L_{safe} , as illustrated below.

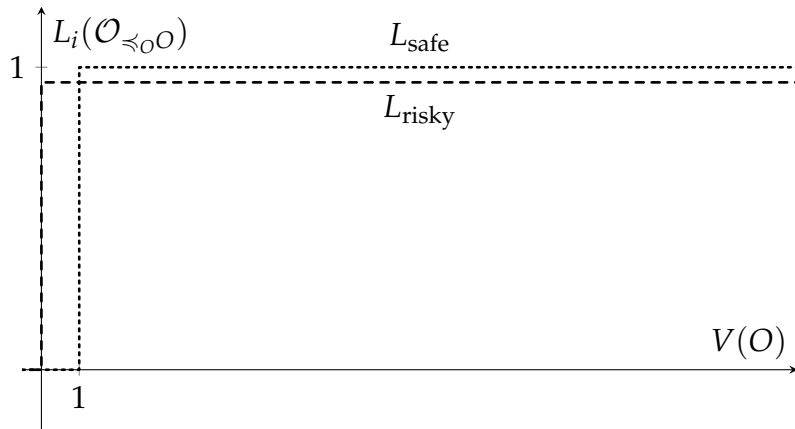


Figure 6.2: Cumulative probability graphs for L_{risky} and L_{safe}

Sometimes one is higher; sometimes the other. So Stochastic Dominance remains silent. And it's a good thing it does—to deny Fanaticism, we must not say that L_{safe} is better than L_{risky} .

But what about $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$? Stochastic Dominance is not necessarily silent about that comparison. In particular, suppose that the lottery B looks like this. (I'll describe the required properties of B below.)

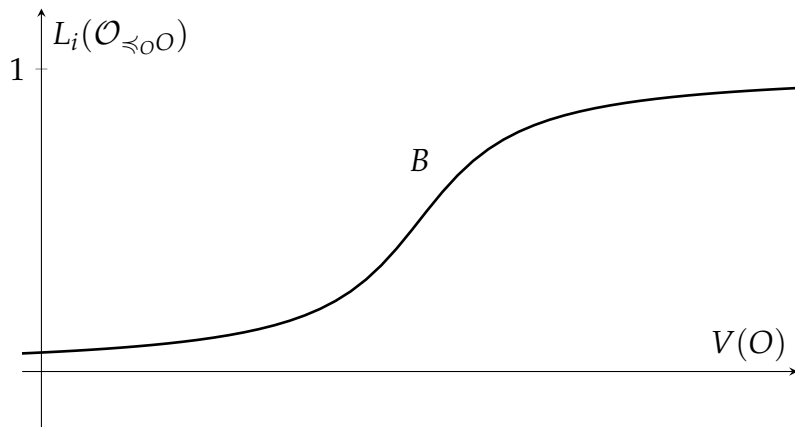


Figure 6.3: An example of background uncertainty B —a Cauchy distribution

When B looks like that, we obtain the following graphs for $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$. Crucially,

the graph for $L_{\text{safe}} + B$ is never higher than that for $L_{\text{risky}} + B$, and sometimes it is strictly lower. So Stochastic Dominance says that $L_{\text{safe}} + B$ is strictly better.

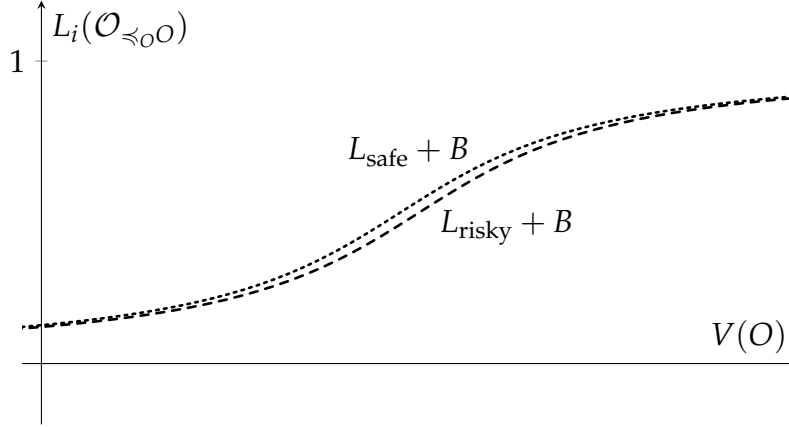


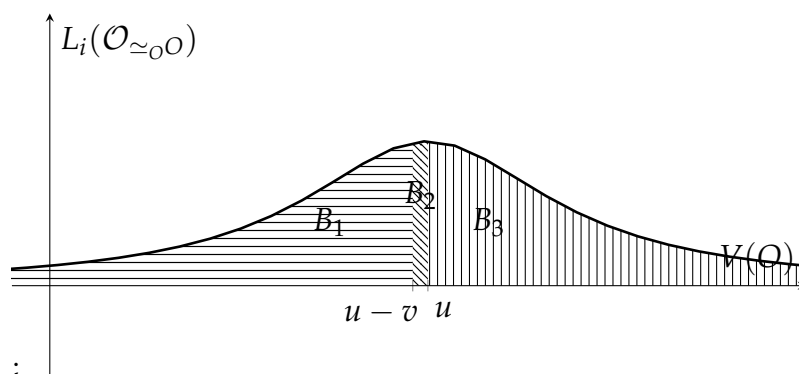
Figure 6.4: Cumulative probability graphs for $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$

But how does this happen?

For Stochastic Dominance to hold, we need $L_{\text{risky}} + B$ to have at least as high a probability as $L_{\text{safe}} + B$ of turning out *at least as good as* an outcome with value u , for *all* possible values u . So take any real value $u < V$. What's each lottery's probability of doing at least that well?

Start with $L_{\text{risky}} + B$. We know that L_{safe} just gives one value (v) with certainty. So the probability that $L_{\text{safe}} + B$ gives value u or better is just the probability that B gives value $u - v$ or better. (This corresponds to the area $B_2 + B_3$ on the graph below.) And then, for $L_{\text{risky}} + B$, we'll get value at least u either if L_{risky} gives value V or if B gives at least value u . Denote the probability that B gives value at least u by B_3 (corresponding to that area on the graph below). Then the probability that $L_{\text{risky}} + B$ turns out at least as good as value u is $\epsilon + B_3(1 - \epsilon)$. And, with a bit of simple arithmetic⁴², we can see that this will be greater than the corresponding probability for $L_{\text{safe}} + B$ if the area B_2 is no greater than $\epsilon \times B_1$.

⁴² $\epsilon + B_3(1 - \epsilon) \geq B_2 + B_3 \Leftrightarrow B_3 + \epsilon(1 - B_3) \geq B_2 + B_3$
 $\Leftrightarrow \epsilon(B_1 + B_2) \geq B_2$
 $\Leftrightarrow B_2 \leq \epsilon B_1 + \epsilon B_2$, for which $B_2 \leq \epsilon B_1$ is a sufficient condition.

Figure 6.5: Probability distribution of B

Here we have a probability graph of one possible B (importantly, not a *cumulative* probability graph), with areas B_1 , B_2 , and B_3 corresponding to the areas mentioned in the last paragraph. If B_2 is small enough compared to B_1 , then $L_{\text{risky}} + B$ has at least as high a probability of u or better as $L_{\text{safe}} + B$ does. And for it to be at least as high for *all* real u , we just need there to be a tiny enough area in the interval between u and $u - v$. And, for that to happen, we just need B to go down slowly enough as we approach $-\infty$, and rise and fall quickly enough as we pass the peak of the curve.

And there are some probability distributions that have this property, including many Cauchy distributions.⁴³ And some is enough. There is some such lottery $L_{\text{risky}} + B$ that is better than $L_{\text{safe}} + B$, even though L_{risky} is no better than L_{safe} .

And so we have a new version of the Egyptology Objection, which I think is slightly milder than the previous version. You may be faced with a decision between two lotteries that, if you exclude the value of events that occurred in ancient Egypt, correspond to L_{risky} and L_{safe} . And which lottery is better overall? That depends on what happened in ancient Egypt, even though you know the same events will have happened there no matter which you choose. If the events of ancient Egypt have value 0 (or b), then the risky lottery is no better than the safe one. But if ancient Egypt might have contained greater or lesser value, and you aren't *certain* how much, then it may be that the risky lottery is better after all. Much like under the classic Egyptology Objection,

⁴³The result holds for any Cauchy distribution with a 'scale factor' of at least $\frac{v}{\epsilon}$. See Tarsney (n.d., §5) and Pomatto et al. (2018).

your evaluation is sensitive to facts that seem irrelevant. But, unlike the original version, the evaluation is not merely sensitive to what *actually* occurred there; instead, it is sensitive just to your *uncertainty* about what happened there. That is perhaps less devastating a problem, as an agent may know enough about events in ancient Egypt to constrain B to a less troublesome distribution. Still, for judgements to be sensitive to the agent's beliefs about events in ancient Egypt *at all* is still, I think, rather absurd—absurd enough that we should be willing to accept Fanaticism to avoid it.⁴⁴

6.6.2 The Indology Objection

But it gets worse. If you reject Fanaticism, you may also face what I'll call the *Indology Objection*.

For context, Indology is the study of the history and culture of India. In particular, classical Indology includes the study of the Indus Valley Civilisation, a Bronze Age civilisation which lasted from 3300 BCE to 1300 BCE. Although less well-known and understood than its contemporaries—ancient Egypt and Mesopotamia—it spanned a greater area than either and rivalled them in population, technology, and urban infrastructure (see (Wright, 2009)). We happen to know even less about what happened in the ancient Indus Valley than in ancient Egypt—archaeological research and excavations of key sites in India began centuries later than similar work in Britain, Italy, and Egypt. So there is likely plenty left to learn in Indology.

Take the same $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$ from above—lotteries such that the former is strictly better, and yet L_{risky} is no better than L_{safe} . Such lotteries must exist, if Fanaticism is false. But suppose that you are no longer uncertain about the events of ancient Egypt; instead, you are uncertain of what occurred in the ancient Indus Valley. The lottery B in $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$ represents only those events.

But, in Indology, there is a great deal more work that can be done! Had you the option, you could spend many years researching the ancient Indus Valley, peeling back that uncertainty and narrowing B down. Let's suppose that with enough years of intensive research you could

⁴⁴This version of the Egyptology Objection loosely resembles one which has been levelled against decision theories that violate the Sure-Thing Principle (by, e.g., (Briggs, 2015)). It turns out that following such a decision theory in each of several consecutive, independent decisions will sometimes be inconsistent with following the theory when deciding on a strategy for all of those decisions at once. Essentially, the combination of two lotteries can be evaluated quite differently to how we evaluate those lotteries separately. And that is troubling. It is unclear whether we should evaluate such decisions together or apart, globally or individually. But the correct verdict depends on which way we do it.

eventually remove that uncertainty entirely and determine exactly which value to assign to the events of the ancient Indus Valley—which value b the lottery B actually results in. If you did first do that research, you would no longer need to choose between $L_{\text{risky}} + B$ and $L_{\text{safe}} + B$. Instead, you would choose between L_{risky} with b added to the value of every outcome and L_{safe} with b added.

Recall Background Independence from above. If it holds, then L_{risky} with any b added to every outcome is better than L_{safe} with the same b added if and *only if* L_{risky} is better than L_{safe} . And if it doesn't, as we saw above, we face an even more severe version of the Egyptology Objection. So we can safely assume that Background Independence holds.

From Background Independence we know that, whatever you might uncover in your research, you would conclude that the risky lottery is no better than the safe lottery. For *any* value of b you might pin down, you'll establish that it's no better to take the risky lottery. To judge otherwise would be to accept Fanaticism. So you know what judgement you would make if you simply learned more, *no matter what* it is you would actually learn. So why do the many years of research? Why not just update your judgement now and say that $L_{\text{risky}} + B$ is no better than $L_{\text{safe}} + B$? You cannot. You must accept that $L_{\text{risky}} + B$ is strictly better, even though you can predict with certainty that you would change your mind if you knew more. To be able to change your mind you must go through those years of gruelling research and digging, even though you already know which judgement you would conclude to be correct.⁴⁵

This ambivalence seems deeply irrational. Surely we can sidestep those years of research into how B turns out, and make the judgement required by every possible value of b . Surely rationality requires that we do so, rather than require that we do not. But, if we deny Fanaticism, we must accept this inconsistency—an inconsistency which, to me, seems far more absurd than

⁴⁵Compare van Fraassen's (1984) principle of *Reflection* as well as its application by van Fraassen (2010, p. 17-8). That form of Reflection implies that, if you know that you would set the probability of some hypothesis to p whether you learned either proposition P or not- P , then its current probability should already be p . The version I am assuming is somewhat similar but need not be quite as strong.

simply accepting Fanaticism and even more absurd than the Egyptology Objection.^{46 47}

Taking stock of this section, we have further reasons to accept Fanaticism. If we deny it then we are forced to accept a version of the Egyptology Objection wherein the judgements of agents like us will be sensitive to their uncertainty about what occurred in ancient Egypt, even though their choices do not affect ancient Egypt at all. If we reject both Fanaticism and Background Independence, then we face an even more severe version of the Egyptology Objection: how such agents should compare their options will be sensitive to not just their uncertainty but also what actually occurred in ancient Egypt. And even if we accept Background Independence (as I think we should), if we try to reject Fanaticism then a bizarre form of inconsistency arises in how agents should compare their options—agents are sometimes forced to make judgements which are inconsistent with *anything* they might learn to resolve their uncertainty about far-off, unrelated events in the ancient Indus Valley. Any one of these implications—let alone any two of them—is absurd. In light of these, it seems a lot less appealing to deny Fanaticism.

⁴⁶Continuing from Footnotes 35 and 36, an analogue of the Indology Objection applies to averageism. (Analogues can also be constructed for other axiologies.) You might be facing two lotteries—one risky and one safe—in which you know that, whatever the outcome, an identical gift will be given to everyone in the world. And you're uncertain of just how much value that gift will bring, with your uncertainty corresponding to B . Then, as we saw above, there are some probability distributions for B that mean that you will judge $L_{\text{risky}} + B$ as strictly better than $L_{\text{safe}} + B$; but, if you opened the gift yourself and discovered what was inside, you already know that L_{risky} with the additional value b would be no better than L_{safe} with the additional value b . And this seems absurd, just as its analogue was for the aggregationist.

⁴⁷A closely related objection against rejecting Fanaticism is that, even if we do reject it, we will not be spared from many verdicts that seem deeply fanatical. As we saw in the case above, with the right uncertainty B about the value of unaffected events you must judge $L_{\text{risky}} + B$ better than $L_{\text{safe}} + B$. In practice, will we often have the right uncertainty about unaffected events to lead us to judge otherwise extremely risky lotteries as better than otherwise safe ones? Tarsney (n.d., §6) argues that we often will, for an epistemically rational agent and for a wide range of cases, including many resembling L_{risky} and L_{safe} (where the probability ϵ is at least 1 in 10^9 , and likely also a lot lower). His reasoning for this is that the moral value of distant events in our universe will be roughly correlated with the number of inhabited planets in the universe; and our cosmological evidence and the Drake equation together motivate a probability distribution over that number of inhabited planets which is relevantly similar to the above B . But even apart from that reasoning, an epistemically rational agent would still have some uncertainty over which probability distribution best characterises our cosmological evidence. So they would still put some non-zero credence in distributions of the right sort. And this is enough—when we take the probability-weighted average of many such distributions to get our *overall* distribution, it will then have the right properties (see *ibid.*: 37-9 for details). Given this, Stochastic Dominance by itself will lead us to making judgements which seem fanatical. So we would gain little from rejecting Fanaticism.

6.7 Summary

These are just some of the compelling arguments for accepting Fanaticism and, with it, fanatical verdicts in cases like Dyson's Wager. There are other compelling arguments too—for instance, all of the arguments for the stronger claims of expected value theory, or for expected utility theory with an unbounded utility function. *A fortiori*, these also justify Fanaticism. For a sampling of these, I refer interested readers to Feller (1968), Ng (2012), Briggs (2015), and Thoma (2018).

Some philosophers still reject expected value theory (e.g., (Buchak, 2013), (Smith, 2014), (Monton, 2019), (Tarsney, n.d.)), and presumably the arguments that imply it. But this is not enough to escape Fanaticism. As I have demonstrated here, there are compelling arguments in its favour—arguments stronger than many of those for expected value theory.

To recap, if we deny Fanaticism then we must deny either Minimal Tradeoffs or transitivity, and accept the counterintuitive verdicts which follow. Likewise, we must accept either violations of Scale Independence, or absurd sensitivity to the tiniest differences in probabilities and value. So too, to deny Fanaticism, we must accept a version of the Egyptology Objection: for agents like ourselves, which judgement is correct can depend on our beliefs about what occurred in far-off, unrelated events, such as those in ancient Egypt. And, unless we wish to accept an even more severe version of the Egyptology Objection, we also face the *Indology* Objection: we sometimes ought to make judgements that we know we would reject if we learnt more, no matter *what* we might learn.

Given all of this, it no longer seems so attractive to reject Fanaticism. As it turns out, the cure is worse than the disease. I would suggest that it is better to simply accept Fanaticism. And with it we should accept fanatical verdicts such as that in Dyson's Wager and those we encounter in the infinite setting more widely. It may have seemed a problem that expected value theory—and that my proposal in earlier chapters—gave such verdicts. But it's a good thing they did. It turns out that denying Fanaticism brings far worse implications.

Conclusion

Recall where we stood at the beginning of this thesis. We faced a seemingly devastating objection against any moral theory that is even *minimally* aggregative—any theory that gives even some consideration to the total moral value an action will produce.

Our best cosmology gives us reason to expect our universe to be infinite—to contain infinitely many instances of every possible phenomenon, including valuable (and disvaluable) human lives. And our standard theories of aggregation cannot handle such outcomes. Even when one's actions only effect finite changes in value, those theories cannot tell us that one action results in a better outcome than another. And so the aggregative part of minimally aggregative theories is entirely impotent.

This objection may seem to put us in a hopeless position, but it does not. We can modify minimally aggregative theories to extend their judgements into the infinite setting. We already have such proposals from Vallentyne (1993), van Liedekerke and Lauwers (1997), Vallentyne and Kagan (1997), Lauwers and Vallentyne (2004), Bostrom (2011), Arntzenius (2014), and Jonsson and Voorneveld (2018). I propose my own in Chapter 2, a form of spatiotemporal expansionism. As I argue there and in Chapter 3, not only is this proposal defensible but it is also more plausible than any of the other proposals so far. Why? None of those other proposals can compare outcomes in practice in a plausible way—they are unable to compare outcomes that differ by long-lasting chaotic differences and, given the ubiquity of chaos in our universe, these include almost every pair of outcomes an agent will ever need to compare.

But my proposal faces problems of its own. For one, its verdicts depend on the spatiotemporal positioning of valuable events, but spatiotemporal positioning in our universe is in some sense relative to the perspective of the observer. Does this mean that my proposal must allow that moral

betterness is also relative to the observer's perspective? Thankfully, it does not. As I showed in Chapter 4, the basic proposal can be modified to avoid this.

Another problem is that spatiotemporal expansionism appears unable to easily provide us with comparisons of *lotteries* over infinite outcomes. And we need to be able to compare such lotteries—in practice, agents rarely know the outcomes of their actions with certainty, and so can at best compare them by the lotteries they produce. But, unlike aggregative theories in the finite setting, spatiotemporal expansionism fails to give us a cardinal measure of the value of each outcome. So, when faced with lotteries over such outcomes, we cannot easily apply expected value theory and evaluate those lotteries by the expectation of their total value. But, as I show in Chapter 5, we can overcome this obstacle—with a little creativity, we can formulate a more general version of expected value theory that works well on outcomes like these. Best of all, if we combine this theory of comparing lotteries with my theory of comparing infinite worlds—spatiotemporal expansionism—we can compare the lotteries that an agent's actions will typically produce in practice, even if the universe is widely chaotic.

You might object to my proposal for comparing lotteries on the basis that it is *fanatical*: it implies that it is better to produce a tiny probability of infinite gain over a certainty of finite gain, no matter how enormous the finite gain or how tiny the probability of infinite gain. You might find this implication implausible, just as many philosophers have found the fanatical verdicts of expected value theory (e.g., that of Pascal, 1670) implausible for centuries. I address this worry in Chapter 6: as it turns out, there are a variety of compelling arguments in favour of allowing fanatical verdicts. To deny all such verdicts is to bring on even more severe objections to one's theory.

Where does this leave us? In a better position than we started, I would suggest. We have a plausible method of extending minimally aggregative theories to the infinite setting, which delivers comparisons in realistic cases and which avoids the most serious problems we encountered along the way. The possibility of an infinite universe no longer poses such a devastating objection to the adoption of minimally aggregative theories. So we can retain minimally aggregative theories, and continue to make moral decisions based on what will promote the good.

Bibliography

- ARNTZENIUS, F., 2014. Utilitarianism, decision theory and eternity. *Philosophical Perspectives*, 28, 1 (2014), pp. 31–58. (cited on pages 5, 9, 29, 32, 53, 78, 90, 92, 99, 118, 120, 121, 123, and 181)
- ARRHENIUS, G. AND RABINOWICZ, W., 2015. Value superiority. *The Oxford Handbook of Value Theory*, (2015), pp. 225–48. (cited on page 165)
- ARROW, K., 1971. *Essays in the Theory of Risk-Bearing*. Markham, Chicago. (cited on pages 150, 151, and 170)
- ARROW, K., 1999. Discounting, morality, and gaming. In *Discounting and Intergenerational Equity*, p. 13–21. Cambridge University Press, Cambridge. (cited on page 90)
- ASHEIM, G.; D’ASPREMONT, C.; AND BANERJEE, K., 2010. Generalized time-invariant overtaking. *Journal of Mathematical Economics*, 46, 4 (2010), pp. 519–33. (cited on pages 19 and 76)
- ASKELL, A., 2019. *Pareto Principles in Infinite Ethics*. PhD dissertation, New York University. (cited on pages 3, 5, 9, 10, 14, 18, 37, 76, 78, and 81)
- BADER, R., n.d. *Person-Affecting Population Ethics*. unpublished manuscript. (cited on page 73)
- BARON, J.; GRANATO, L.; SPRANCA, M.; AND TEUBAL, E., 1993. Decision making biases in children and early adolescents: Exploratory studies. *Merrill Palmer Quarterly*, 39 (1993), pp. 23–47. (cited on page 153)
- BASU, K. AND MITRA, T., 2007. Utilitarianism for infinite utility streams: A new welfare criterion and its axiomatic characterization. *Journal of Economic Theory*, 133, 1 (2007), pp. 350–73. (cited on page 119)
- BECKSTEAD, N., 2013. *On the Overwhelming Importance of Shaping the Far Future*. PhD dissertation, Rutgers University. (cited on pages 150, 152, and 161)

-
- BECKSTEAD, N. AND THOMAS, T., n.d. A paradox for tiny probabilities and enormous values. Unpublished manuscript. (cited on pages 161 and 170)
- BENCI, V. AND NASSO, M. D., 2003. Numerosities of labelled sets: a new way of counting. *Advances in Mathematics*, 173, 1 (2003), pp. 50–67. (cited on page 26)
- BOSTROM, N., 2009. Pascal’s mugging. *Analysis*, 69, 3 (2009), pp. 443–5. (cited on page 152)
- BOSTROM, N., 2011. Infinite ethics. *Analysis and Metaphysics*, 10 (2011), pp. 9–59. (cited on pages 3, 5, 9, 33, 92, 118, 120, 121, 123, 150, 152, and 181)
- BRIGGS, R., 2015. Costs of abandoning the sure-thing principle. *Canadian Journal of Philosophy*, 45 (2015), pp. 827–40. (cited on pages 177 and 180)
- BROOME, J., 2004. *Weighing Lives*. Oxford University Press, Oxford. (cited on page 14)
- BROOME, J., 2019. Against denialism. *The Monist*, 102, 1 (2019), pp. 110–29. (cited on page 64)
- BUCHAK, L., 2013. *Risk and Rationality*. Oxford University Press, Oxford. (cited on pages 127, 163, and 180)
- CAIN, J., 1995. Infinite utility. *Australasian Journal of Philosophy*, (1995). (cited on pages 10, 17, 53, 67, and 90)
- CARROLL, S., 2004. *Spacetime and Geometry: An Introduction to General Relativity*. Addison Wesley, San Francisco. (cited on page 113)
- CARROLL, S., 2017. Why Boltzmann brains are bad. *arXiv preprint*, (2017). (cited on page 4)
- CHALMERS, D., 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17, 9 (2010), pp. 7–65. (cited on pages 89 and 148)
- CHEN, D.; MOSKOWITZ, T.; AND SHUE, K., 2016. Decision making under the gambler’s fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, 131, 3 (2016), pp. 1181–242. (cited on page 152)
- CHUNG, K. AND FUCHS, W., 1951. On the distribution of values of sums of random variables. *Memoirs of the American Mathematical Society*, 6 (1951), pp. 12. (cited on pages 41, 69, and 84)

-
- CLARK, M., n.d. Infinite ethics, intrinsic value and the Pareto principle. Unpublished manuscript. (cited on page 73)
- COHN, E., 1904. Zur elektrodynamik bewegter systeme ii [on the electrodynamics of moving systems ii]. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, 43, 2 (1904), pp. 1404–16. (cited on pages 89 and 96)
- COLYVAN, M., 2008. Relative expectation theory. *Journal of Philosophy*, 105, 1 (2008), pp. 37–44. (cited on page 118)
- COLYVAN, M. AND HÁJEK, A., 2016. Making ado without expectations. *Mind*, 125, 499 (2016), pp. 829–57. (cited on page 118)
- DASGUPTA, P., 2008. Discounting climate change. *Journal of Risk and Uncertainty*, 37, 2-3 (2008), pp. 141–69. (cited on pages 64 and 90)
- DE BUFFON, G.-L. L., 1777. Essai d'arithmétique morale. *Supplement a l'Histoire Naturelle*, vol. 4 (1777). (cited on pages 149, 151, 152, 156, and 163)
- DE SIMONE, A.; GUTH, A.; LINDE, A.; NOORBALA, M.; SALEM, M.; AND VILENKIN, A., 2010. Boltzmann brains and the scale-factor cutoff measure of the multiverse. *Physical Review D*, 82, 6 (2010), 063520. (cited on page 4)
- DREIER, J., 2019. World-centered value. In *Consequentialism: New Directions and Problems* (Ed. C. SEIDEL), p. 31–50. Oxford University Press, Oxford. (cited on page 14)
- DYSON, F., 1979. Time without end: Physics and biology in an open universe. *Reviews of Modern Physics*, 51 (1979). (cited on page 148)
- DYSON, F., 1981. Life in the universe. Unpublished Darwin Lecture, Darwin College, Cambridge, Nov 10. (cited on page 148)
- EBERHARD, P. H. AND ROSS, R. R., 1989. Quantum field theory cannot provide faster-than-light communication. *Foundations of Physics Letters*, 2 (1989), pp. 127–49. (cited on page 107)
- EINSTEIN, A., 1905. Zur elektrodynamik bewegter körper. *Annalen der Physik*, 322, 10 (1905), pp. 891–921. (cited on pages 89 and 95)

- FELDMAN, D., 2019. *Chaos and Dynamical Systems*. Princeton University Press, Princeton. (cited on page 60)
- FELLER, W., 1968. *An Introduction to Probability Theory and Its Applications* (3rd ed.). Wiley, New York. (cited on pages 157 and 180)
- GARRETT, B.; MICHELL, G.; AND SCURICH, N., 2018. Comparing categorical and probabilistic fingerprint evidence. *Journal of Forensic Sciences*, 63, 6 (2018), pp. 1712–7. (cited on page 153)
- GARRIGA, J. AND VILENKIN, A., 2001. Many worlds in one. *Physical Review D*, 64, 4 (2001), 043511. (cited on page 4)
- GILOVICH, T.; TVERSKY, A.; AND VALLONE, R., 1985. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17, 3 (1985), pp. 295–314. (cited on page 152)
- GLEICK, J., 1987. *Chaos: Making a New Science*. Viking Books, New York. (cited on page 60)
- GOTT, J. R., 2008. Boltzmann brains—I’d rather see than be one. *arXiv preprint*, (2008). (cited on page 4)
- GREAVES, H., 2016. Cluelessness. *Proceedings of the Aristotelian Society*, 116, 3 (2016), pp. 311–39. (cited on pages 39, 49, 58, 62, 66, 86, and 140)
- GRIMMETT, G. AND STIRZAKER, D., 1992. *Probability and Random Processes*. Clarendon Press, Oxford. (cited on page 62)
- GURMANKIN, L. AND BARON, J., 2005. How bad is a 10% chance of losing a toe?: Judgments of probabilistic conditions by doctors and laypeople. *Memory and Cognition*, 33 (2005), pp. 1399–406. (cited on page 153)
- GUTH, A., 2007. Eternal inflation and its implications. *Journal of Physics A: Mathematical and Theoretical*, 40, 25 (2007), pp. 6811. (cited on page 4)
- HAFELE, J. AND KEATING, R., 1972. Around-the-world atomic clocks: Observed relativistic time gains. *Science*, 177, 4044 (1972), pp. 168–70. (cited on page 110)

-
- HÁJEK, A., 2014. Unexpected expectations. *Mind*, 123, 490 (2014), pp. 533–67. (cited on pages 151, 152, and 156)
- HÁJEK, A., n.d. Most counterfactuals are false. Unpublished manuscript. (cited on page 152)
- HÁJEK, A. AND NOVER, H., 2006. Perplexing expectations. *Mind*, 115, 459 (2006), pp. 703–20. (cited on page 3)
- HALBERSTAM, H. AND ROTH, K. F., 2012. *Sequences*. Clarendon Press, Oxford. (cited on page 25)
- HAMKINS, J. AND MONTERO, B., 2000. With infinite utility, more needn't be better. *Australasian Journal of Philosophy*, 78, 2 (2000), pp. 231–40. (cited on page 3)
- HANOCH, Y.; ROLISON, J.; AND FREUND, A., 2019. Reaping the benefits and avoiding the risks: Unrealistic optimism in the health domain. *Risk Analysis*, 39, 4 (2019), pp. 792–804. (cited on page 153)
- HUEMER, M., 2008. In defence of repugnance. *Mind*, 117, 468 (2008), pp. 899–933. (cited on page 14)
- ISAACS, Y., 2016. Probabilities cannot be rationally neglected. *Mind*, 125, 499 (2016), pp. 759–62. (cited on pages 152 and 156)
- JACKSON, F., 1991. Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics*, 101, 3 (1991), pp. 461–82. (cited on pages 71, 86, 117, and 140)
- JONSSON, A. AND VOORNEVELD, M., 2018. The limit of discounted utilitarianism. *Theoretical Economics*, 13, 1 (2018), pp. 19–37. (cited on pages 5, 9, 92, 120, 138, and 181)
- KAHNEMAN, D. AND TVERSKY, A., 1982. Evidential impact of base rates. In *Judgment under uncertainty: Heuristics and biases* (Eds. D. KAHNEMAN; P. SLOVIC; AND A. TVERSKY), p. 153–60. Cambridge University Press, Cambridge. (cited on page 153)
- KOOPMANS, T., 1972. Representation of preference orderings over time. *Decision & Organization*, 57 (1972). (cited on page 23)
- KRIPKE, S., 1980. Basil Blackwell, Oxford. (cited on page 12)

-
- LASKAR, J., 1989. A numerical experiment on the chaotic behavior of the solar system. *Nature*, 338, 6212 (1989), pp. 237–8. (cited on page 65)
- LAUWERS, L., 2010. Ordering infinite utility streams comes at the cost of a non-Ramsey set. *Journal of Mathematical Economics*, 46, 1 (2010), pp. 32–7. (cited on page 14)
- LAUWERS, L. AND VALLENTYNE, P., 2004. Infinite utilitarianism: more is always better. *Economics & Philosophy*, 20, 2 (2004), pp. 307–30. (cited on pages 5, 9, 78, 119, and 181)
- LAUWERS, L. AND VALLENTYNE, P., 2016. Decision theory without finite standard expected value. *Economics and Philosophy*, 32, 3 (2016), pp. 383–407. (cited on page 127)
- LE ROND D’ALEMBERT, J., 1784. *Opuscles Mathématiques ou Mémoires sur différens sujets de Géométrie, de Méchanique, d’Optique, d’Astronomie, Tome Second*. David, Paris. (cited on pages 149 and 151)
- LENMAN, J., 2000. Consequentialism and cluelessness. *Philosophy and Public Affairs*, 29, 4 (2000), pp. 342–70. (cited on pages 39 and 66)
- LEWIS, D., 1971. Counterparts of persons and their bodies. *Journal of Philosophy*, 68 (1971), pp. 203–11. (cited on page 12)
- LIGHTHILL, M. AND WHITHAM, G., 1955. On kinematic waves ii: A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229, 1178 (1955), pp. 317–45. (cited on page 58)
- LINDE, A., 2007. Sinks in the landscape, Boltzmann brains and the cosmological constant problem. *Journal of Cosmology and Astroparticle Physics*, (2007). (cited on page 4)
- LIPKIN, L., 2003. Tossing a fair coin. *The College Mathematics Journal*, 34, 2 (2003), pp. 128–33. (cited on pages 49 and 86)
- LORENZ, E., 1963. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20, 2 (1963), pp. 130–41. (cited on page 63)
- LORENZ, E., 1972. Does the flap of a butterfly’s wings in Brazil set off a tornado in Texas? (1972). (cited on page 63)

-
- MACASKILL, W. AND MOGENSEN, A., n.d. The paralysis argument. Unpublished manuscript, available at <https://globalprioritiesinstitute.org/william-macaskill-andreas-mogensen-the-paralysis-argument/>. (cited on page 58)
- MALHOTRA, R.; HOLMAN, M.; AND ITO, T., 2001. Chaos and stability of the solar system. *Proceedings of the National Academy of Sciences*, 98, 22 (2001), 12342–3. (cited on page 64)
- MCCLELLAND, G.; SCHULZE, W.; AND COURSEY, D., 1993. Insurance for low-probability hazards: A bimodal response to unlikely events. In *Making Decisions About Liability and Insurance*, p. 95–116. Springer, Dordrecht. (cited on page 153)
- McMAHAN, J., 1981. Problems of population theory. *Ethics*, 92, 1 (1981), pp. 96–127. (cited on page 169)
- MEACHAM, C., 2020. Too much of a good thing: Decision-making in cases with infinitely many utility contributions. *Synthese*, (2020). (cited on pages 118 and 121)
- MELCHERS, M.; MONTAG, C.; REUTER, M.; SPINATH, F.; AND HAHN, E., 2016. How heritable is empathy? differential effects of measurement and subcomponents. *Motivation and Emotion*, 40, 5 (2016), pp. 720–30. (cited on page 62)
- MICHELSON, A. A. AND MORLEY, E. W., 1887. On the relative motion of the earth and of the luminiferous ether. *Sidereal Messenger*, 6 (1887), pp. 306–10. (cited on page 95)
- MILL, J. S., 1861. *Utilitarianism*. Oxford University Press, New York. (cited on page 22)
- MOGENSEN, A., 2021. Maximal cluelessness. *Philosophical Quarterly*, 71, 1 (2021), pp. 141–62. (cited on page 169)
- MONTON, B., 2019. How to avoid maximising expected utility. *Philosophers' Imprint*, 19 (2019). (cited on pages 149, 151, 152, 155, 156, 157, and 180)
- NEBEL, J., 2018. The good, the bad, and the transitivity of *better than*. *Noûs*, 52, 4 (2018), pp. 874–99. (cited on page 14)
- NELSON, M., 1991. Utilitarian eschatology. *American Philosophical Quarterly*, 28, 4 (1991), pp. 339–47. (cited on page 3)

- NG, Y.-K., 2012. A case for happiness, cardinalism, and interpersonal comparability. *The Economic Journal*, 107 (2012), pp. 1848–58. (cited on page 180)
- NOVER, H. AND HÁJEK, A., 2004. Vexing expectations. *Mind*, 113, 450 (2004), pp. 237–49. (cited on page 3)
- OFFICE FOR NATIONAL STATISTICS, 2018. Childbearing for women born in different years, England and Wales: 2017. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/conceptionandfertilityrates/bulletins/childbearingforwomenbornindifferentyearsenglandandwales/2017>. (accessed: 02.09.2020). (cited on page 61)
- OFFICE FOR NATIONAL STATISTICS, 2019. National life tables: UK. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables>. (accessed: 02.09.2020). (cited on page 61)
- PALMER, T.; DÖRING, A.; AND SEREGIN, G., 2014. The real butterfly effect. *Nonlinearity*, 27, 9 (2014), R123. (cited on page 64)
- PARFIT, D., 1984. *Reasons and Persons*. Oxford University Press, Oxford. (cited on pages 22, 23, 62, 151, and 169)
- PARFIT, D., 1997. Equality and priority. *Ratio*, 10 (1997), pp. 202–21. (cited on page 2)
- PASCAL, B., 1670. *Pensées*. Translated by W. F. Trotter, Dent, London, 1910. (cited on pages 147, 151, and 182)
- PEMANTLE, R., 2007. A survey of random processes with reinforcement. *Probability Surveys*, 4 (2007), pp. 1–79. (cited on page 70)
- PIVATO, M., 2014. Additive representation of separable preferences over infinite products. *Theory and Decision*, 77, 1 (2014), pp. 31–83. (cited on pages 33 and 123)
- PLOMIN, R. AND DEARY, I. J., 2015. Genetics and intelligence differences: five special findings. *Molecular Psychiatry*, 20, 1 (2015), 98–108. (cited on page 62)

-
- POINCARÉ, H., 1890. *The Three-Body Problem and the Equations of Dynamics: Poincaré's Foundational Work on Dynamical Systems Theory*. Popp, Bruce D. (trans.), Springer International Publishing (2017). (cited on page 64)
- POMATTO, L.; STRACK, P.; AND TAMUZ, O., 2018. Stochastic dominance under independent noise. *arXiv e-prints*, (2018). (cited on page 176)
- PUMMER, T., 2016. Whether and where to give. *Philosophy & Public Affairs*, 44, 1 (2016), pp. 77–95. (cited on page 1)
- PUTENSEN, C. AND WRIGGE, H., 2007. Tidal volumes in patients with normal lungs: one for all or the less, the better? *Anesthesiology*, 106 (2007), pp. 1085–7. (cited on page 64)
- RABIN, M., 2000. Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68, 5 (2000), pp. 1281–92. (cited on page 163)
- RAILTON, P., 1984. Alienation, consequentialism, and the demands of morality. *Philosophy and Public Affairs*, 13 (1984). (cited on page 71)
- RAMSEY, F. P., 1964. Truth and probability (1926). In *The Foundations of Mathematics and other Logical Essays*, p. 156–98. (cited on page 163)
- RAWLS, J., 1971. *A Theory of Justice*. Harvard University Press, Cambridge. (cited on page 1)
- RUBINSZTEJN, A. Demonstrating chaos with a double pendulum. https://en.wikipedia.org/wiki/File:Demonstrating_Chaos_with_a_Double_Pendulum.gif. (cited on page 60)
- SANDBERG, A., n.d. *Grand Futures*. Unpublished manuscript. (cited on page 148)
- SCHOENFIELD, M., 2014. Decision making in the face of parity. *Philosophical Perspectives*, 28 (2014). (cited on pages 126, 127, and 160)
- SCHWARZSCHILD, K., 1916. On the gravitational field of a mass point according to Einstein's theory. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, (1916). (cited on page 111)
- SCHWITZGEBEL, E., 2017. 1% skepticism. *Noûs*, 51, 2 (2017), pp. 271–90. (cited on page 152)

-
- SEIDENFELD, T.; SCHERVISH, M.; AND KADANE, J., 2009. Preference for equivalent random variables: A price for unbounded utilities. *Journal of Mathematical Economics*, 45 (2009), pp. 5–6. (cited on page 127)
- SIDGWICK, H., 1907. *The Methods of Ethics*, 7th edn. Macmillan, London. (cited on page 22)
- SINGER, P., 1972. Famine, affluence, and morality. *Philosophy and Public Affairs*, 1 (1972), pp. 299–343. (cited on page 22)
- SMITH, N., 2014. Is evaluative compositionality a requirement of rationality? *Mind*, 123, 490 (2014), pp. 457–502. (cited on pages 127, 149, 151, 152, 155, 163, and 180)
- SVENSSON, L.-G., 1980. Equity among generations. *Econometrica*, 48, 5 (1980), pp. 1251–6. (cited on page 14)
- TANG, J.; NICOLLE, A.; KLETTNER, C. A.; PANTELIC, J.; WANG, L.; AND TAN, A. B. S. A., 2013. Airflow dynamics of human jets: sneezing and breathing - potential sources of infectious aerosols. *PLOS ONE*, 8, 4 (2013), e59970. (cited on page 64)
- TARSNEY, C., n.d. Exceeding expectations: Stochastic dominance as a general decision theory. Unpublished manuscript. (cited on pages 150, 151, 152, 176, 179, and 180)
- TEMKIN, L., 2014. *Rethinking the Good*. Oxford University Press, Oxford. (cited on page 14)
- TEMKIN, L., 2015. Rationality with respect to people, places, and times. *Canadian Journal of Philosophy*, 45, 5-6 (2015), pp. 576–608. (cited on page 90)
- THOMA, J., 2018. Risk aversion and the long run. *Ethics*, 129, 2 (2018), pp. 230–53. (cited on page 180)
- TIPLER, F., 1986. Cosmological limits on computation. *International Journal of Theoretical Physics*, 25 (1986). (cited on page 148)
- TVERSKY, A. AND KAHNEMAN, D., 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185 (1974), pp. 1124–31. (cited on page 153)
- TVERSKY, A. AND KAHNEMAN, D., 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, (1983). (cited on page 152)

-
- VALLENTYNE, P., 1993. Utilitarianism and infinite utility. *Australasian Journal of Philosophy*, 71, 2 (1993), pp. 212–7. (cited on pages 5, 9, 92, and 181)
- VALLENTYNE, P. AND KAGAN, S., 1997. Infinite value and finitely additive value theory. *Journal of Philosophy*, 94, 1 (1997), pp. 5–26. (cited on pages 5, 9, 29, 56, 78, 119, 120, 138, and 181)
- VAN FRAASSEN, B., 1984. Belief and the will. *Journal of Philosophy*, 81 (1984), pp. 235–56. (cited on page 178)
- VAN FRAASSEN, B., 2010. Evidential symmetry and mushy credence. *Oxford Studies in Epistemology*, 3 (2010), pp. 161–86. (cited on page 178)
- VAN LIEDEKERKE, L., 1995. Should utilitarians be cautious about an infinite future? *Australasian Journal of Philosophy*, 73, 3 (1995), pp. 405–7. (cited on pages 73 and 77)
- VAN LIEDEKERKE, L. AND LAUWERS, L., 1997. Sacrificing the patrol: Utilitarianism, future generations and infinity. *Economics & Philosophy*, 13, 2 (1997), pp. 159–74. (cited on pages 5, 9, and 181)
- VON WEIZSÄCKER, C., 1965. Existence of optimal programs of accumulation for an infinite time horizon. *The Review of Economic Studies*, 32, 2 (1965), pp. 85–104. (cited on page 83)
- WAINSCHEIN, P.; JAIN, D. P.; YENGO, L.; ZHENG, Z.; CUPPLES, L. A.; SHADYAB, A. H.; AND MCK-NIGHT, B., 2019. Recovery of trait heritability from whole genome sequence data. *BioRxiv*, (2019), 588020. (cited on page 62)
- WALD, R., 1983. Asymptotic behavior of homogeneous cosmological models in the presence of a positive cosmological constant. *Physical Review D*, 28, 8 (1983), pp. 2118. (cited on page 4)
- WILKINSON, H., 2021. Infinite aggregation: Expanded addition. *Philosophical Studies*, 178, 6 (2021), pp. 1917–49. (cited on page 9)
- WILKINSON, H., n.d.(a). Chaos, add infinitum. Unpublished manuscript. (cited on page 9)
- WRIGHT, R., 2009. *The Ancient Indus: Urbanism, Economy, and Society*. Cambridge University Press, Cambridge. (cited on page 177)

ZAME, W., 1999. Can intergenerational equity be operationalized? *Theoretical Economics*, 2 (1999).
(cited on page 14)

© Hayden Wilkinson 2021

All rights reserved.