Addleton Academic Publishers
Analysis and Metaphysics
Volume 10, 2011, pp. 9–59, ISSN 1584-8574

INFINITE ETHICS

NICK BOSTROM
nick.bostrom@philosophy.ox.ac.uk
Future of Humanity Institute
Faculty of Philosophy & Oxford Martin School
Oxford University

ABSTRACT. Aggregative consequentialism and several other popular moral theories are threatened with paralysis: when coupled with some plausible assumptions, they seem to imply that it is always ethically indifferent what you do. Modern cosmology teaches that the world might well contain an infinite number of happy and sad people and other candidate value-bearing locations. Aggregative ethics implies that such a world contains an infinite amount of positive value and an infinite amount of negative value. You can affect only a finite amount of good or bad. In standard cardinal arithmetic, an infinite quantity is unchanged by the addition or subtraction of any finite quantity. So it appears you cannot change the value of the world. Modifications of aggregationism aimed at resolving the paralysis are only partially effective and cause severe side effects, including problems of "fanaticism", "distortion", and erosion of the intuitions that originally motivated the theory. Is the infinitarian challenge fatal?

Keywords: aggregative, consequentialism, ethics, infinite, cosmology, value

1. The Challenge

1.1. The Threat of Infinitarian Paralysis

When we gaze at the starry sky at night and try to think of humanity from a "cosmic point of view", we feel small. Human history, with all its earnest strivings, triumphs, and tragedies can remind us of a colony of ants, laboring frantically to rearrange the needles of their little ephemeral stack. We brush such late-night rumination aside in our daily life and analytic philosophy. But, might such seemingly idle reflections hint at something of philosophical significance? In particular, might they contain an important implication for our moral theorizing?

If the cosmos is finite, then our own comparative smallness does not necessarily undermine the idea that our conduct matters even from an impersonal perspective. We might constitute a minute portion of the whole, but that does not detract from our absolute importance. Suppose there are a hundred thousand other planets with civilizations that had their own holocausts. This does not alter the fact that the holocaust that humans caused contributed an enormous quantity of suffering to the world, a quantity measured in millions of destroyed lives. Maybe this is a tiny fraction of the total suffering in the world, but in absolute terms it is unfathomably large. Aggregative ethics can thus be reconciled with the finite case if we note that, when sizing up the moral significance of our acts, the relevant consideration is not how big a part they constitute of the whole of the doings and goings-on in the universe, but rather what difference they make in absolute terms.

The infinite case is fundamentally different. Suppose the world contains an infinite number of people and a corresponding infinity of joys and sorrows, preference satisfactions and frustrations, instances of virtue and depravation, and other such local phenomena at least some of which have positive or negative value. More precisely, suppose that there is some finite value ε such that there exists an infinite number of local phenomena (this could be a subset of e.g. persons, experiences, characters, virtuous acts, lives, relationships, civilizations, or ecosystems) each of which has a value $\geq \varepsilon$ and also an infinite number of local phenomena each of which has a value $\leq (-\varepsilon)$. Call such a world *canonically infinite*. Ethical theories that hold that value is aggregative imply that a canonically infinite world contains an infinite quantity of positive value and an infinite quantity of negative value. This gives rise to a peculiar predicament. We can do only a finite amount of good or bad. Yet in cardinal arithmetic, adding or subtracting a finite quantity does not change an infinite quantity. Every possible act of ours therefore has the same net effect on the total amount of good and bad in a canonically infinite world: none whatsoever.

Aggregative consequentialist theories are threatened by *infinitarian paral ysis*: they seem to imply that if the world is canonically infinite then it is always ethically indifferent what we do. In particular, they would imply that it is ethically indifferent whether we cause another holocaust or prevent one from occurring. If any non-contradictory normative implication is a *reductio ad absurdum*, this one is.

Is the world canonically infinite or not? Recent cosmological evidence suggests that the world is probably infinite.[1] Moreover, if the totality of physical existence is indeed infinite, in the kind of way that modern cosmology suggests it is, then it contains an infinite number of galaxies, stars, and planets. If there are an infinite number of planets then there is, with probability one, an infinite number of people.[2] Infinitely many of these people are happy,

infinitely many are unhappy. Likewise for other local properties that are plausible candidates for having value, pertaining to person-states, lives, or entire societies, ecosystems, or civilizations – there are infinitely many democratic states, and infinitely many that are ruled by despots, etc. It therefore appears likely that the actual world is canonically infinite.

We do not know for sure that we live in a canonically infinite world. Contemporary cosmology is in considerable flux, so its conclusions should be regarded as tentative. But it is definitely not reasonable, in light of the evidence we currently possess, to assume that we do *not* live in a canonically infinite world. And that is sufficient for the predicament to arise. Any ethical theory that fails to cope with this *likely empirical contingency* must be rejected. We should not accept an ethical theory which, conditional on our current best scientific guesses about the size and nature of the cosmos, implies that it is ethically indifferent whether we cause or prevent another holocaust.[3]

1.2. Which Theories Are Threatened?

Infinitarian paralysis threatens a wide range of popular ethical theories. Consider, to begin with, hedonistic utilitarianism, which in its classical formulation states that you ought to do that which maximizes the total amount of pleasure and minimizes the total amount of pain in the world. If pleasure and pain are already infinite, then all possible actions you could take would be morally on a par according to this criterion, for none of them would make any difference to the total amount of pleasure or pain. Endorsing this form of utilitarianism commits one to the view that, conditional on the world being canonically infinite, ending world hunger and causing a famine are ethically equivalent options. It is not the case that you ought to do one rather than the other.

The threat is not limited to hedonistic utilitarianism. Utilitarian theories that have a broader conception of the good – happiness, preference-satisfaction, virtue, beauty-appreciation, or some objective list of ingredients that make for a good life – face the same problem. So, too, does average utilitarianism, mixed total/average utilitarianism, and prioritarian views that place a premium on the well-being of the worst off. In a canonically infinite world, average utility and most weighted utility measures are just as imperturbable by human agency as is the simple sum of utility.

Many non-utilitarian ethical theories are also imperiled. One common view is that in determining what we ought to do we should take into account the difference our acts would make to the total amount of well-being experienced by sentient persons even though we must also factor in the special obligations that we have to particular individuals (and perhaps various deontological side-constraints). If our actions never make any difference to the amount of well-being in the world, the maximizing component of such hybrid

theories becomes defunct. Depending on the structure of the theory, the components that remain in operation may – *or may not* – continue to generate sensible moral guidance.

Moorean views, which claim that value resides in "organic unities", are also vulnerable. If the relevant unities supervene on some medium-sized spacetime regions, such as societies or planets, then there might well be infinitely many such unities. If, instead, the relevant unity is the universe itself, then it is unclear that we could change its total value by modifying the infinitesimal part of it that is within our reach.[4]

For simplicity, we will focus most of the discussion on purely consequentialist theories (even though, as we have seen, the problems affect a much larger family of ethical systems). However, not all consequentialist theories are threatened. The vulnerability infinitarian paralysis arises from the combination of two elements: consequentialism and aggregationism. By "aggregationism" we refer to the idea that the value of a world is (something like) the sum or aggregate of the values of its parts, where these parts are some kind of local phenomena such as experiences, lives, or societies. By consequentialism we refer to the idea that the rightness or wrongness of an action is (somehow) determined on the basis of considerations about whether its consequences increase or decrease value. We shall later explore how various more precise explications of "aggregationism" and "consequentialism" fare in relation to the threat of infinitarian paralysis and associated challenges.

The challenge addressed in this paper is related to – but also crucially different from – Pascal's wager, the St. Petersburg paradox, the Pasadena problem, the Heaven and Hell problem, and kindred *prudential* "infinite" decision problems.[5] Related, because in each case there is, purportedly, the prospect of infinite values to be reckoned with. Different, because one important escape route that is available in the prudential cases is blocked in the ethical case. This is the route of denying that infinite values are really at stake. One way of responding to Pascal's wager, for instance, is by taking it to show that we do not in fact have an infinitely strong preference for spending an eternity in Heaven. The attractiveness of this response would be enhanced by the finding that the alternative is to accept highly counterintuitive consequences. In a revealed-preference paradigm, this is anyway a perfectly natural view. If we accept a theory of rationality that grounds what we have reason to do in our preferences (whether raw or idealized) then we have a simple and plausible answer to Pascal: Yes, if one had an infinitely strong preference for eternal life in Heaven, then it would be rational to forego any finite pleasure on Earth for any ever-so-slight increase in the odds of salvation (at least if one assumes that there would be no chance of obtaining an infinite good if one did not accept the wager, and no chance that accepting it might backfire and result in an infinite bad). However, if one does not have an infinitely strong preference for Heaven, then Pascal's argument does not

show that one is irrational to decline the wager. The fact that most people would on reflection reject the wager would simply show that most people do not place an infinite value on Heaven. The analogous response is *not* available to the ethical aggregationist, who is committed to the view that the total value of a world is the aggregate of the value of its parts, for this *entails* placing an infinite value on certain kinds of world. If a world has an infinite number of locations, and there is some finite value v such that an infinite number of the locations have an ethical value greater than v, then that world has an infinite ethical value. This is a core commitment of aggregationism; giving it up means giving up aggregationism. So the possibility of an infinite world presents a graver problem for aggregative ethics than it does for prudential rationality.[6]

1.3. Modifiable Components of Aggregative Ethics, and Adequacy Criteria

The aggregative consequentialist ethical theories that we shall investigate can be dissected into four components:

Substantive component
• A *value rule*, specifying what kinds of local phenomena have value, and how much positive or negative value each instance of such a phenomenon has.

Formal components
• A *domain rule*, specifying what the relevant domain is;
• An *aggregation rule*, specifying how to sum or aggregate the local values in the domain into a total value; and
• A *selection rule* specifying how the set of right (and wrong) actions is selected, from among the available acts, on the basis of the result of the aggregation.

In the standard rendition of aggregative consequentialism, the formal components have the following default settings: the domain rule is that we should aggregate uniformly over everything (the entire cosmos); the aggregation rule is that the total of value of the domain is the cardinal sum of the value of its parts; and the selection rule is the one given by standard decision theory applied to ethics: you ought to perform one of those available acts for which the expected value of the world is maximal. It is easy to see that these specifications lead immediately to paralysis. Assigning a finite probability to the world being canonically infinite, we find that the expected amount of positive value (and also the expected amount of negative value) in the world is the same for all humanly possible actions. In fact, if the total positive value is of the same infinite cardinality as the total of the negative value, the net value of the world is undefined. Since no humanly possible action gets

assigned a value that is higher than a value assigned to any other humanly possible action, the selection rule has no basis on which to discriminate right and wrong actions: all come out as being ethically on a par.

Attempting to salvage aggregative consequentialism by modifying the substantive component is not promising, because any plausible kind of local phenomenon is infinitely instantiated in a canonically infinite world. We shall therefore explore how we might modify one or more of the formal components, to see whether it is possible to meet the infinitarian challenge. To be successful, a solution would have to meet several criteria and desiderata, including the following:

Criteria and desiderata

• *Resolving infinitarian paralysis.* It must not be the case that all humanly possible acts come out as ethically equivalent.

• *Avoiding the fanaticism problem.* Remedies that assign lexical priority to infinite goods may have strongly counterintuitive consequences.[7]

• *Preserving the spirit of aggregative consequentialism*. If we give up too many of the intuitions that originally motivated the theory, we in effect abandon ship.

• *Avoiding distortions*. Some remedies introduce subtle distortions into moral deliberation. (This will be illustrated later.)

One open methodological question is over what range of situations an acceptable ethical theory must avoid giving prescriptions that are intuitively implausible. By the most stringent standard, an ethical theory should be rejected if there is any possible case about which it makes some (sufficiently) implausible prescription. According to those who hold to this standard, an ethical theory can be refuted by (coherently) describing a case – be it ever so improbable, far-fetched, or even physically impossible – and showing that the theory implies something shocking or perverse about that case. Others may adopt a more lenient standard and be willing to accept an ethical theory so long as it gives intuitively sensible advice in cases that are at least somewhat realistic, including all the cases that there is a non-trivial probability of us actually encountering. These people may be willing to trade off some goodness of fit to case-specific intuitions in return for a gain in some other theoretical virtue such as simplicity or completeness. An even laxer standard would require only that a moral theory normally gives non-perverse recommendations in the cases that we are most likely to encounter.

In its standard version, aggregative consequentialism fails to meet even this laxest standard of acceptability. In *all* actual situations that we are encountering, we should assign a non-zero probability to the world being canonically infinite; and as we have seen, this leads to infinitarian paralysis. Modified versions of aggregative consequentialism can be judged as more or less suc-

cessful depending on (a) the degree to which they satisfy the abovementioned criteria and desiderata, and (b) the scope of this degree of satisfaction, i.e., whether it obtains in all possible worlds, or only in physically realistic situations, or only in the empirically most plausible and typical cases.

Can the infinitarian challenge be met by being more careful about how we formulate aggregative ethics (and the other imperiled ethical theories)? Does the challenge revolve around some mere technicality that can be overcome by enlisting some more suitable formalism? Or does the challenge drive a fatal stake through the heart of a large family of ethical theories that have been widely discussed and widely embraced for several hundred years? By contrast to more familiar objections against utilitarianism (and other aggregative consequentialist theories), the alleged consequence that it is ethically indifferent whether we cause or prevent another holocaust is a bullet that even its most hardened supporters are presumably unwilling to bite. If one must either accept *that* or switch to some other ethical theory, the choice should be easy.

To answer these questions, we must patiently analyze the various responses available to the aggregative consequentialist. The next three sections examine the following possible modifications of the theory's formal components:

Modifying the aggregation rule
Default setting:
• Cardinal arithmetic
Candidate modifications:
• Extensionist program
• Value-density
• Hyperreals

Modifying the domain rule
Default setting:
• Universal domain
Candidate modifications:
• Discounting
• Causal approach

Modifying the selection rule
Default setting:
• Standard decision theory
Candidate modifications:
• Buck-passing
• Extended decision rule
• Infinity shades
• Class action

A final section examines the effects of combination therapies that involve the concurrent modification of multiple formal components.

2. Modifying the Aggregation Rule

We shall consider three possible substitutes for the standard cardinal aggregation rule: the extensionist program, the value-density approach, and the hyperreal framework.

2.1. The Extensionist Program

One approach to modifying the aggregation rule attempts to extend axiology by introducing rules for ranking worlds that contain infinite goods. This extensionist program is the only strategy for reconciling aggregative ethics with the possibility of an infinite world that has received significant attention in the literature.[8]

A value-bearing part of a world is called a *location*. Candidates include experiences, acts, persons, space-time regions, and lives. Consider a world that contains an infinite set of locations each having some finite non-zero positive value k, and another infinite set of locations each having a finite negative value $-k$. In cardinal arithmetic, the sum of value in such a world is undefined.[9] The same holds for worlds that in addition to these two sets of locations also contain locations of varying values, and for many worlds that do not contain an infinite number of locations of some constant value.[10] Canonically infinite worlds fall into the category of worlds in which the net cardinal value is undefined.

To see how the extensionist program tries to avoid paralysis, let us first consider the simple case presented in Example 1. It represents two possible worlds, each containing one immortal person who each day enjoys either a moderate or a high level of well-being. The locations are days in this person's life, each of which has either one or two units of value.

w1: 2, 2, 2, 2, 2, 2, 2, 2, …
w2: 1, 1, 1, 1, 1, 1, 1, 1, …

Example 1

There is an intuitive sense in which w1 is better than w2. Clearly, most people would prefer to live in w1, where one's level of well-being is greater. The two worlds have the same locations, and w1 has everywhere strictly more value than w2. For analogous reasons, if we change the gloss on Example 1 so that the locations, instead of being days in the life of an immortal, represented the entire lives of an infinite number of individuals, a plausible verdict would be that w1 is still better than w2. The worlds, in this alternative example, would have the same people and everybody would be better off in w1 than in w2.

Here is a simple principle that captures this intuition:[11]

> *Basic Idea*. If w1 and w2 have exactly the same locations, and if, relative to any finite set of locations, w1 is better than w2, then w1 is better than w2.

The Basic Idea is weak (although not uncontroversial).[12] Consider Example 2, where one location is a notch better in the second world:

w1: 2, 2, 2, 2, 2, 2, 2, 2, …
w3: 1, 3, 1, 1, 1, 1, 1, 1, …
Example 2

Since neither of these two worlds is better than the other relative to all finite sets of locations, the Basic Idea falls silent. (E.g., w1 is better relative to the singleton set of the first location, while w3 is better relative to the singleton set of the second location.)

To deal with cases like Example 2, Peter Vallentyne and Shelly Kagan, in an elegant paper that built on and extended the earlier literature, proposed several strengthenings of the Basic Idea, the first one of which can be re-formulated as follows, omitting a technical complication that is irrelevant to our investigation.[13]

> *SBI1* (strengthened basic idea 1): If (1) w1 and w2 have exactly the same locations, and (2) for any finite set of locations there is a finite expansion such that for all further expansions, w1 is better than w2, then w1 is better than w2.

This ranks w1 better than w3, because w1 is better than w3 relative to any finite set that includes at least three locations. The point of SBI1 is that it enables us to judge one world as better than another even if it is worse at a finite number of locations, provided that it is sufficiently better at the other locations to compensate for its regional inferiority.

SBI1 is still quite feeble. In particular, it fails to rank world pairs in which each world is better than the other at infinitely many locations. This possibility is illustrated in Example 3, where we have also added a time index for the days in the immortal's life.

w4: 3, 2, 3, 2, 3, 2, 3, 2, …
w5: 4, 0, 4, 0, 4, 0, 4, 0, …
Time: 1, 2, 3, 4, 5, 6, 7, 8, …
Example 3

Vallentyne and Kagan propose a strengthening of SBI1 that applies to cases in which the locations have what they call an "essential natural order." They suggest that spatial and temporal regions, but not people or states of nature, arguably have such an order.[14] Let us suppose that the locations in

Example 3, i.e. the days in the life of the immortal, have an essential natural order. It is intuitively plausible, if one had to choose between w4 and w5, that one ought to choose w4. One reason that could be given for this judgment is that for any time after the third day, the immortal will have enjoyed strictly more well-being in w4 than in w5. This reason, however, does not apply to the closely related case (Example 4) where the immortal has always existed, i.e. is everlasting in the past as well as the future time direction.

w6: …, 3, 2, 3, 2, 3, 2, 3, 2, …
w7: …, 4, 0, 4, 0, 4, 0, 4, 0, …
Time: …, -2, -1, 0, 1, 2, 3, 4, 5, …
Example 4

In Example 4, there is no time such that the immortal has enjoyed a greater amount of well-being by that time in w6 than in w7. At any time, a countably infinite quantity of well-being has already been enjoyed in both worlds. Nevertheless, there is an intuitive ground for holding that w6 is better than w7. The temporal density of value (average value per unit of time) in w6 is 2.5, while in w7 it is merely 2. In any finite (continuous) time interval lasting at least four days, w6 contains strictly more value than does w7. The proposed strengthening of SBI1, to deal with such cases, can be rendered in simplified form as follows.

SBI2: If (1) w1 and w2 have exactly the same locations, and (2) for any bounded region of locations there is a bounded regional expansion such that for all further bounded regional expansions w1 is better than w2, then w1 is better than w2.

This principle judges w6 as better than w7. SBI2 ranks a fairly broad set of pairs of worlds containing infinite quantities of value, and it does so in a way that is intuitively plausible. SBI2 is the about strongest principle that the extensionist program has offered to date.[15]

2.2. Shortcomings of the Extensionist Program

As a response to the threat of infinitarian paralysis, the extensionist program suffers from at least three shortcomings.

First, SBI2 applies only when values are tied to locations that have an essential natural order. Yet for many aggregative ethical theories, the primary value-bearers are not spatial or temporal locations, but experiences, preference satisfactions, people, lives, or societies. It is not at all clear that these value-bearers have an ethically relevant essential natural order. It is true that people (and experience etc.) are located in time and space, and that time and space arguably have an essential natural order.[16] But the fact that people exist in spacetime does not imply that an ethical theory that says that people are

locations of good is therefore entitled to help itself to the supposed essential natural ordering of these times and places where the people are. To attach fundamental ethical significance to the spatiotemporal ordering of people is a substantial commitment – one that is not necessarily compatible with other core features of ethical theories.

For example, it is fair to say that classical utilitarianism rejects (in spirit, if not explicitly) the notion that any fundamental ethical significance is attached to facts about *where* people live. One central motivating intuition in traditional utilitarian thinking is that "everybody counts for one and nobody for more than one", i.e. that characteristics such as skin color, position in society, or place of birth are of no fundamental ethical importance; and that what matters, rather, is something like pleasure and pain, or happiness and unhappiness, or preference satisfaction and frustration. To admit spatiotemporal location as a morally relevant consideration in addition to these traditional factors would be to make a substantial departure from the intuitions that originally inspired the theory.

To appreciate what is at stake here, note that the betterness relation specified by SBI2 is not invariant under permutation of the values of locations. Consider Hotel Hilbert (example 5), which has infinitely many rooms, each containing one occupant. Each occupant has a level of well-being corresponding to either zero or one unit of value. Let us suppose, in order to be able to appeal to principles relying on the notion of an essential natural ordering, that we take the locations of this world to be the rooms in Hotel Hilbert and that the sequence of the room numbers constitutes a natural order.

w8: 1, 1, 0, 1, 1, 0, 1, 1, 0, …
w9: 1, 0, 0, 1, 0, 0, 1, 0, 0, …
Room #: 1, 2, 3, 4, 5, 6, 7, 8, 9, …

Example 5

SBI2 says that w8 is better than w9. This might seem intuitively plausible since in w8 two out of every three rooms have happy occupants whereas in w9 only every third room has a happy occupant. Yet there exists a bijection (a one-to-one mapping) between the local values of w8 and w9.[17] That is to say, w8 can be obtained from w9 by having the guests swap rooms in such a way that all guests are still accommodated, no new guests are admitted, all rooms continue to be occupied, *and everybody's well-being remains exactly the same as it was before*. Applying SBI2 to Example 5 commits one to the view that a world can be worsened or improved without making anybody in the least bit better or worse off. This implication is in direct conflict with classical utilitarianism and other welfarist theories.

Some aggregative ethical theories, however, might be able to accommodate this implication. Infinitudes are notoriously counterintuitive. Maybe we should

see it as a relatively minor concession for aggregative ethics to admit that the spatiotemporal pattern of goods and bads can make a moral difference in certain infinite contexts. A theory that admitted this could still preserve many other features associated with a welfarist outlook. It could maintain, for example, that no person has greater ethical status than any other, that personal identity lacks fundamental moral significance (anonymity), and that in the finite case the value of a world is simply the sum of the values of its locations. All this would be consistent with the claim that relocating an infinite number of people can make a moral difference.

An alternative is to reject SBI2 and fall back on SBI1. Since SBI1 makes no use of the notion of an essential natural order, its rankings are invariant under one-to-one permutations of the values assigned to locations. Yet while this would ease some of the tensions, it would also remove the theory's ability to handle a wide range of problem cases, including examples 3-5. SBI1 fails to resolve the paralysis.

Even if we allow ourselves to use the stronger principle SBI2, many world-pairs will remain unranked. This is the second shortcoming of the extensionist approach. Even the strongest principles available fail to rank all possible worlds in order of goodness. Vallentyne and Kagan suggest a way to further strengthen SBI2 to cover some cases in which the locations have an essential natural order in more than one dimension, and some – but not all – cases in which the two worlds to be compared do not have exactly the same locations. Additionally, their strongest principle is silent about cases in which a single location has infinite value. It is also silent about cases of the sort illustrated in Example 6, where the essential natural order has a more complex structure.

w10: 2, 2, 2, 2, …, …, 1, 1, 1, 1
w11: 1, 1, 1, 1, …, …, 1, 2, 1, 1

Example 6

The worlds in example 6 have order type $\omega + \omega^*$.[18] Intuitively, w10 is better than w11, since it is better (by one unit of value) at an infinite number of locations and worse (also by one unit) at only a single location. Yet SBI2 is silent, because there is a bounded region (e.g. the one consisting of the sole location where w11 has value 2) for which there is no bounded regional extension such that w10 is better than w11 relative to that expansion (or any further bounded regional expansions thereof). So w10 is not ranked as better than w11; nor, of course, is w11 is ranked as better than w10.[19]

The extensionist program, therefore, has not provided a general cure for paralysis even if we accept that the spatiotemporal distribution of value can have ethical significance. Some things could be said in defense of the extensionist program at this point. One could express hope that further progress will be made. Assuming that our intuitions about the relative goodness of

worlds with infinite values are coherent, it is possible to construe the extensionist program as an open-ended project that could in principle codify all the rankings that are implicit in our intuitions. Whenever we encounter a world-pair not yet addressed by our principles, we could simply add a clause expressing our intuitive judgment about the new case. Even if we never find an explicit principle that covers all possible cases about which we have definite intuitions, this failing could written off as a symptom of our cognitive limitations rather than counted as a fundamental problem for the underlying theoretical framework. Furthermore, one might think that even if the extensionist program does not succeed in addressing all possible cases, it would still provide at least a partial remedy if it managed to cover a wide range of cases including the cases we are most likely actually to confront.

Whatever hope these remarks might suggest, however, is disappointed when we notice that even if the extensionist program somehow managed to succeed on its own terms, all it would have produced is an *ordinal* ranking of worlds. A completed extensionist program would give a criterion, which, for any pair of possible worlds, would say either which world is better or that they are equally good. But it would fail to tell us *how much* better one world is than another. This is the program's third shortcoming.

Since we are not omniscient beings, we make our moral choices under uncertainty. When doing so, we need to take into account not only the outcomes that would actually result from any acts we are choosing between but also the range of possible outcomes that we think would have some non-zero probability of resulting. More specifically, we must consider the conditional probabilities of the various possible outcomes given that a particular act is performed. Standard decision theory tells us that we should multiply these conditional probabilities with the value associated with the corresponding outcomes, and that we ought to do one of the acts for which the expectation value is maximal. (We postpone to later sections discussion of alternative decision rules or domains of evaluation.) In order to perform this operation we need a cardinal measure of the value of worlds. A mere ordinal ranking, telling us which worlds are better than which but not by how much, does not combine in the right way with probabilities, and fails to enable us to calculate the conditional expectation of value.

To illustrate this point, consider example 7.

w12: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, …
w13: 1, 2, 7, 4, 5, 10, 7, 8, 13, 10, 11, 16, …
w2: 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …

Example 7

By SBI2, w12 is better than w13; so if the choice is simply between these two worlds, we know that we ought to choose w12. But suppose the choice

is between the two acts *A* and *B*. Act *A* is guaranteed to realize w13. Act *B* will realize world w12 with probability *p* and w2 with probability (1 – *p*). In order to decide what we ought to do, we need to know how large *p* has to be for the choice of *B* to be as good as, or better than, the choice of *A*. The ordinal rankings given by principles like SBI2 do not provide this information.

Without a cardinal representation of the values of worlds, therefore, we have no expected values of acts – or even an ordinal ranking of acts in order of their moral goodness – when we are uncertain what results our actions will have. Nor would it help much if we had a cardinal measure of the value of worlds for a proper subset of the worlds, such as for the worlds whose value is finite. Vicious gaps in our expected value calculations would then arise whenever we assigned a non-zero probability to the world being the sort for which no cardinal value measure was available. To obtain useable moral guidance, more is therefore needed than the ordinal ranking of worlds that the extensionist program aims to provide.

In the case of individual prudential decision-making, a classic result by von Neumann and Morgenstern showed that it is possible to construct a cardinal utility function for an individual on the basis of a sufficiently rich set of ordinal preferences.[20] This result is a cornerstone of social choice theory. The method used by von Neumann and Morgenstern relies on individuals having preferences not only about specific outcomes, but also about gambles that may deliver any of a range of outcomes with varying odds. Given certain assumptions, it is possible to derive a utility function from such preferences about gambles. In order to apply a similar approach to construct a cardinal scale of the ethical value of different possible worlds, we would need to assume the existence of an ordinal ranking of the ethical value of various gambles over possible worlds (such as in example 7). But to assume such an ordinal ranking would be to help ourselves to precisely what is problematic. In the ethical case, by contrast to the case of individual prudential decision-making, we cannot appeal to a simple revealed preference account. Not everybody might be disposed to make the same evaluations of the ethical merits of the possible gambles. The requisite ranking must instead come from an axiological theory; and if that axiology is aggregative, we face the threat of paralysis.

We conclude: (1) existing principles provided by the extensionist program fail to rank all possible worlds; (2) in order for there to be any hope of it providing such a complete ranking (in a way consistent with our intuitions about the betterness relation), ethical significance would have to be attached to the spatiotemporal distribution of goods and bads; (3) for many classical aggregative theories, it is doubtful that they are compatible with the stipulation that the spatiotemporal pattering of local values has ethical significance; and (4) even if we had a complete ordinal ranking of all specific possible worlds

(excluding situations involving complex moral gambles), we still would not have solved the problem of infinitarian paralysis.[21]

2.3. The Value-density Approach

One way to get a cardinal measure of the value of some canonically infinite worlds is by looking at their "average value", their value-density. To apply this idea, we need to assume that the cosmos is, at a sufficiently large scale, approximately homogeneous. We can then define its value-density as follows. Arbitrarily select a spacetime point *p*, and consider a hypersphere centered on *p* with a finite radius *r* (where *r* is a spacetime interval). If V^+ is the (finite) amount of positive value within this hypersphere, and V^- is the (finite) amount of negative value, we can define the value-density of the sphere to be $\hat{V}(p,r) = (V^+ - V^-)/|r|$, where $|r|$ is the magnitude of *r*. If there is some constant *k* such that for any *p* we have

$$\lim_{r\to\infty} \hat{V}(p,r) = k\,,$$

then we can regard *k*, the value-density, as an index of the value of such a world, for the purpose of comparing it with other homogeneous canonically infinite worlds of the same order-type. Since value-density is a cardinal, it can be multiplied with probabilities and combined with standard decision theory (with value-density taking the place of utility).

Aggregationism, however, is committed to the view that the total amount of good matters. Not all worlds with the same value-density are equally good. A large world with positive value-density is better than a small world with the same value-density. Unless one accepts a non-aggregative view such as average utilitarianism, therefore, one could not in general identify the value of a world with its value-density. Canonically infinite worlds with positive value-density would have to be ranked as lexicographically superior to all finite-value worlds; and canonically infinite worlds with negative value-density as lexicographically inferior to all finite-value worlds.

Just as did SBI2, the value-density approach places ethical significance on the spatiotemporal distribution of value. It would thus require sacrificing a common aggregationist intuition. Furthermore, the value-density approach fails to apply to inhomogeneous infinite worlds (such as w12 and w13), because value-density is undefined for such worlds. If a single location can have infinite value, it also fails for worlds that have such singularities. It also fails for worlds that have a more complicated order-type (such as w10 and w11).

In conclusion, the value-density approach, by contrast to the extensionist program, at least provides a cardinal measure. Yet it is woefully incomplete, and it entails placing ethical significance on the spatiotemporal arrangement

of values. It also faces a fanaticism problem similar to that of the EDR approach, which we shall discuss later. Nevertheless, even though the value-density approach is a non-starter considered on its own, it will be worth considering as an ingredient in some of the combination therapies that we shall examine in section 5.

2.4. Introducing Hyperreal Numbers

Howard Sobel ends a recent book chapter on Pascal's Wager with a comment on a paper by Roy Sorenson in which the latter discussed some problems for infinite decision theory:

> It is remarkable that he [i.e. Sorenson] does not consider the hyperreal option by which decision theory can, without any adjustments, be reinterpreted to accommodate the very big, and for that matter the very small. Of that option, I sing, Let it be, for it works and is done.[22]

We are in possession of a well-developed mathematical theory of the so-called hyperreal numbers, numbers that can be infinitely large or infinitesimally small. Hyperreals can be multiplied by, divided by, added to, or subtracted from ordinary real numbers (such as probabilities) in a natural manner. But Sobel's remark concerns the application to decision theory, where the desirabilities of basic outcomes are exogenous variables. That is, decision theory pertains to well-formulated decision problems where the payoff function is already given. If the payoff function has hyperreal values in its domain, it may be easy to transpose decision theory into a hyperreal framework so that it can process these values. The task for aggregative ethics is more complicated, however, for it must also specify a mapping from worlds to the (ethical) value of these worlds. Placing aggregative ethics in the framework of the hyperreals is *not* a completed project; indeed, it has perhaps never even been attempted. In the next sub-section, however, we will make such an attempt and see how far it can get us.[23]

While we lack the space for a thorough introduction to hyperreal numbers, it may still be useful to start with a thumbnail sketch of some of their properties before we consider the application to infinitarian paralysis.[24]

The study of hyperreal numbers, their functions and properties, is known as nonstandard analysis. The hyperreals are an extension of the reals. Among the hyperreals, there are many (infinitely many) different "infinitely small" numbers, "infinitesimals" – numbers that are greater than zero but smaller than any non-zero real number. There are also infinitely many different infinitely large hyperreal numbers that are greater than any real. In addition, for every real number r, there are infinitely many hyperreal numbers r' that are "infinitely close" to r, i.e. such that the difference $|r-r'|$ is infinitesimal.

(By contrast, any two different real numbers are at a finite distance from one another.) Hyperreals thus stand very densely packed together, and they extend all the way up to infinitely large sizes as well as all the way down to infinitesimally small quantities. These properties make them a promising framework for analyzing ethical problems involving infinite values.

Hyperreals are defined in such a way that all statements in first-order predicate logic that use only predicates from basic arithmetic and that are true if we quantify only over reals are also true if we extend the domain of quantification to include hyperreals. For example, for any numbers a, b, c, that are in the field ${}^*R$ of hyperreals, we have the following familiar properties:

1. Closure
If a and b are in ${}^*R$, then $a + b$ and $a * b$ are both in ${}^*R$

2. Commutativity
$a + b = b + a$ and $a * b = b * a$

3. Associativity
$(a + b) + c = a + (b + c)$ and $(a * b) * c = a * (b * c)$

4. Distributivity
$a * (b + c) = (a * b) + (a * c)$

5. Existence of identity or neutral elements
There exist elements z and e in ${}^*R$ such that $a + z = a$ and $a * e = e$

6. Existence of inverses
There exist elements, $(-a)$ and a^{-1}, for every a such that $a + (-a) = 0$ and [for a $\neq 0$] $a * (a^{-1}) = 1$

Another example of a statement that carries over to nonstandard analysis is that if you add 1 to a hyperreal you get a bigger number:

7. $a < a + 1$

Nevertheless, R and ${}^*R$ do not behave identically. For instance, in ${}^*R$ there exists an element w that is larger than any finite sum of ones:

8. $1 < w$, $1 + 1 < w$, $1 + 1 + 1 < w$, $1 + 1 + 1 + 1 < w$, ...

There is, of course, no such number w in R. Notice that the nonexistence of w cannot be expressed as a first-order statement (as doing so would require an infinite conjunction).

The hyperreals can be constructed as countably infinite sequences of reals in such a way as to satisfy the above axioms. This construction has the convenient feature that we can identify the real number r with the sequence $(r, r, r, \ldots)$. Addition of two hyperreals can then be defined as $(a_0, a_1, a_2, ...) + (b_0, b_1, b_2, ...) = (a_0 + b_0, a_1 + b_1, a_2 + b_2, ...)$, and, analogously for multiplication, $(a_0, a_1, a_2, ...) * (b_0, b_1, b_2, ...) = (a_0 * b_0, a_1 * b_1, a_2 * b_2, ...)$.

Using this construction, here is one example of an infinite hyperreal:

(1, 2, 3, …)

and here is an infinitesimal hyperreal, its inverse:

(1/2, 1/3, 1/4, …)

The product of these two numbers equals the (finite) number (1, 1, 1, …). As implied by (6), for any infinite hyperreal, there is an infinitesimal hyperreal such that their product equals unity.

There are many differently sized infinite and infinitesimal hyperreals. For example, the following two hyperreals are, respectively, strictly larger and strictly smaller than the above couple:

(3, 4, 5, …)
(1/10, 1/12, 1/14…)

To compare the size of two hyperreals, we make a pairwise comparison of their elements. We want to say that one hyperreal is larger than another if it is larger in at least "almost all" places, and we want the resulting ordering to be complete, so that for every two hyperreals a, b, it is the case that $a > b$, or $b > a$, or $a = b$. The definition of "almost all" needed to make this work, however, is technically somewhat complicated and involves the selection of a so-called non-principal (or "free") ultrafilter. Independently of which ultrafilter is chosen, we have that if a is larger than b everywhere except for a finite number of places, then $a > b$. Similarly, if a and b are identical in all places save for a finite number of places, then $a = b$. But for some choices of a and b, a may be larger than b in an infinite number of places, and b may be larger (or equal) to a in an infinite number of other places. For instance, this is the case for the pair

a = (1, 0, 1, 0, …)
b = (0, 1, 0, 1, …)

The role of the non-principal ultrafilter (whose technical definition need not concern us here) is to adjudicate such cases so that we get a complete ordering of all hyperreals.

This quick survey might be enough to convey some feel for the hyperreals. Their main use in mathematics is in providing an alternative ("nonstandard") foundation for analysis, developed by Abraham Robinson in the 1960s, which is closer to the original ideas of Newton and Leibniz than the "epsilon-delta limit" approach that is the common fare in introductory calculus courses today. Some people find this alternative approach more intuitive, and some theorems are easier to prove within the nonstandard framework. But to return to the concern of this paper, let us consider how the introduction of the hyperreals might help solve the problem of imperturbable infinities.

2.5. The Hyperreal Approach

For a start, we need a way to map a world containing some distribution of local values to a corresponding hyperreal that represents the total value of that world. The most straightforward way of doing this would be by mapping the value at a location to a real number in the sequence of a hyperreal. To illustrate, we reuse an earlier example:

w1: 2, 2, 2, 2, 2, 2, 2, 2, …
w2: 1, 1, 1, 1, 1, 1, 1, 1, …
Example 1

The simple-minded suggestion is that we should assign to these two worlds overall values equal, respectively, to the hyperreals (2, 2, 2,…) and (1, 1, 1,…). Since (2, 2, 2,…) > (1, 1, 1,…), this would imply that w1 is better than w2. So far, so good.

Unfortunately, this approach quickly gets stuck on the fact that hyperreals whose sequences differ in only a finite number of places are of the same magnitude. Thus, for instance, the hyperreal associated with

w3: 1, 3, 1, 1, 1, 1, 1, 1, …

is of exactly the same magnitude as that associated with w2. In fact, '(1, 1, 1, …)' and '(1, 3, 1, 1, 1, …)' are merely different names for the same hyperreal, just as '1/3' and '9/27' are different names for the same real. If we can change the value of a world in at most a finite number of locations, then on this approach we cannot change the total value of a world at all. The value of a canonically infinite world is as imperturbable as ever.

If non-standard analysis is to be of any help, we need a different way of mapping worlds to values. A promising approach is to modify the previous idea by postulating that each real in the sequence of the hyperreal should be the sum of the value of the world at the corresponding location and of the real in the preceding place in the hyperreal's sequence.[25] If the local values in a world have a one-dimensional essential natural order which is infinite in one direction, (v_1, v_2, v_3, v_4, …), its value will thus be represented by the hyperreal (v_1, v_2+v_1, $v_3+v_2+v_1$, $v_4+v_3+v_2+v_1$, …).

To illustrate this, the values of w1, w2, and w3 would be represented by the following hyperreals, respectively:

Value(w1) = (2, 2+2, 2+2+2, 2+2+2+2, …) = (2, 4, 6, 8, …) = ω * 2
Value(w2) = (1, 1+1, 1+1+1, 1+1+1+1, …) = (1, 2, 3, 4, …) = ω
Value(w3) = (1, 1+3, 1+3+1, 1+3+1+1, …) = (1, 4, 5, 6, …) = ω + 2

If we consider a world that is like w3 except its extra-good location is moved one step to the right:

w3*: 1, 1, 3, 1, 1, 1, 1, 1, …

we find that its value is identical to that of w3:

Value(w3*) = (1, 1+1, 1+1+3, 1+1+3+1, …) = (1, 2, 5, 6, …) = ω + 2

This approach also handles worlds with unboundedly large local values, such as the following:

w14: 1, 3, 5, 7, 9, …

which gets assigned the hyperreal value (1, 4, 9, 16, 25, …) = ω^2.

By assigning hyperreal values to worlds in this manner, some ethical decision problems could be easily resolved. For a simple example, consider the choice between act *A* which with certainty realizes w3 and act *B* which with probability *p* realizes w2 and with probability (1 – *p*) realizes the following world:

w15: 1, 4, 1, 1, 1, 1, 1, …

Since w15 is assigned the hyperreal value ω + 3, the expected values of the two acts are:

EV(*A*) = ω + 2
EV(*B*) = (ω * *p*) + (ω + 3)(1 – *p*)

Consequently, *B* is better than *A* if and only if *p* is less than 1/3.

This "finite-sum" version of the hyperreal approach thus has some things going for it. We can extend it to deal with some cases where the locations do not have the essential natural order-type of the natural numbers. To handle worlds where the past as well as the future is infinite (e.g. w6 and w7) as well as multidimensional worlds, we can use the following procedure. First, consider a hypersphere of finite volume *v*, centered on the decision-making agent, and let the sum of the values of locations within this hypersphere define the first place of the hyperreal. Then, expand the hypersphere to volume 2*v*, and let the sum of value within this larger hypersphere define the second place of the hyperreal; and so forth, so that the value within the hypersphere of volume *n***v* defines the *n*th place of the hyperreal. (If this expansion hits a boundary of the spacetime manifold, then simply continue to expand the hypersphere in the directions that remain open.)

2.6. Costs and Limitations of the Hyperreal Approach

Some cases are not covered even when the hyperreal approach is augmented with the finite-volume expansion method. These include cases where individual locations have infinite value (whether $\aleph_0$ or of an even higher cardinality). Worlds with the order-type of w16,

w16: 7, 7, 7, 7, 7, 7, …, 7, 7, 7, 7, 7, 7, …

can be accommodated by assigning them a value equal to the sum of the value of the their two parts (in this case, (ω*7) + (ω*7) = ω*14), but if we consider a world like w17, where the decision-maker is located in the first segment of ones,

w17: 1, 1, 1, 1, 1, 1, …, …, -2, -1, 0, 1, 2, 3, …

we run into trouble because we then lack a preferred location in the second segment on which to center the constant-volume expansion. Without such a preferred location, the value of the segment is undefined. To see this, suppose we start expanding in both directions from the location that has the value –1. We then get the following hyperreal value (for the second segment):[26]

(-1, -3, -5, -7, …) = (-ω * 2) + 1

whereas if we start from the location that has the value +1, we get instead an infinitely larger hyperreal:

(1, 3, 5, 7, …) = (ω * 2) - 1

There are thus some gaps in the hyperreal approach, although these are smaller than for the value-density approach.[27]

Another problematic aspect of the hyperreal approach concerns the use of a non-principal ultrafilter. In non-standard analysis, the choice of ultrafilter is arbitrary. For the purposes of pure mathematics, this multiple instantiability causes no problem. It might be thought undesirable, however, to have such arbitrariness embedded in the foundations of axiology. Depending on the choice of ultrafilter, two worlds can come to be ranked as equally good, or as the first being better than the second, or as the second being better than the first. This is illustrated in example 8.

w18: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
w19: 1, -2, 1, 1, -2, 1, 1, -2, 1, 1, -2, 1, 1, …

Example 8

These two worlds are assigned the following hyperreal values:

Value(w18) = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …)
Value(w19) = (1, -1, 0, 1, -1, 0, 1, -1, 0, 1, -1, 0, …)

The values in w18 are equal to those in w19 at an infinite number of locations, greater at an infinite number of locations, and smaller at an infinite number of locations. Depending on which ultrafilter we select, we can therefore get either Value(w18) = Value(w19), or Value(w18) > Value(w19), or Value(w18) < Value(w19).

A fan of the hyperreal approach could argue that this indeterminacy is actually a virtue in disguise because it matches a similar indeterminacy in our intuitions about the relative merits of w18 and w19. (This judgment would conflict with the value-density approach, which would rank w18 and w19 as determinately equally good.) Alternatively, the proponent of hyperreals could issue a promissory note stating that additional constraints could be specified that would remove the indeterminacy and select a unique ultrafilter. But note that once a particular ultrafilter is selected, it will be possible to alter the value assigned to a world by rearranging (an infinite number of) its values. For example, suppose an ultrafilter is selected that agrees with the value-density approach (whenever the latter make definite verdicts). Given such an ultrafilter, w18 and w19 would be assigned the same hyperreal. But if the values of w19 were rearranged in such a way as to alter its value-density, then the resulting world would be assigned a different hyperreal, making it either better or worse than w18. Like SBI2 and the value-density approach, therefore, the hyperreal approach incurs the cost of placing ethical significance on the spatiotemporal distribution of values as soon as we fix on a specific ultrafilter.

Among paralysis remedies that focus on altering the aggregation rule alone, the hyperreal approach is the most powerful one to date. We now turn our attention to a class of remedies that focus on modifying the domain rule.

3. Modifying the Domain Rule

Rather than aggregating all local values, we could postulate that only a subset of them should be aggregated, or that local values should be weighted in some way before the aggregation rule is applied to them. Clearly, some rationale would have to be given for such domain modifications – a completely arbitrary and unnatural restriction would destroy whatever intuitive support aggregative consequentialism can boast. Here we shall examine two alternative domain rules: discounting and the causal approach.

3.1. Discounting

By discounting values that are spatiotemporally remote from the decision-maker, one could prevent dangerous infinities from arising in certain cases. The cost of this option, however, is forbidding. To be effective, it would have to involve both temporal and spatial discounting. Temporal discounting, as an alleged feature of fundamental axiology, rather than as a merely practically convenient proxy, is often viewed with great suspicion. *Spatial* discounting has been seen as a patent absurdity. Thus, Derek Parfit:

> Remoteness in time roughly correlates with a whole range of morally important facts. So does remoteness in space… But no

> one suggests that, because there are such correlations, we should adopt a Spatial Discount Rate. No one thinks that we would be morally justified if we cared less about the long-range effects of our acts, at some rate *n* percent per yard. The Temporal Discount Rate is, I believe, as little justified.[28]

In any case, adopting a spatiotemporal discount factor would do little to solve the paralysis problem. For any given discount factor, we can consider worlds that have, centered on the decision-maker, a sequence of locations whose values increase at a faster rate than the discount factor discounts them, so that the sum of discounted values is infinite. To avoid this, we would have to postulate that the discount rate at some point becomes infinite, creating an ethics-free zone at some finite distance from the decision-maker – making a travesty of aggregative consequentialism (and even this would not help if it is possible for a single location to have infinite value). We shall therefore not further pursue the discounting approach.[29]

3.2. The Causal Approach

Instead of maximizing the expected goodness of the world, we could aim to maximize the expected goodness of the causal consequences of our acts.

The basic idea here is simple: while evidential decision theory suffers ethical paralysis in a canonically infinite world, causal decision theory might avoid having to confront infinities if we are certain that the values at the locations that we can causally affect are finite. Causal decision theory, as we shall interpret it here, directs us to evaluate the *changes* we can bring about.[30] Since the values of these changes may be finite even if the value of the world as a whole is infinite, a domain rule that restricts the evaluation to the changes we affect may thus prevent the ethical significance of our acts from being washed away in the limitless sea of values that stretch out beyond our sphere of influence.

Consider example 9. There is a line, infinite in both directions, of alternately happy and unhappy people. You have the choice between increasing the well-being of person p_3 by one unit (w20) and leaving things as they are (w21).

w20:	…, 1, -1, 1, -1, 2, -1, 1, -1, 1, …
w21:	…, 1, -1, 1, -1, 1, -1, 1, -1, 1, …
Person:	…, p_{-1}, p_0, p_1, p_2, p_3, p_4, p_4, p_5, p_6, …

Example 9

Since you can affect the value of only one location (p_3), the causal domain rule allows aggregative consequentialism to deliver the verdict that you ought to choose the act that realizes w20.

3.3. Problems with the Causal Approach

Although this causal domain rule is itself quite agreeable, its marriage with aggregative consequentialism would not be entirely without tension. One consequence of the causal approach is that there are cases in which you ought to do something, and ought to not do something else, even though you are certain that neither action would have any effect at all on the total value of the world. Example 9 is a case in point, since the cardinal sum of value of both w20 and w21 is undefined. The implication that you ought to "do good" even when doing so does not make the world better must, from the standpoint of the aggregative consequentialist, be regarded as a liability of the causal approach. But suppose we are willing to pay this price – how much protection against ethical paralysis does it buy?

An advocate for the causal approach might point out that, according to relativity theory, nobody can influence events outside their future light cone. Cosmology suggests that the number of value-bearing locations (such as lives, or seconds of consciousness etc.) in our future light cone is finite. Given our best current physics, therefore, the causal approach appears to avoid paralysis.

Not so fast. Basing our ethics on an empirical fact about the laws of nature means that it cannot satisfy the highest methodological standard (cf. section 1). Well, we might be able to live with that. But the situation is much worse: the causal approach fails even in the situation we are actually in, thus failing to meet even the lowest possible acceptability criterion for a moral theory. This is because reasonable agents might – in fact, should – assign a finite non-zero probability to relativity theory and contemporary cosmology being wrong. When a finite positive probability is assigned to scenarios in which it is possible for us to exert a causal effect on an infinite number of value-bearing locations (in such a way that there is a real number $r>0$ such that we change the value of each of these location by at least r), then the expectation value of the causal changes that we can make is undefined.[31] Paralysis will thus strike even when the domain of aggregation is restricted to our causal sphere of influence.

3.4. Tweaking the Causal Approach

We could attempt to avoid this problem arising from our subjective uncertainty about the correctness of current physics by stipulating that the domain of aggregation should be restricted to our future light cone even if, contrary to special relativity, we could causally affect locations outside it. With this stipulation, we could ignore the physically far-fetched scenarios in which faster-than-light influencing is possible.

This tweak is not as good as it may appear. If, contrary to what current physics leads us to believe, it is in fact possible for us (or for somebody else,

perhaps a technologically more advanced civilization) to causally influence events outside our (their) future light cone, moral considerations would still apply to such influencing. According to the present proposal, we should not factor in such considerations even if we thought superluminal propagation of causal influence to be quite likely; and that is surely wrong.

Moreover, even if the propagation of our causal effects is limited by the speed of light, it could still be possible for us to influence an infinite number of locations. This could happen, for instance, in a spatially infinite cyclic space-time or in a steady-state cosmology.[32]

In conclusion, the causal approach entails that we should do good even when doing so does not make the world better; and even if we accept this, we still do not avoid paralysis. Adopting the causal domain rule, while leaving the other formal components of aggregative consequentialism in their standard form, fails to cure the problem.[33]

4. Modifying the Selection Rule

The third of the formal components in aggregative consequentialism is the rule that selects, on the basis of the aggregation of local value in the designated domain, what an agent ought to do. The default version of this component is the selection rule given by standard decision theory applied to the ethical case: it says that we ought to perform one of the available actions that maximize expected aggregate value. In this section, we shall examine three possible replacements for this selection rule: the extended decision rule ("EDR"), the aggregate act approach ("class action"), and ignoring small probabilities ("infinity shades"). But first, we devote a subsection to considering the radical strategy of simply omitting the selection rule in aggregative ethics.

4.1. Passing the Buck from Ethics to Decision Theory

In view of all the problems discussed above, it might be tempting to scale back our ambition. Maybe aggregative ethics should content itself with specifying an ordinal ranking of the goodness of worlds, such as the one the extensionist program aims to provide. A right act could then be defined as one that *in fact* leads to as good a world as any other available act does. Ethics would lay down the success criteria for an act being morally right but would leave it to the judgments of individuals, aided perhaps by decision theory, to figure out how best to go about trying to meet these criteria. On this view, the fact that the practical decision problem has not yet been solved for all infinite cases should not be held against aggregative ethics.

This buck-passing strategy fails to address the underlying problem. If decision theory works fine in finite cases but not in infinite cases, then as far as decision theory is concerned the right conclusion might simply be that

all the values entering into the decision procedure in any particular case must have finite upper and lower bounds. Such a requirement need not even be externally imposed. As we have already noted, it emerges naturally in subjectivist decision theory founded on the concept of revealed preferences. If an individual is unwilling to gamble everything on a tiny (but finite) chance of getting to Heaven, then that means that she does not have an infinitely strong preference for Heaven. Decision theory simply reflects this fact. Aggregative ethical theories, on the other hand, defy finitistic strictures because their core commitments imply that certain worlds and outcomes be assigned infinite value. This implication is not an innocent or neutral feature that can be ignored in an evaluation of the plausibility of these theories. If some ethical theories refuse to play ball by making impossible demands on decision theory, then that must count against those theories and in favor of other ethical theories that can be integrated in workable ways with decision theory.

What if we went further and invoked not only an actualist ethics but also an actualist decision theory saying simply that we ought to decide to perform one of the actions that in fact would have the best moral results? Then so long as the results can be ranked in order of their moral goodness, a complete specification would have been given in the sense that for each decision problem, including infinite ones, there would be a set of correct choices. But this is a plainly pseudo-solution. It is easy enough to stipulate that we ought to decide to do an act that will in fact have the best moral consequences, yet this injunction is non-actionable if we lack an effective way of figuring out which of the feasible acts are better in this actualist sense, and which are worse. At one place or another, the subjectively possible outcomes (including ones that would not actually occur) must be taken into account, along with their subjective probabilities. This is necessary for all real-world agents, who are operating under conditions of uncertainty about what the world is like and about what the consequences of their actions would be. The problem of paralysis results from the ethical claim that values are aggregative. While the problem can be delegated to decision theory, or to some account of practical deliberation, it must ultimately be confronted. And aggregative ethics, having made the original claim, must accept responsibility if the problem turns out to be unsolvable.[34]

Since we cannot, then, omit the action rule, let us consider how we may revise it.

4.2. The Extended Decision Rule

This idea begins with the observation that even though we might have reason to think that the world is infinite, and even if the world is in fact infinite, we should still assign some positive probability to it being finite. If it really does not matter what we do in the infinite case – because of our inability to

change the infinite values that would be present in such a case –then maybe we ought to focus instead on the finite case. Could aggregative ethics salvage itself by clinging to this slender reed, the mere subjective possibility that the world is finite? This strategy requires that we let go of the view that a right act is one that maximizes the expected value of the world. We then need some other way of deciding what we ought to do.

Consider two possible acts: A^+, which is intuitively good (feeding the starving), and A^-, which is intuitively bad (committing genocide). Next, consider two possibilities: S_{fin}, the world contains only a finite amount of positive and negative value, and S_{inf}, the world is canonically infinite. For simplicity, let us assume that these are the only feasible acts and the only possible ways for the world to be. If S_{inf} obtains then we cannot affect the total value of the world and, by assumption, it is ethically irrelevant what we do; that is, conditional on S_{inf}, neither of A^+ and A^- is morally preferable to the other. But if S_{fin} obtains then our conduct does make an ethically significant difference; for conditional on S_{fin}, A^+ is strictly better than A^-. Hence A^+ dominates A^-, since in no contingency is A^+ worse than A^- and in some contingency is A^+ strictly better than A^-. This, one could argue, constitutes a moral reason for doing A^+ rather than A^-.

The dominance principle underpinning this reasoning boils down to this: "If you can't make any difference to the value of the world if the world is infinite, then focus on the finite case and do what would maximize expected value given that the world is finite." We can generalize that idea into the following decision rule.

The Extended Decision Rule (EDR)

Let $P(\infty^+ \mid A)$ be a subjective probability assigned by the agent to the proposition that the world contains an infinite amount of good and at most a finite amount of bad, conditional on act A. Let $P(\infty^- \mid A)$ be the probability that the world contains an infinite amount of bad and at most a finite amount of good, conditional on A. Let Ω be the set of feasible acts for the agent at the time of the decision; we assume that this set is finite.[35]

1. For each act $A_i \in \Omega$, consider $P(\infty^+ \mid A_i) - P(\infty^- \mid A_i)$, and let $\Omega^* \subseteq \Omega$ be the subset of acts for which this difference is maximal. If Ω^* contains only one act, then the agent ought to perform that act (and other acts are wrong).
2. If there is more than one act in Ω^* then consider, for each such act $A_i \in \Omega^*$, the expected value of the world conditional on A_i & S_{fin}. Then all acts for which this expected value is maximal are right (and other acts are wrong).

EDR draws on two intuitions. First, it incorporates a generalized version of a principle suggested by George Schlesinger in the context of Pascal's Wager, that "when each available line of action has infinite expected value, then

one is to choose that which is most probable to secure the reward."[36] EDR generalizes (or corrects) this by stating one should rather maximize *the difference* between the probability of securing an infinite reward and the probability of incurring an infinite penalty. Second, EDR allows finite values to serve as tiebreakers when considerations about infinite values cancel out. EDR can be further generalized to cover cases where values of different infinite orders are at stake. This could be done by adding stages to the decision procedure such that values of greater infinite cardinality are always given lexical priority over lower-level values, the latter serving as tiebreakers.[37]

Does EDR solve the problem of infinitarian paralysis? The hope would be that, with EDR, aggregative consequentialism can avoid paralysis even if we believe, truly and with good reason, that the world is canonically infinite. So long as there is a non-zero subjective probability of the world being finite, and our feasible acts are on a par vis-à-vis infinite values, EDR directs our attention to the possibility that only finite values are involved. Relative to this possibility, intuitively bad acts such as genocide will generally be rated as inferior to intuitively good acts such as feeding the starving. Thus, plausible ethical advice could ensue. Moreover, EDR seems consistent with the basic motivations behind aggregative ethics. It can be viewed as a conservative adaptation rather than a radical revision of traditional positions.

4.3. The Fanaticism Problem

One feature of EDR that is attractive at first sight but disturbing on closer inspection is the strict priority given to maximizing the probability that the world will contain some infinite good and minimizing the probability that it will contain some infinite bad. *Any* shift in the difference $P(\infty^+ \mid A_i) - P(\infty^- \mid A_i)$, however tiny, will justify the sacrifice of any finite value, however large. If there is an act such that one believed that, conditional on one's performing it, the world had a 0.00000000000001% greater probability of containing infinite good than it would otherwise have (and the act has no offsetting effect on the probability of an infinite bad), then according to EDR one ought to do it even if it had the certain side-effect of laying to waste a million human species in a galactic-scale calamity. This stupendous sacrifice would be judged morally right even though it was practically certain to achieve no good.

We are confronted here with what we may term *the fanaticism problem*. One aspect of this problem is that EDR would in some cases force us to judge that it would be morally wrong to decline gambles like the one just described. It is difficult to bring oneself to endorse this (if one seriously contemplates what it would involve). Yet if it is only in extremely improbable and far-fetched scenarios that we would have to countenance such gambles, then we might perhaps – while conceding that the EDR would give counter-

intuitive advice in such abnormal cases – nevertheless accept aggregative consequentialism (with the EDR modification), provided that we hold a methodological view of the acceptability conditions for an ethical theory that has tolerance for such rare "failures" (cf. section 1.3).

It is therefore relevant to consider a second aspect of the fanaticism problem: it seems that EDR recommends that *in our actual situation* we should become almost obsessively concerned with speculative infinite scenarios.[38]

If EDR were accepted, speculations about infinite scenarios, however unlikely and far-fetched, would come to dominate our ethical deliberations. We might become extremely concerned with bizarre possibilities in which, for example, some kind of deity exists that will use its infinite powers to good or bad ends depending on what we do. No matter how fantastical any such scenario would be, if it is a logically coherent and imaginable possibility it should presumably be assigned a finite positive probability,[39] and according to EDR, the smallest possibility of infinite value would smother all other considerations of mere finite values.

Aggregative consequentialism is often criticized for being too "coldly numerical" or too revisionist of common morality even in the more familiar finite context. Suppose that I know that a certain course of action, though much less desirable in every other respect than an available alternative, offers a one-in-a-million chance of avoiding catastrophe involving x people, where x is finite. Whatever else is at stake, this possibility will overwhelm my calculations so long as x is large enough. Even in the finite case, therefore, we might fear that speculations about low-probability-high-stakes scenarios will come to dominate our moral decision making if we follow aggregative consequentialism.

Yet the infinite vista, even though it might seem like a more remote concern, in fact makes the fanaticism problem worse. It makes it worse in two ways. First, it increases the likelihood of our actually confronting a situation in which low-probability-high-stakes scenarios overwhelm our calculations. The suggestion that we must make some great moral sacrifice for the sake of realizing some vast yet finite good (such as saving a billion people) is in practice often defeated by the consideration that the chances of success are too small. But if the potential good to be gained is infinite, then no finite positive success probability is too small. To show that the potential gain would not be worth the sacrifice, one would have to show that the probability of success is smaller than any positive real number; and it doubtful that such extreme confidence in our impotence could *ever* be justified. And second, the infinite vista may also make the fanaticism problem worse by increasing the counterintuitiveness of the actions that would be recommended by aggregative consequentialism in those (perhaps ubiquitous) situations wherein the infinite considerations dominate. In other words, the allegedly fanatic acts that aggregative consequentialism would prescribe in the infinite case may

be "crazier" than the corresponding fanatic acts it would prescribe in the finite case.

Proponents of the EDR could bite the bullet and resolve to accept the implications of giving absolute priority to infinite considerations whatever these implications might be. But if these implications are counterintuitive, they count against the theory; and if they are *too* counterintuitive, they count decisively against the theory. So what, precisely, would one be committed to do if one embraced the EDR?

Proponents of the EDR – or of other variations of aggregative consequentialism, such as the hyperreal approach, that also give absolute priority to infinite concerns – may argue that the practical upshot of adopting the theory is not as radical as we might fear. In particular, they might argue that considerations about different infinite possibilities cancel each other out. For each bizarre scenario in which an available act that is intuitively wrong leads to the realization of an infinite good, we can imagine an equally probable and opposite scenario in which the same act leads to the realization of an infinite bad. Only if we had some discriminating information about infinite scenarios would taking them into account alter the outcome of our deliberations. Since in fact we lack such information (so the argument would go), we can safely ignore infinite scenarios and focus our attention on the finite scenarios about which we do have some relevant information, and use these to determine what we ought to do. On the basis of these ordinary finite considerations, we can then safely conclude, for instance, that gratuitous genocide is wrong and that feeding the starving is a much more promising candidate for being morally right.[40]

This argument delivers the reassuring conclusion that giving absolute priority to infinite concerns does not engender unacceptably fanatic prescriptions about what we ought to do. However, the argument rests on a dubious assumption. It presupposes that the infinite scenarios we could concoct cancel each other out *exactly*. Each scenario in which some action increases the probability of obtaining an infinite good must be matched by some other scenario in which the same act decreases the probability of the obtainment of the infinite good by the same amount (or offset the difference by an increase in the probability of obtaining an infinite bad). This cancellation of probabilities would have to be *perfectly accurate*, down to the nineteenth decimal place and beyond.

While it might sound plausible to say that, in the real world, we have no discriminating information about which infinite scenario might materialize conditional on our taking some particular action rather than another, this is true only if we speak roughly. The epistemic probabilities that enter into the calculation can be sensitive to a host of imprecise and fluctuating factors: the estimated simplicity of the hypotheses under consideration, analogies (more or less fanciful) derived from other domains of our changing experience, the

pronouncements of miscellaneous authorities, and all manner of diffuse hunches, inklings, and gut feelings. It would seem almost miraculous if these motley factors, which could be subjectively correlated with infinite outcomes, always managed to conspire to cancel each other out without remainder. Yet if there is a remainder – if the balance of epistemic probability happens to tip ever so slightly in one direction – then the problem of fanaticism remains with undiminished force. Worse, its force might even be *increased* in this situation, for if what tilts the balance in favor of a seemingly fanatical course of action is the merest hunch rather than any solid conviction, then it is so much more counterintuitive to claim that we ought to pursue it in spite of any finite sacrifice doing so may entail. The "exact-cancellation" argument threatens to backfire catastrophically.

A different strategy for arguing that the upshot of EDR is not as revolutionary as it may appear is by appealing to (what we shall term) an *empirical stabilizing assumption*. Instead of claiming that the infinite considerations cancel out, this strategy involves making the diametrically opposite claim, namely, that we have empirical grounds for thinking that infinite considerations decisively favor some courses of action over others. The idea here is that certain infinite scenarios are more likely than others, and the actions that would turn out to be best in the most likely scenarios in which these actions bring about infinite outcomes are the same actions that common morality recommends in the finite context (or are, at any rate, not too radically different from those sanctioned by common morality). We call the kind of premiss deployed in this argument an "empirical stabilizing assumption", because it serves to "stabilize" our infinite deliberations so that they point in an intuitively safe direction; and because it the assumption is "empirical" – it is false in some possible worlds and we have reason to believe it is true in the actual world only on grounds of empirical information. If a stabilizing assumption is true, then the EDR does not prescribe implausibly fanatical courses of action in the actual case. This would diminish the fanaticism problem. Perhaps it would diminish the problem sufficiently to render the resulting theory acceptable, at least if we adopt a relatively lax acceptability condition for our ethics (cf. section 1.3).

4.4. Empirical Stabilizing Assumptions

We will not attempt comprehensively to review the possible stabilizing assumptions that some person or other might be tempted to accept. A brief discussion of two candidate assumptions – one theological, the other naturalistic – will illustrate the general idea.

A theological stabilizing assumption. Suppose we were convinced that the (by far) most likely scenario involving infinite values is one featuring a deity

akin to the Judeo-Christian God but who operates on a principle of collective responsibility. In this scenario, *Homo sapiens* is the only intelligent species in a finite universe. If we collectively exceed a certain threshold of (traditionally understood) moral merit, God will reward us all with infinitely long lives in Heaven; but if we fail to do so, He will soon bring on the apocalypse and end the world. Given this view, EDR seems to suggest that we ought to act in accordance with traditional morality and encourage others to do likewise. Disturbing or counterintuitive forms of fanaticism are avoided.

A naturalistic stabilizing assumption. Suppose we were convinced that the (by far) most likely scenario involving infinite values goes something like follows: One day our descendants discover some new physics which lets them develop a technology that makes it possible to create an infinite number of people in what otherwise would have been a finite cosmos.[41] If our current behavior has some probabilistic effect, however slim, on how our descendants will act, we would then (according to EDR) have a reason to act in such a way as to maximize the probability that we will have descendants who will develop such infinite powers and use them for good ends. It is not obvious which courses of action would have this property. But it seems plausible that they would fall within the range acceptable to common sense morality. For instance, it seems more likely that ending world hunger would increase, and that gratuitous genocide would decrease, the probability that the human species will survive to develop infinitely powerful technologies and use them for good rather than evil ends, than that the opposite should be true. More generally, working towards a morally decent society, as traditionally understood, would appear to be a good way to promote the eventual technological realization of infinite goods. Note that the relevant magnitude here is not the absolute probability of success but the relative probability compared to that attainable through alternative courses of action. We need *not* assume that it is probable that our descendants could develop infinitely powerful technologies, *nor* that it is probable that we could determine whether they would do so if they were able to, *nor* that we could influence how they would use them if they did develop such technologies.

These are two possible stabilizing assumptions. Unless, however, one is willing to nail one's colors to the mast of some such assumption, the fanaticism problem remains a threat. Since it is difficult to determine what course of action would be best from the infinite standpoint, one would need to brace oneself for the possibility of unpalatable deviations from traditional morality.

One could argue that in our current situation, where we are still so much in the dark about infinite scenarios, what EDR would recommend is that we make the investigation of such scenarios a top priority. We cannot rule out that further attention to infinite prospects from such fields as cosmology, theology, philosophy, or futurology would produce some useful ideas. Even

the slightest illumination of this topic would, given the EDR, be of enormous value as it would help us refine our understanding of what our practical aims ought to be. However, the EDR would not necessarily imply that we should drop everything else in order to study these fields; for indirect contributions would also count. Contributing to a fairer, richer, and more educated society may be an efficient way of promoting the long-term objective that an informed understanding of remote infinite prospects will eventually come to guide public policy.

All things considered, therefore, the practical upshot of accepting the EDR may not be all that radical.[42] Nevertheless, even if the particular acts that would be recommended fall within the bounds of acceptability, it may be counterintuitive that the ultimate justification for these acts would be grounded in the imperative that we seek enlightenment about infinite prospects. This kind of justification may seem too brittle and too rationalistic to be the real reason why, say, it is wrong to commit genocide.

In defense, a proponent of the EDR could argue that there may be nothing especially problematic about the esoteric nature of this ultimate justification. Were we to require that our ethical theories be such that most people already understand and accept them and use them to guide their moral conduct, we would have set ourselves an impossible task – most people have never even heard of Kantian ethics, perfectionism, emotivism, contractarianism, or any other systematic ethical or meta-ethical theory. Moreover, even according to an aggregative consequentialism that used the EDR, at least the *proximate* grounds for the morality of particular acts would be *exoteric* and available to common sense. For example, genocide is wrong because it unjustly inflicts great harms on large numbers of people, leads to strife and war and large-scale destruction of resources, forfeits opportunities for cooperation, trade, and friendship, and so forth. Only the ultimate ground is esoteric, i.e. that these consequences of genocide are bad because they would decrease the magnitude $[P(\infty^{+} \mid A_i) - P(\infty^{-} \mid A_i)]$.

In conclusion, the problem with the EDR (and other principles giving lexical priority to infinitarian considerations) is twofold. First, it makes everything depend on tenuous speculations about infinite possibilities. This is in itself disturbing. Suppose we obtained increasingly strong evidence that the world is canonically infinite and that the already weak probabilistic link between our acts and the world's expected amount of goodness and badness becomes even more attenuated. It is counterintuitive to claim that whether any act would be morally wrong would depend on there still being a tiny subjective probability that the universe is (despite all appearance) finite or that our acts correlate, however weakly, with infinite values. Could the wrongness of genocide really dangle on such a dainty thread? Second, there is the problem of fanaticism, which is resolved or mitigated only by making some

empirical stabilizing assumption, and even then it is unclear whether the resulting prescriptions are acceptably close to our common morality.

4.5. Infinity Shades

Having played with subtle considerations involving far-fetched scenarios in which our acts improbably correlate with infinite outcomes, let us now examine a very different selection rule, one that dispenses with such considerations altogether. Could we meet the challenge of infinitarian paralysis by postulating that, in determining what we ought to do, we should simply ignore all very unlikely possibilities?[43]

As a piece of pragmatic advice, the notion that we should ignore small probabilities is often sensible. Being creatures of limited cognitive capacities, we do well by focusing our attention on the most likely outcomes. Yet even common sense recognizes that whether a possible outcome can be ignored for the sake of simplifying our deliberations depends not only on its probability but also on the magnitude of the values at stake. The ignorable contingencies are those for which the *product* of likelihood and value is small. If the value in question is infinite, even improbable contingencies become significant according to common sense criteria. The postulation of an exception from these criteria for very low-likelihood events is, at the very least, theoretically ugly.

If one were nevertheless tempted by this maneuver, the question would arise of just how small a probability must be in order for it to be negligible. If the threshold is set too high, it will have the unacceptable implication that low-probability events such as nuclear accidents should be ignored in moral reasoning. But if the threshold is low enough to include this type of event, it is also low enough to include scenarios involving infinite values. We reasonably assign a greater probability to the world being canonically infinite than to any particular nuclear reactor undergoing a core melt.

There is another way to remove infinite values from consideration: rather than introduce a low-probability threshold, simply stipulate that all possibilities involving infinite values be ignored, independently of their probability. On this view, a right act would be one that maximizes the expected value of the world, but where this expectation is calculated by omitting all possible worlds that contain infinite value. We can dub this selection rule "infinity shades." A proponent of this theory could argue that since our moral intuitions were formed in contexts that did not involve confrontation with infinite values, it would be unsurprising if these intuitions could be captured fairly well by a theory that brackets off infinite values from consideration.

The exemption clause specified by infinity shades adds an ungraceful epicycle to the standard selection rule. Yet so long as we consider only intuitions about what we ought to do – as opposed to intuitions about what

the theoretical structure of our ethical theories should be – it seems plausible that the theory may succeed in matching intuitions fairly well (setting aside, of course, any mismatches that that do not stem specifically from the possibility of infinite values but arise from other aspects of aggregative consequentialism.) Discrepancies would arise in hypothetical cases in which there is a close and direct connection between our acts and infinite values. In such cases, clearly, it is counterintuitive to say that we ought to disregard infinite considerations. However, if we are unlikely ever to confront such cases, then the theory might at least pass the muster of a moderate acceptability standard.

Pretending that worlds with infinite value have zero probability can, however, also lead to another more subtle kind of moral distortion. Consider the following thought experiment:

The Research Council
Your task is to allocate funding for basic research, and you have to choose between two applications from different groups of physicists. The Oxford Group wants explore a theory that implies that the world is canonically infinite. The Cambridge Group wants to study a theory that implies that the world is finite. You believe that if you fund the exploration of a theory that turns out to be correct you will achieve more good than if you fund the exploration of a false theory. On the basis of all ordinary considerations, you judge the Oxford application to be slightly stronger. But you use infinity shades. You therefore set aside all possible worlds in which there are infinite values (the possibilities in which the Oxford Group tends to fare best), and decide to fund the Cambridge application. Is this right?

If the answer is "No", then the suggestion that we ignore possibilities involving infinite values fails to meet even a low methodological standard because it gives the wrong verdict in cases like Research Council which are quite realistic (and quite likely actually to arise).

Its theoretical unsightliness combined with the fact that infinity shades would fail completely in some far-fetched hypothetical cases and would introduce distortions in more realistic cases like Research Council means that we should be very reluctant to accept this selection rule.

4.6. Class Action

The last selection rule we shall consider attempts to get leverage on infinite values by focusing not on individual acts but on some larger units to which our acts are in some suitable way related. These larger units might either be *rules*, whose general acceptance can have wide-ranging consequences which could form a basis for evaluating individual instances of rule-following, or they could be some kind of *aggregate* of individual acts or decision processes. We shall refer to both variations of this idea as "class action".

Consider first how rule-consequentialists can cope with some of the situations that paralyze act-consequentialists. The original problem was this: if we can only do a finite amount of good or bad, it seems we cannot change the total value of a canonically infinite world. Let us suppose that each agent can in fact only make a finite difference. It could nevertheless be possible for an infinitude of agents, as a collective, to make an infinite difference that changes the value even of a canonically infinite world. Rule-consequentialism, in its most rudimentary form, claims that an act is morally right if it is recommended by the set of moral rules whose "general acceptance" would have the best consequences.[44] If the population is infinite then the general acceptance of a rule could easily have infinite consequences, which we could try to use as a basis for evaluating the moral rightness of individual acts.

Another, alternative, class action selection rule might be more congenial to act-consequentialists; call it the "aggregate act approach." Consider a case in which there is an infinite number of exact copies of you spread throughout an infinite cosmos. (This case is not far-fetched; it is, in fact, empirically plausible.[45]) Now, suppose that we conceive of "you" in a broader sense than usual – as not just this particular creature but instead as the aggregate of all physical copies of you throughout the cosmos. Let us refer to this distributed aggregate entity by using capitalized letters: "YOU." Then, even though your actions may have only finite consequences, YOUR actions will be infinite. If the various constituent person-parts of YOU are distributed roughly evenly throughout spacetime, then it is possible for YOU to affect the world's value-density. For example, if each person-part of YOU acts kindly, YOU may increase the well-being of an infinite number of persons such that the density of well-being in the world increases by some finite amount.[46]

One positive argument that could be made on behalf of this aggregate act approach is that it can draw support from evidential decision theory. What one part of YOU decides is relevant information about what other parts decide. The expected (evidential) value of your saving a drowning child includes not only the value of this particular child's rescue but also the expected value deriving from the evidential linkage between your decision and the decisions of other parts of YOU, decisions that may have consequences for an infinite number of drowning children. Your deciding to save this child gives you evidence that the other parts of YOU will make analogous decisions to save drowning children in similar circumstances throughout the world. The expected value-density of the world conditional on your saving this particular child can therefore exceed (by a non-infinitesimal amount) the expected value-density of the world conditional on your not saving this child. Evidential decision theory, combined with the value-density aggregation rule, then enables aggregative consequentialism to deliver the implication that you have moral reason for saving the child, notwithstanding that the world is assumed to be canonically infinite. (Causal decision theorists will of course be unmoved

by this argument, but if the strategy could cure ethical paralysis they might embrace a suitably modified version as a distinctively moral postulate, just as they could accept ethical rule-consequentialism without supposing that this theory receives any special support from causal decision theory.)

Neither the aggregate act nor the rule-consequentialist version of the class action selection rule would, *on its own*, cure infinitarian paralysis. The cardinal sum of value in a canonically infinite world is undefined, and would remain undefined even under the action of an infinite collective of similarly situated agents. At best, class action would be useful as an adjunct to some other treatment modality, such as the value-density aggregation rule. In the next section, we shall study this and some other potential combination therapies.

5. Combination Therapies

We have considered several single-point interventions. If we prune the least promising ones (the extensionist program, the discounting approach, and the buck-passing strategy), we retain the following shortlist:

Aggregation rules
• Cardinal arithmetic (default)
• Value-density
• Hyperreals
Domain rules
• Universal domain (default)
• Causal
Selection rules
• Standard decision theory (default)
• Extended decision rule (EDR)
• Infinity shades
• Class action

Even this shortlist offers a large number of combinatorial possibilities. Before surveying the intricate effects that arise from some of these potential multi-modal interventions, let us recall the key evaluation criteria. These are: resolving infinitarian paralysis, avoiding the fanaticism problem, preserving the spirit of aggregative consequentialism, and avoiding distortions. We have already seen that selecting the default options throughout the list fails to satisfy the first criterion, leading to total ethical paralysis.

Our survey in this section must be highly condensed since there are so many possibilities to consider. We can divide the task into two parts, corresponding to whether significance is attached to the spatiotemporal patterning of values.

5.1. If Spatiotemporal Pattern Is Not Morally Significant…

Suppose that we refuse to assign ethical significance to the spatiotemporal distribution of local values. Then we must reject the value-density and the hyperreal approaches, leaving us with the default aggregation rule, cardinal arithmetic.

The introductory section of this paper showed that the combination of cardinal arithmetic with the other two default options leads to ethical paralysis, and sections 2-4 considered the effects of combining cardinal arithmetic with single modifications of either the domain rule or the selection rule. It remains to examine what happens if we make several simultaneous modifications of the domain or selection rules. In this context, we can set aside the class action selection rule, since it does not get along well with cardinal arithmetic. This leaves us with the following four possible combinations:

(1) Cardinal arithmetic + EDR + Infinity shades
(2) Cardinal arithmetic + EDR + Causal
(3) Cardinal arithmetic + Infinity shades + Causal
(4) Cardinal arithmetic + EDR + Infinity shades + Causal

Note that whenever infinity shades are used, EDR reduces to standard decision theory. Only options (2) and (3), therefore, present substantially new alternatives.

The addition of the causal domain rule has parallel effects in (2) and (3): it incurs a cost and produces a potential benefit.

The cost is the same in both cases: the spirit of aggregative consequentialism is compromised to some degree. The restriction of the domain of aggregation to the causal consequences of our actions entails that an action might be definitely right or wrong even though it is known to have no impact whatever on the total value of the world.

The potential benefit, in both cases, is that the addition of the causal domain restriction enables the resultant theory to make do with a different set of empirical stabilizing assumptions. In order for this to be a real benefit, however, the new stabilizing assumptions would have to be substantially more plausible than the stabilizing assumptions required by the original theory; and it is very questionable whether this is so.

To see why, first consider (2). The stabilizing assumption needed if we use the EDR *without* the causal restriction is that we are unlikely to find ourselves in a situation in which infinite values are subjectively correlated with our actions in such a way as to make the theory's prescriptions unacceptably "fanatical" in the sense discussed earlier. *With* the causal restriction, the required stabilizing assumption is that we are unlikely to find ourselves in a situation in which causal effects involving infinite values are subjectively correlated with our actions in such a way that fanatical prescriptions ensue.

These two types of inadmissible situation are subtly different, yet it is hard to see any reason for believing that a situation of the latter type is much less likely to arise than one of the former. In lieu of a reason for believing this, adding the causal restriction to the EDR does not provide any significant benefit.

A similar point can be made with regard to (3). If we use infinity shades *without* the causal restriction, we must make the stabilizing assumption that we are unlikely to find ourselves in a situation in which the bracketing-off of the possibility that the world contains infinite value leads to unacceptable distortions of our pre-theoretic common moral reasoning. Situations featuring such distortions arise, for example, in scenarios in which there is a close link between the realization of infinite value and our actions, and in scenarios like Research Council. *With* the causal restriction, the required stabilizing assumption is that we are unlikely to find ourselves in a situation in which the bracketing-off of the possibility that we could *causally affect* infinite value would lead to distortions. The latter stabilizing assumption might be more plausible than the former. A reason for thinking so is that it is less probable that we could cause infinite values than that the world contains infinite values.

In summation, under the assumption that the spatiotemporal distribution of value is irrelevant, the only potentially useful combination therapy is adjoining the causal domain rule to the infinity shades selection rule. Cost: some additional compromise of the spirit of aggregative consequentialism. Benefit: some lessening of the demands on the required stabilizing empirical assumption.

5.2. If Spatiotemporal Pattern Is Morally Significant…

If the value of a world is allowed to depend on the patterning of local values, then the value-density and the hyperreal aggregation rules become permissible options.

Let us start with combinations involving the class action selection rule. Note that the value-density approach is useless without class action, since a human individual is almost certainly incapable of affecting the value-density of a canonically infinite world. Only when combined with class action does value-density become a contender.

The possible gain from using the class action rule is subtle and relates to the fanaticism problem. The class action rule does not prevent an arbitrarily small chance of an infinite value from smothering all considerations of finite value. What it does is change the nature of the scenarios wherein infinite values are at stake. *Without* class action, the most likely such scenarios are far-fetched – in our earlier discussion we mentioned a theological and a techno-futuristic scenario. *With* class action, the most likely scenarios in which infinite values are at stake are more ordinary: namely, that the world is canonically

infinite and contains an infinite set of agents whose activities constitute the class action. This could make a difference. If the class action rule is adopted, then while the far-fetched infinite scenarios still smother the finite scenarios, these far-fetched infinite scenarios are in turn smothered by mundane "class action-infinite" scenarios. And what makes sense given the mundane "class action-infinite" scenarios is, one might hope, similar to what makes sense given that the world is finite. For example, your saving of a drowning child would be probabilistically linked, given the class action rule, to an infinite value (the saving of an infinite number of children throughout the canonically infinite world). Since this link is much stronger than any speculative link between our current actions and, say, the decisions made by a hypothetical future civilization that has developed infinitely powerful technologies, we can ignore considerations about such far-fetched scenarios for most practical purposes. This reduces the severity of the fanaticism problem.

This advantage would be annulled if we allowed the possibility of infinite values greater than those that are linked in mundane ways to the class action of the infinite population. If the possibility of such higher-order infinite values were admitted, and were allowed to trump lower-order infinite values, then the class action rule's rationale for saving the drowning child would again be smothered by speculative considerations about far-fetched scenarios in which our actions are correlated with higher-order infinite values. This would bring the fanaticism problem back in an especially malignant form.

If we are interested in using the class action selection rule, therefore, we might want to bar higher-order infinite values (and hence reject EDR). To do this, we could use a modified version of infinity shades, e.g. a version that postulates that we should ignore, not all infinite scenarios, but all infinite scenarios containing higher-order infinite values. As we saw earlier, the use of infinity shades risks creating distortions. However, the risk is reduced when only scenarios involving higher-order infinite values are filtered out, since we are less likely to find ourselves in situations in which such higher-order infinite values are directly and strongly linked to our actions.

It thus remains to analyze what happens if we add the causal domain rule to one of the following four combinatorial possibilities:

(1) Value density + Class action + Higher-order infinity shades
(2) Hyperreal + Class action + Higher-order infinity shades
(3) Hyperreal
(4) Hyperreal + Higher-order infinity shades

Addition of the causal domain rule appears to be more or less redundant in (1) and (2). Restricting the domain of aggregation to agents' causal effects would not prevent the domain from containing infinite values when the class action rule is used. Nor would it significantly change the nature of the most

probable scenarios in which infinite values are at stake, since, given the class action rule, these will be of a mundane nature anyway.

Finally, consider (3) and (4). Combination (3), using the hyperreal modification alone (with default settings for the domain rule and the selection rule), is, strictly speaking, not a well-defined option, since there are possible worlds to which the hyperreal approach, at least such as it has been developed to date, does not apply. We noted this problem in section 2.6 and gave the examples of worlds that have an unusual order-type or have single locations containing infinite values. The simplest remedy for these kinds of gap is to invoke a kind of infinity shades that specifically postulate that those possible worlds which the hyperreal aggregation cannot handle should be ignored. This gives us (4).

With combination (4), scenarios wherein infinite values are at stake – no matter how improbable and far-fetched – trump all finite scenarios. The fanaticism problem is therefore a major concern, and considerations similar to those outlined in section 4.3 pertain in this context too. Adding the causal domain rule to (4) does little to change the situation. Since the absolute probability of the problematic infinite contingencies does not matter (so long as the probability is finitely greater than zero), we would gain no security from ignoring possible evidential correlations between our acts and infinite values outside our sphere of causal influence. For this would still leave in play the contingencies where our acts causally affect infinite values.

To summarize this sub-section, the only interesting combinatorial possibilities we have identified that involve the value-density or the hyperreal aggregation rule are (1), (2), and (4); i.e. combining either of these aggregation rules with class action and higher-order infinity shades, or combining the hyperreal aggregation rule with a version of infinity shades *sans* class action.

6. General Assessment and Conclusion

The problem of infinitarian paralysis threatens to take a large class of ethical theories out of the running. From the default interpretation of aggregative consequentialism it follows that it is always ethically indifferent what we do. This should count as a *reductio* by everyone's standards.[47] Infinitarian paralysis is not one of those moderately counterintuitive implications that all known moral theories have, but which are arguably forgivable in light of the theory's compensating virtues. The problem of infinitarian paralysis must be solved, or else aggregative consequentialism must be rejected. And if aggregative consequentialism is felled by this problem, it drags down with it a much larger class of ethical theories that contain a maximizing aggregative component.

To have any chance of avoiding this fate, aggregative consequentialism must be modified. Modifications that do not resolve the paralysis are useless.

Modifications that do resolve the paralysis face further challenges. To varying extents, these modifications compromise the spirit of aggregative consequentialism, create distortions in our moral reasoning, and induce a vulnerability to the fanaticism problem. All the cures we have examined have serious side effects.

Suppose we take the path that involves assigning moral significance to the spatiotemporal distribution of local values. This would imply that the value of a world could be changed by having people (or whatever entities are local value-bearers) swap places even though nobody is made the least bit better or worse off. If we go for this option, we then have a choice between the hyperreal and the value-density aggregation rule. We can optionally add the class action selection rule if we choose the hyperreal aggregation rule, and we must add class action if we choose value-density. But this is not enough: we must also add some kind of infinity shades to filter out those possible worlds to which the chosen aggregation rule fails to apply. (Adding the causal domain rule provides no significant benefit.)

Itemized billing (if patterning is morally significant…)
(a) Compromising the spirit of aggregative consequentialism by according ethical significance to the patterning of value
(b) The arbitrariness of the selection of a non-principal ultrafilter (hyperreal aggregation rule only)
(c) Possible further compromising of aggregative consequentialism by adding class action?
(d) The possibility of distortions resulting from the use of infinity shades
(e) The threat of a fanaticism problem
To cope with (d) and (e), we need to add an empirical stabilizing assumption to the effect that we are unlikely to encounter situations where distortions or fanaticism arise, wherefore:
(f) The resultant theory cannot meet the strictest acceptability criteria because it fails in those possible worlds in which the empirical assumption is false.

Suppose that we instead take the alternative path and refuse to accord moral significance to the spatiotemporal patterning of values. One option then is to rely on infinity shades to prevent infinitarian paralysis. But this maneuver is highly *ad hoc* and creates potentially serious distortions. Distortions can arise even in mundane cases, such as Research Council. The empirical assumption that we will not actually encounter any such situations in which distortions occur is implausibly strong. A theory patched up with such an assumption is apt to fail us occasionally. The introduction of the causal domain rule would weaken the needed empirical assumption. But this benefit would come at the cost of further compromising the spirit of aggregative consequentialism, since the causal domain rule implies that we can be obligated to perform

certain actions and to abstain from others even though what we do has no effect on the total value of the world.

Another option is to rely on the EDR. This is the strategy that would best respect the spirit of aggregative consequentialism. However, the EDR makes all moral reasoning depend on speculations about far-fetched scenarios involving correlations between our acts and infinite values; and this must be regarded as a liability. Moreover, the EDR invites a potentially pernicious fanaticism problem. A strong empirical stabilizing assumption would be needed, which means that the theory could at best satisfy only a lax methodological criterion for what makes a moral theory acceptable. It is not even entirely clear that the required empirical assumption is true. (If the strongest true empirical assumption is too weak, then the theory gives fanatical or "crazily counterintuitive" prescriptions for what we ought to do – not just in some hypothetical situation but in our *actual* situation!)

Itemized billing (if patterning is not morally significant…)
(g) Using infinity shades only: *ad hoc*, and serious distortions
(h) Using infinity shades and causal domain rule: *ad hoc*, reduced distortions, and additional compromise of aggregative consequentialism
(i) Using EDR only: all moral reasoning dependent on speculations about far-fetched scenarios, and a severe fanaticism problem.

To ward off the specters of distortions and fanaticism, we need to make an empirical assumption to the effect that these issues will not cause too much trouble in the situations we are likely to encounter, whence
(j) The resultant theory cannot meet the strictest acceptability criteria.

We see that all the solutions have substantial costs. How one weighs these up might be a matter of subjective judgment. Whether aggregative consequentialism is worth the sacrifices needed to salvage it from the infinitarian challenge depends on the merits of alternative moral theories and on other considerations that are beyond the scope of this paper.

The situation for mixed ethical theories that include non-consequentialist side-constraints in addition to an aggregative consequentialist component is slightly more hopeful. For example, a theory constructed along these lines might say that "permissible" acts are those that satisfy deontological side-constraints (no unjustified killing, lying, cheating, stealing, etc.), and that a right act is a permissible act that scores highest on the aggregative consequentialist criterion among the available permissible acts. The side-constraints would serve as a buttress, reducing the theory's dependence on empirical assumptions to avoid the fanaticism problem and the distortion problem. The theory could thus make do with a weaker empirical assumption, which means that it could potentially meet a higher methodological acceptability standard. However, unless a great many side-constraints were added, some

sort of empirical assumption would probably still be needed. How this would play out would depend on the precise nature of the deontological component. We lack the space to explore these possibilities here.

Were one to reject aggregative consequentialism as a fundamental moral theory, one may still find an important place for its core idea as a lower-level moral principle – a principle of limited validity that would be embedded in a more encompassing normative framework and that would come into play only in particular circumscribed contexts. For example, one might hold that certain institutions ought to take an aggregative maximizing stance with regard to the interests of their constituencies; or, more weakly, that such a stance reflects one type of consideration that some institutions have a duty to incorporate into their decision making. This would lead to none of the difficulties described in this paper, so long as the constituencies are necessarily finite and there is an upper bound on the amount of harm or benefit that can be imposed on any member. To the extent that utilitarian and other aggregationist ideas make their way into the world outside philosophy departments, it is usually in such a circumscribed capacity. Social choice theory often finds it convenient to proceed on the assumption that policies and social institutions exist to serve finite populations of individuals whose interests are defined in terms of their preference structures in such a way as to avoid problematic infinities. We also find echoes of aggregationist consequentialism in such policy tools as cost-effectiveness analysis, impact statements, and QALY-based evaluations of health care policies. Properly circumscribed and qualified, these real-world applications are immune to infinitarian paralysis.[48]

NOTES

1. In the standard Big Bang model, assuming the simplest topology (i.e., that space is singly connected), there are three basic possibilities: the universe can be open, flat, or closed. Current data suggests a flat or open universe, although the final verdict is pending. If the universe is either open or flat, then it is spatially infinite at every point in time and the model entails that it contains an infinite number of galaxies, stars, and planets. There exists a common misconception which confuses the universe with the (finite) "observable universe". But the observable part – the part that could causally affect us – would be just an infinitesimal fraction of the whole. Statements about the "mass of the universe" or the "number of protons in the universe" generally refer to the content of this observable part; see e.g. [1].

Many cosmologists believe that our universe is just one in an infinite ensemble of universes (a multiverse), and this adds to the probability that the world is canonically infinite; for a popular review, see [2]. The "many worlds" of the Everett version of quantum physics, however, would not in any obvious way amount to the relevant kind of infinity; both because whether the "world"-count reaches infinity or merely a large finitude might be an artifact of convenient formalism rather than reflecting of physical reality, and also because the ethical significance of each Everettian "world"

should, plausibly, be weighted by its associated measure (amplitude squared), which is a normalized; see e.g. [3].

2. If there are an infinite number of planets then there will be – with probability one – and infinite number of people, since each planet has finite non-zero chance of giving rise to intelligent life. In fact, in an infinite universe it seems that there will be an infinite number of people spontaneously materializing in gas clouds or from black hole radiation, since the probability of such occurrences, although *extremely* small, is nevertheless finitely greater than zero and would thus be expected to happen infinitely many times in a universe that contains an infinite number of gas clouds and black holes. (See e.g. [4], p. 19; but also [5].) Of course, it is vastly more probable for intelligent creatures to evolve on a planet than to spontaneously materialize by random combination of elementary particles. Infinitely many of these people will be happy, infinitely many will be unhappy, and likewise for other such local properties that pertain to person-states, lives, or entire societies or civilizations – there will be infinitely many democratic ones, infinitely many ruled by malevolent dictators, etc.

3. David Lewis' modal realism may likewise seem to imply that it is ethically irrelevant what we do. Robert Adams [42] developed an objection to Lewis' theory along such lines, arguing that our value judgments reflect the absoluteness of actuality. Lewis responded [43] that modal realism can respect many of our ethical intuitions if we construe them as containing an indexical component. Recently, Mark Heller [44] argued that Lewis' response is not fully successful. We consider some indexicalizing maneuvers in section 3.

4. For Moore, the total value, the value "on the whole", is the sum of the value of the parts and the value they have "as a whole". The infinitarian paralysis thus threatens to set in if *either* the sum of the values of the parts is infinite *or* the value that the parts have as a whole is infinite [6].

5. For some recent discussion on the Pasadena problem, see [45–46].

6. *Some* theories of prudential rationality do not get off the hook so easily.

7. Permitting extremely large *finite* values might also engender some bizarre results; see [47].

8. See e.g. [7–14]; for the "locations" terminology, see also [15].

9. To get a feel for why this is so, consider two different ways of adding up the values of the infinite number terms $+k$ and an infinite number of negative terms $-k$. If we perform the operation $(k + k - k) + (k + k - k) + \ldots$, then each bracket has the value k, and the sum of these brackets becomes infinite positive. If we instead perform the operation $(k - k - k) + (k - k - k) + \ldots$, then each bracket has the negative value $-k$, and the sum becomes infinite negative. In both these operations, a (countable) infinite number of positive and negative terms $\pm k$ would be included. By fiddling with the brackets, it is easy to see that one could make the terms add up to any positive or negative multiple of k.

10. E.g., the series (1) + (-1) + (½) + (- ½) + (¼) + (- ¼) +, …, $(\pm(1/n)2)$, … converges (to zero), and in fact it does so independently of the ordering of the terms. But if a world is such that there is some finite number $m > 0$ and an infinite number of locations with value greater than m, and an infinite number of locations with value less than $-m$, then the sum of values in that world does not converge. This is the case in a canonically infinite world.

11. [7].

12. See e.g. [16].

13. The original version is:

> *SBI1* (strengthened basic idea 1): If (1) w1 and w2 have exactly the same locations, and (2) for any finite set of locations there is a finite expansion and some positive number, *k*, such that, relative to all further finite expansions, w1 is *k*-better than w2, w1 is better than w2.

A world is "k-better" than another, relative to a given "expansion" (i.e., a set of locations) if the total value of its locations in that expansion exceeds that of the other world by at least *k* units. The complication in the second clause is designed to deal with cases involving the possibility of asymptotically converging series of values.

14. [7], p. 9.

15. Modulo some further refinements which do not affect any of the points made in this paper.

16. Or rather, *a metric*. But whether space and time really do have this property is hard to tell, because Vallentyne and Kagan do not provide any clear definition of what they mean by "essential natural" order.

17. This follows trivially from the fact that, in both w8 and w9, the cardinality of the set of ones and the cardinality of the set of zeros is the same, ℵ0.

18. The order type ω + ω* can be represented as consisting first of the natural numbers and then the natural numbers "with a tag" in reverse order, where any number with a tag is defined to be greater than any number without a tag. That is, the order type can be written as follows: $1 < 2 < 3 < 4 < \ldots < \ldots\ 4' < 3' < 2' < 1'$. This order has a smallest element, 1, and a greatest element, 1'. Starting from, e.g., element 4', one has to descend an infinite number of steps before reaching any of the untagged numbers.

19. To see why there is no bounded regional expansion containing the location where w11 has the value 2 such that w10 is better than w11 relative to this expansion, consider that by the time the expansion has reached the part where w10 is better, the expanded region has already grown to infinite size, such that both worlds have the same (infinite) amount of value in it, whence adding a finite amount of value to this infinite amount will fail to make a difference.

20. [17].

21. The hypothetically completed extensionist program would enable the determination of *objectively* right actions: these are, on an act-consequentialist view, the actions that in fact make things go best. But when we *use* a moral theory, we need to determine what is *subjectively* right do to; that is, we need to decide on an action on the basis of our actual (limited) knowledge, not on the basis of complete specification of the value of all its consequences – something a human agent never has. For a recent argument for the indispensability of a subjectivized notion of "right" (referred to by the authors as a "decision-ought"), see [37]. That paper argues against absolutist deontological theories on grounds that they cannot accommodate this decision-ought. See also section 4.1 below.

22. [18], appendix 2.2.

23. A different way of extending classical mathematics, which we shall not discuss here, is by constructing the so-called surreal numbers, first introduced by John Conway [19].

24. The *locus classicus* is [20]. For an accessible primer, see [21].

25. I am indebted here to Toby Ord (personal communication).

26. We permit ourselves some sloppiness here for the sake of ease of exposition. Since w17 has discrete locations, we use a volume of one location for the first stage in the "finite-volume" expansion, and an increment of two locations in each subsequent step.

27. Yet further complications may arise if a world contains many segments of ordered locations, but where these segments are not themselves ordered. This could be the case if there are many universes which are not anchored in any background space or externally ordered in any other way.

28. [22], p. 486.

29. Another way to discount would be not on the basis of spatiotemporal distance but on the basis of how much value a world contains at other locations. The contribution to a world's overall value that the content of a location would make would, on this view, depend on the contents of other locations, so value would not really be local. We could express this view (somewhat awkwardly) by saying that locations carry "utility" and that the world derives diminishing marginal value from utility, analogous to the way that individuals typically derive diminishing marginal utility form consumption. By postulating such a law of diminishing marginal value from utility, we could easily ensure that the total value of a world is finite, for example by stipulating that $V = e^{-U^-} - e^{-U^+}$, where U- is the amount of negative utility in the world, U+ the amount of positive utility in the world, and where the two terms are to be computed separately before they are added together. However, although infinites are here suppressed in the final assessment of the value of a world, they are still present in what we termed the utility of locations. The effect is that the value of a canonically infinite world be unalterable by us despite it being finite. (If there is already an infinite amount of positive and negative utility in the world, we cannot change it.)

30. Evidential and causal decision theory are commonly thought to coincide except for special cases, like the Newcomb problem, where our choice of action would give us information about the world that is not mediated by the causal consequences of our acts. However, they can also depart in the infinite case even when no Newcomb-like setup is involved. This point, however, is accidentally obscured by the standard formalizations of causal decision theory (e.g., [23]), which were not designed to deal with infinite cases.

31. This expectation value is the sum of terms corresponding to the set of maximally specific scenarios. Each of these terms is the product of the value of the changes we make in that specific scenario and the probability of the scenario obtaining. A term is undefined if the probability is non-zero and the value of the changes we make in that scenario is undefined. The expectation is undefined if one of its terms is undefined.

32. Although it should be noted that current data suggests a positive cosmological constant, in which case we may permanently lose causal contact with all but a rather small finite part of the cosmos within a few billion years [24].

33. Additionally, it might fail in situations (if such be possible) in which we would have an infinite number of alternatives to choose between, with finite but unboundedly good consequences, and also if there is an infinite number of alternatives with boundedly good consequences but such that for each act there is one act that is at least slightly better. (An omnipotent God choosing between different creation acts

provides an illustration of this predicament. If for every possible world there is a better one, it might seem that everything God could do would be wrong. This consequence has led some to revise the above criterion for right action (see e.g. [25]). Maybe instead we should say that any act by somebody in that position would be right, or that any act that is expected to have at least "rather good" results would count as right.

34. There are additional grounds for not expunging all aspects of probabilistic decision making from the purview of consequentialist ethics. Some acts seem wrong by almost everybody's standards even though they happen to produce good results. Suppose someone fires a bullet into an innocent man's chest with the intent of murder. The man survives and, fortuitously, the bullet hits and destroys a budding malignancy that would otherwise have killed the man within three months. The outcome is beneficial, yet most people, consequentialists included, would maintain that it was morally wrong to attempt the murder. A natural starting point for a consequentialist account of why it was wrong is that the (reasonably) expected consequences were bad even though the actual consequences were good. For a discussion of some related themes, see [26, 37].

35. In this paper we shall assume that there are only finitely many feasible acts for an agent at any one time. This is likely the case for us in the actual world, although for some possible beings (God?) this might not be true, and additional problems are known to arise for decision theory in such cases (see e.g. [27]). The subjective probabilities referred to in the EDR could be qualified as "reasonable" or "rational" credences if we wish to maintain that an ideally motivated but imperfectly rational agent might fail to choose to do the right thing. The EDR sets the bar for right action quite high: any act other than one of those for which there is no better act is classified as wrong. Defenders of aggregative consequentialism might want to supplement their theory with some account that connects it more closely to everyday notions of right and wrong, for example by taking the line that acts that are "sufficiently" good compared to the alternatives, even if they are not the very best, may often be regarded as "right" for practical purposes (albeit not perfectly right "strictly speaking"). This issue also arises for aggregative consequentialism in the finite case, so we will not pursue it here.

36. [28], p. 154.

37. Extending the EDR to recognize infinite values of different cardinalities is straightforward. It is less clear how to extend the EDR to recognize difference in infinite values of the same cardinality, since it may be not always be plausible to differentiate between these lexicographically. But a proponent of the current approach might hope that it does not make any practical difference how we deal with these finer discriminations, for reasons outlined later in this section.

38. Fanaticism problems can also arise in the finite case. Suppose a stranger approaches you on the street and offers to create x units of utility for you in return for one dollar. Does the probability that the stranger will hold up his end of the bargain always diminish at least linearly for large values of x?

39. David Lewis argues that we not only should not, but that we *could not* assign a zero probability such possibilities. See [23], p. 14.

40. For discussion of some related issues in the finite case, see [38–40].

41. Speculative scenarios of this kind have been described; see e.g. [24, 29–31]. For a parallel with the finite case, see [32].

42. Some have seen it as a problem for views such as utilitarianism that they are too demanding, quite apart from infinite considerations, because these views seem to imply that we ought devote practically all our time and resources to the world's most pressing problems. The fanaticism refereed to in the text, by contrast, does not concern the quantity of effort that aggregative theories seem to demand of us, but rather the direction in which this effort should be exerted. Traditional responses, such as stipulating that meeting a lower threshold of moral effort qualifies an agent for praise, do not address this directional form of the fanaticism problem.

43. The idea that small probabilities don't count is advocated in [33]; it is criticized in [34].

44. For a recent, more sophisticated development of rule-consequentialism, see [35].

45. See [36].

46. Refinements of this idea are possible. For instance, rather than focusing on YOU, the aggregate of all persons that are qualitatively identical to yourself, one could instead focus on the aggregate of all instantiations of a decision process that are sufficiently similar to the instantiation of the decision process whereby you are currently making your moral choice to count as instantiations of (qualitatively) "the same process". We could call these aggregates of decision-instantiations "YOUR DECISION", and proceed in the same way as suggested in the text. (I have benefited from discussions with Eliezer Yudkowsky on this point.)

47. Quentin Smith, however, seems to accept this conclusion; [41].

48. For helpful discussions and comments, I am very grateful to John Broome, Jeremy Butterfield, Tyler Cowen, Guy Kahane, Robin Hanson, Brad Hooker, Daniel Isaacson, John Leslie, Toby Ord, Mitch Porter, Rebecca Roache, Peter Singer, Howard Sobel, David Wallace, Timothy Williamson, Alex Wilkie, Eliezer Yudkowsky, and to the audience at the Aristotelian Society/Mind Association Joint Session, Canterbury, 9–12 July, 2004, and at the Oxford Moral Philosophy Seminar, 14 February 2005, where earlier versions of this paper were presented.

REFERENCES

1. Martin, J.L. (1995), *General Relativity*, 3 edn. London: Prentice Hall.

2. Tegmark, M. (2003), "Parallel Universes," *Scientific American* 288(5): 41–53.

3. Wallace, D. (2002), "Everett and Structure," *Studies in History and Philosophy of Modern Physics* 34B(1): 87–105.

4. Hawking, S.W., and W. Israel (eds.) (1979), *General Relativity: An Einstein Centenary Survey*. Cambridge: Cambridge University Press.

5. Belot, G., J. Earman, and L. Ruetsche (1999), "The Hawking Information Loss Paradox: The Anatomy of a Controversy," *British Journal for the Philosophy of Science* 50(2): 189–229.

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

7. Vallentyne, P., and S. Kagan (1997), "Infinite Value and Finitely Additive Value Theory," *Journal of Philosophy* 94(1): 5–26.

8. Segerberg, K. (1976), "A Neglected Family of Aggregation Problems in Ethics," *Nous* 10: 221–244.

9. Nelson, M.T. (1991), "Utilitarian Eschatology," *American Philosophical Quarterly* 28(4): 339–347.

10. Cain, J. (1995), "Infinite Utility," *Australasian Journal of Philosophy* 73(3): 401–404.

11. Lauwers, L. (1995), *Social Choice with Infinite Populations*. Dissertation 101. Leuven: Monitoraat E.T.E.W.

12. Mulgan, T. (2002), "Transcending the Infinite Utility Debate," *Australasian Journal of Philosophy* 80(2): 164–177.

13. Campbell, D.E. (1985), "Impossibility Theorems and Infinite Horizon Planning," *Social Choice and Welfare* 2(4): 283–293.

14. Hamkins, J.D. and B. Montero (2000), "Utilitarianism in Infinite Worlds," *Utilitas* 12(1): 91–96.

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

16. Hamkins, J.D., and B. Montero (2000), "With Infinite Utility, More Needn't Be Better," *Australasian Journal of Philosophy* 78(2): 231–240.

17. von Neumann, J., and O. Morgenstern (1944), *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

18. Sobel, H. (2004), *Logic and Theism: Arguments For and Against Beliefs in God*. Cambridge: Cambridge University Press.

19. Conway, J. (1976), *On Numbers and Games*. New York: Academic Press.

20. Robinson, A. (1966), *Non-standard Analysis*. Amsterdam: North-Holland.

21. MathForum, *Nonstandard Analysis and the Hyperreals*. http://mathforum.org/dr.math/faq/analysis_hyperreals.html

22. Parfit, D. (1984), *Reasons and Persons*. Oxford: Clarendon Press.

23. Lewis, D. (1981), "Causal Decision Theory," *Australasian Journal of Philosophy* 59(1): 5–30.

24. Cirkovic, M., and N. Bostrom (2000), "Cosmological Constant and the Final Anthropic Hypothesis," *Astrophysics and Space Science* 274(4): 675–687.

25. Adams, R.M. (1984), "Must God Create the Best?" in *Ethics and Mental Retardation*, J. Moskop and L. Kopelman (eds.). Dordrecht: Reidel Publishing Company: 127–140.

26. Broad, C.D. (1914), "The Doctrine of Consequences in Ethics," *International Journal of Ethics* 24(3): 293–320.

27. Sorenson, R. (1994), "Infinite Decision Theory," in *Gambling on God: Essays on Pascal's Wager*, J. Jordan (ed.). Savage, MD: Rowman & Littlefield, 139–159.

28. Schlesinger, G. (1988), *New Perspectives on Old-time Religion*. Oxford: Clarendon Press.

29. Tipler, F. (1994), *The Physics of Immortality*. New York: Doubleday.

30. Linde, A. (1988), "Life After Inflation," *Physics Letters B* 211: 29–31.

31. Dyson, F. (1979), "Time without End: Physics and Biology in an Open Universe," *Reviews of Modern Physics* 51(3): 447–460.

32. Bostrom, N. (2003), "Astronomical Waste: The Opportunity Cost of Delayed Technological Development," *Utilitas* 15(3): 308–314.

33. Gorovitz, S. (1979), "The St. Petersburg Puzzle," in *Expected Utility Hypothesis and the Allais Paradox*, M. Allais and O. Hagen (eds). Dordrecht: Reidel: 259–270.

34. Weirich, P. (1984), "The St. Petersburg Gamble and Risk," *Theory and Decision* 17(2): 193–202.

35. Hooker, B. (2000), *Ideal Code, Real World: A Rule-consequentialist Theory of Morality*. Oxford: Oxford University Press.

36. Bostrom, N. (2002), "Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation," *Journal of Philosophy* 99(12): 607–623.

37. Jackson, F. and Smith, M. (2006), "Absolute Moral Theories and Uncertainty," *Journal of Philosophy* 103(6): 267–283.

38. Lenman, J. (2000), "Consequentialism and Cluelessness," *Philosophy and Public Affairs* 29(4): 342–370.

9. Bostrom, N. (2007), "Technological Revolutions: Ethics and Policy in the Dark," in *Nanoscale: Issues and Perspectives for the Nano Century*. Cameron, N. and Mitchell, E. (eds.). Hoboken, NJ: John Wiley, 129–152.

40. Cowen, T. (2006), "The Epistemic Problem Does Not Refute Consequentialism," *Utilitas* 18(4): 383–399.

41. Smith, Q. (2003), "Moral Realism and Infinite Spacetime Imply Moral Nihilism," in *Time and Ethics: Essays at the Intersection*, H. Dyke (ed.). Dordrecht: Kluwer Academic Publishers, 43–54.

42. Adams, R. (1979), "Theories of Actuality," in Michael Loux (ed.), *The Possible and the Actual*. Ithaca: Cornell University Press.

43. Lewis, D. (1986), *On the Plurality of Worlds*. Cambridge, MA: Basil Blackwell.

44. Heller, M. (2003), "The Immorality of Modal Realism, or: How I Learnt to Stop Worrying and Let the Children Drown," *Philosophical Studies* 114: 1–22.

45. Nover, H., and A. Hayek (2004), "Vexing Expectations," *Mind* 111(450): 237–249.

46. Colyvan, M. (2006), "No Expectations," *Mind* 115(459): 695–702.

47. Bostrom, N. (2009), "Pascal's Mugging," *Analysis* 69(3): 443–445.