The Paralysis of Aggregative Ethics in our Possibly Infinite Universe

Nick Bostrom
Faculty of Philosophy, Oxford University
Homepage: http://www.nickbostrom.com

Abstract

Cosmology indicates that we might well be living in an infinite universe that contains an infinite number of happy and sad people. Given some assumptions, aggregative ethics implies that such a world would contain an infinite amount of positive value and an infinite amount of negative value. Yet you could presumably do only a finite amount of good or bad. Since an infinite cardinal quantity is not changed by the addition or subtraction of a finite quantity, it looks as if we incapable of making any difference to the total amount of value in the world. Aggregative consequentialism would then entail the absurd consequence that it is morally indifferent what we do. This paper examines a number of responses to this challenge and argues that none is completely satisfactory. The problem of imperturbable infinite values is serious, under-appreciated, and potentially fatal difficulty for many ethical theories which include an aggregative consequentialist component.

1. Ethics for small fish in an infinite pond

We take the morality of human affairs to be a serious matter, and yet when we look up at the starry night sky and try to think of humanity from a "cosmic point of view", then human history and all our earnest strivings, tragedies and triumphs take on an absurd and almost ridiculous appearance – as if we were ants laboring frantically to rearrange a stack of needles. In the hurly-burly of daily life and analytic philosophy, such late-night ruminations on our humble role in the scheme of things tend to be brushed aside. But is it possible that these seemingly idle thoughts could point to something of philosophical significance? In particular, might they perhaps contain some useful implication for our moral theorizing?

If the cosmos is finite then (even if it is very big) our own comparative smallness does not necessarily undermine the idea that our conduct matters even from an impersonal perspective. For even though we may make up only a small portion of the whole, in absolute terms we are still significant enough. There may be a hundred thousand other planets with civilizations that each have their own holocaust, but the holocaust that humans made still contributed an enormous quantity of suffering to the

world, a quantity measured in millions of destroyed lives. This might be tiny portion of the total amount of suffering in the world, but in absolute terms it is unfathomably large. Ethical theories can be reconciled with this finite case by pointing out 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 doings and goings-on in the world, but rather what difference they make in absolute terms.

The infinite case is a different matter. Suppose the world contains an infinite number of people and a corresponding infinite number 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 (exceeding some low threshold). Ethical theories that hold that value is aggregative are then committed to the view that such a world contains infinite values. Now a very peculiar predicament arises. Whatever joy, virtue, and satisfaction we might add or subtract does not change the total amount there is of these goods. We can only affect a finite amount of good or bad. In ordinary cardinal arithmetic, adding or subtracting a finite quantity does not change an infinite quantity. Every possible act we could take then has the same net effect on the total amount of good and bad in the world: none whatsoever.

This problem of imperturbable infinities presents an under-appreciated challenge for many ethical theories. Consider, for example, one classical formulation of hedonistic utilitarianism which states that you ought to do that which maximizes the total amount of pleasure and minimizes the total amount of suffering in the world. If the amounts of pleasure and pain in the world are already infinite, then on this hedonistic criterion all possible actions you could take would be morally on a par, for none of them would make any difference to the total amount of pleasure or suffering. One who endorses this form of utilitarianism is committed to the view that, conditional on the world being of this infinite sort, then ending world hunger and starting another holocaust are ethically equivalent options; it is not the case that you ought to do one rather than the other. This consequence is a reductio ad absurdum if any non-contradictory normative conclusion is.

Hedonistic utilitarianism is far from the only ethic imperiled by the possibility that we are living in an infinite world. Utilitarian theories that have a broader conception of the good – such as happiness, preference-satisfaction, virtue, beauty-appreciation, or some objective list of features that make for a good life – face the same problem. So, too, does average utilitarianism and mixed total/average utilitarian views that place some value on equality in the distribution of well-being. In an infinite world, average utility (where the average may be weighted in almost any which way, for example by giving some priority to the worst off) is as imperturbable by us as is the simple sum of utility.

Many non-utilitarian ethical theories are also affected. For instance, one popular 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 this consideration must be supplemented by consideration of the special obligations we might have to particular people as well as with miscellaneous side-constraints that prohibit certain modes of conduct. If none of our acts ever makes any difference to the amount of well-being in the world, then the maximizing component of

such a theory becomes non-functional. Depending on the detailed structure of the theory, the components that remain operational may – or may *not* – continue to generate sensible ethical guidance. Further, a Moorian view according to which value resides in "organic unities" may face the same problem: if the relevant unities supervene on, say, planets, there might well be infinitely many of them, while if the relevant unity is the universe itself, then it is unclear that modifying the infinitesimal part of it that we can reach would change its value.[1]

The problem of imperturbable infinities thus threatens a large family of important ethical theories. For simplicity, we shall focus most of our discussion on purely consequentialist theories, but as we noted, many hybrid theories also inhabit the danger zone. On the other hand, not all consequentialist theories are imperiled. The vulnerability to the problem of imperturbable infinities arises from the combination of a consequentialist element with an aggregationist element. Aggregationism, as we shall use the term here, is the view 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 entire societies. The bulk of the paper will discuss and assess various countermoves whereby defenders of aggregationist consequentialism could try to avoid the reductio that it ethically doesn't matter what we do.

The problem confronting us here is related to but distinct from that of Pascal's wager, the St. Petersburg paradox, the Pasadena problem, the Heaven and Hell problem, and kindred "infinite" decision problems. They are related because in each case there are, purportedly, the prospects of infinite values to be reckoned with. They are different because one escape route that is available in the prudential problem cases is blocked in the ethical problem of imperturbable infinities. This is the route of denying that infinite values are really at stake. One way of responding to Pascal's wager, for example, 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. Moreover, in a revealed-preference paradigm, this is a perfectly natural conclusion. If we accept a theory of rationality that grounds what we have reason to do in our preferences (whether raw preferences or some kind of idealized, perfectly informed preferences) then a plausible way of answering Pascal is by saying that, 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 probably that one would go to heaven (at least if one assumes that there would be no chance of obtaining an infinite good if one did not accept Pascal's 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 eternity in heaven then Pascal's argument does not show that one is irrational to reject his wager. The fact that most people would on reflection reject his wager could be taken to show simply that most people do not in fact place an infinite value on eternity in heaven. But the analogous response does not work for the consequentialist aggregationist, for he is committed to the view that the total value of a world is the aggregate of the value of its parts, and 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 *v* such that an infinite number of the locations in the world each have an ethical value greater than *v*, then that world has an infinite ethical value. This is a core commitment of aggregationism, and giving it up means giving up aggregationism. So the ethical problem of imperturbable infinities is a more serious one for aggregationists than Pascal's wager and similar decision problems are for people who are willing to accept some kind of subjective-preference view of rationality.

Before beginning our investigation of various possible solutions of the problem of imperturbable infinities, we shall first briefly review some cosmological evidence suggesting that we probably do live in an infinite world – one that contains an infinite number of instantiations pleasure, satisfaction, virtue, beauty, and almost any quality that can supervene on human-scale physical stuff.

2. We might well be living in a canonically infinite world

One methodological question that we shall not seek to answer in this paper is what range of cases an acceptable ethical theory must avoid making wrong assertions about. By the most stringent standards, an acceptable ethical theory must not make incorrect assertions about any possible case. According to those who hold to this standard, an ethical theory can be refuted by (coherently) describing a case – be it every so improbable, farfetched, or even physically impossible – and showing that the theory implies a false statement about that case. Others might adopt a lower standard and accept an ethical theory so long as it works in cases that are at least somewhat realistic, including all the cases to which we assign a non-trivial probability. The most modest standard would require only that the theory normally work in the cases we are most likely to encounter. The challenge of imperturbable infinities threatens to show that important ethical theories fail to meet even this lowest standard of adequacy.

As a preliminary observation we may note that cosmology is a discipline that is in some flux. The last two decades have seen vigorous theory-formulation fueled by a wealth of new data whose collection has been made possible by improved instrumentation. Important revisions to our scientific understanding of the cosmos seem to occur almost every year. From a meta-level standpoint, it would seem wise to reserve at least some non-trivial degree of credence for the possibility that the coming years might see changes as drastic as the ones we have witnessed in the recent past, including changes to our beliefs about whether the cosmos is finite or infinite. Even if the cosmos is finite, we certainly do not currently have good grounds for being extremely confident that it is. On the contrary, current evidence suggests that the totality of physical existence is infinite.

On the standard Big Bang model, assuming the simplest topology (i.e. that space is singly connected), there are three fundamental possibilities: the universe can be open, flat, or closed. Current data suggests a flat or open universe, although the final verdict is still pending. If the universe is either open or flat, then it is spatially infinite at every

point in time and the Big Bang model entails that it contains an infinite number of galaxies, stars, and planets. There is a common popular misconception about this point, which confuses the universe with the (finite) "observable universe". But the observable part – i.e. the part that could causally affect us – would just be 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.[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.[3] 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. We shall call a world that contains such an infinite abundance of locally instantiated goods and bads *canonically infinite*.

Many cosmologists believe that our universe is just one in an infinite ensemble of universes, a multiverse. That there should be such an infinite multiverse is a consequence of many versions of quantum inflation theory. The possibility of a multiverse adds to the probability that the world is canonically infinite.

The upshot, then, is that it seems more likely than not that we are living in a canonically infinite world. At the very least, this must be considered a serious empirical possibility. Any ethical theory that fails to cope with this contingency should be rejected.

The "many worlds" or "branches" of the universal wavefunction referred to in the Everett version of quantum mechanics, however, do not constitute a multiverse of the kind that would necessarily imply that the totality of physical existence is canonically infinite. For a branch in the Everett ontology comes with an associated quantity, its amplitude, whose square is identified as the probability that we will observe the corresponding events located at that branch. In the most natural transposition of ordinary ethics into the Everett framework, the ethical significance of occurrences in a particular branch should likewise weighted by that branch's amplitude squared. Since the squares of the amplitudes of the branches sum to one, the weighted total amounts of good and bad occurrences would be infinite only if at least one individual branch (whose measure is greater than zero) contains such infinitudes. While it might not *follow* from traditional ethical theories that ethical significance should be weighted according to square of the amplitude (unsurprisingly, since the issue couldn't even be conceptualized before Everett!), it would be more appropriate to view this maneuver as a maximally conservative extension or reformulation rather than as a radical departure from the original intent, so at least in this respect the Everett theory would not need to be considered morally revolutionary. In this paper, we shall therefore set aside issues arising from quantum physics and focus instead on the possibility of a world that is canonically infinite either because the universe is infinitely big or because there is an infinite cosmological multiverse.

One rather desperate strategy for dealing with the problem of imperturbable infinities would be to reason that since the assumption that the world is canonically infinite has absurd implications when combined with aggregative ethics therefore the

world is not canonically infinite. It is commonly believed, however, that ethical theory should not dictate empirical science. Whenever an ethical theory finds itself on collision course with our best current scientific understanding of some matter, it is normally the ethical theory that must yield, not the science. On the most stringent methodological standard mentioned above, according to which an acceptable ethical theory must work in all possible cases, it of course true in particular that an ethical theory must be rejected if has absurd implications about cases in the world that empirical science tells us is the actual one, assuming this world is at least possible. If one instead adopts some lower methodological standard, it is not entirely clear why science should always get epistemic priority over ethics. But whatever the full story might be about that, it is highly unlikely that advocates of aggregative ethics would win much sympathy were they to base their defense on the claim that cosmologists should take cosmology lessons from ethicists.

3. Comparing infinite values: the extensionist program

Having outlined the basic problem of imperturbable infinities, we begin our examination of various possible responses. The first approach that we shall consider seeks to extend axiology by introducing rules for ranking worlds that contain infinite goods. We will term this the "extensionist program". A small literature exists on this topic, which we shall use as our starting point.

Following what has become the standard terminology, we shall call a value-bearing part of a world a *location*.[4] Experiences, acts, persons, space-time regions, and lives may be candidate locations. 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 ordinary arithmetic, the sum of value in this world is undefined.[5] 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 constant value.[6] According to aggregative ethical theories, canonically infinite worlds fall into this category, because such worlds contain an infinite number of happy people and an infinite number of unhappy people (and likewise for other kinds of local phenomena that might be considered as having positive or negative value). The sum of value in canonically infinite worlds is therefore undefined. It seems, then, that the injunction that we should maximize total value in such worlds fails to give any recommendation about what we should do, since the total value is not defined.

To see how the extensionist program tries to get out of this impasse, 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, and they each have a value of either one or two units.

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 seems 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 exactly the same locations, and w1 has strictly more value at all locations than w2. For analogous reasons, if we changed the gloss on Example 1 so that rather than the locations being days in the life of an immortal person they instead 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 contain the same people and everybody would be better off in w1 than in w2.

Peter Vallentyne and Shelly Kagan have proposed a principle that captures these intuitions:[7]

> *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.[8] Consider Example 2, where one location is a little bit 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.

In order to deal with cases like Example 2, Vallentyne and Kagan provide several strengthenings of the Basic Idea, the first one of which can be reformulated as follows. (We omit a technical complication in the original formulation that is designed to deal with certain cases that are not relevant to purposes of this paper.[9])

> *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 will rank w2 better than w3, because for any set that includes at least three locations, w1 is better than w3 relative to that set. The point of SBI1 is that it enables us to judge one world as better than another even if there is a finite number of locations at which it is worse, provided that it is sufficiently better at the other locations to compensate for this regional inferiority.

SBI1 is still quite feeble. In particular it fails to rank world pairs in which each world is better than the other in an infinite number of locations. A case illustrating this possibility is presented in Example 3 (where we have added a time index for the days in the immortal person'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, may arguably have such an order.[10] Let us suppose that the locations in Example 3, i.e. the days in the life of the immortal being, have an essential natural order. It is intuitively plausible that, if one is forced to make a choice between w4 and w5, one ought to choose w4. One reason that could be given for this is that for any time after the third day, the immortal person will have enjoyed strictly more well-being in w4 than in w5. This reason, however, fails to apply to the closely related case (Example 4), where the immortal being has always existed (so that she is everlasting in the past as well as in 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 every time, a countably infinite amount of well-being has been enjoyed in both w6 and w7 up to that time. Nevertheless, there is an intuitive ground for holding that w6 is better than w7. The temporal density of value in w4 is 2.5, while in w5 it is merely 2. For any finite (and continuous) time-period of at least four days, w6 contains strictly greater value than w7. The strengthening of SBI1 that Vallentyne and Kagan propose to deal with such cases can, in simplified form, can be rendered as follows.[11]

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

This principle judges w6 to be better than w7. SBI 2 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 come up with to date, modulo some further refinements suggested by Vallentyne and Kagan which do not affect any of the points we shall make in this paper.[12]

As a solution to the problem of imperturbable infinities, the extensionist program suffers from at least three shortcomings.

First, SBI 2 applies only when the values in question 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, and it is not at all clear that these value-bearers have an ethically relevant essential natural order. It is true that people, and the things that people do or experience, are located in time and space, and that time and space arguably have an essential natural order.[13] But the fact that people (and experiences etc.) 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 – a commitment that is not necessarily consistent with other core features of ethical theories. For example, it is fair to say that the classical utilitarianism rejects (in spirit, if not explicitly) the notion that any fundamental ethical significance is attached to facts about *where* somebody lives. One central motivating intuition in traditional utilitarian thinking is that "everybody counts for one and nobody for more than one", that features such as somebody's skin color, position in society, or place of birth are of no fundamental ethical importance; and that what matters, rather, are something like pleasure and pain, or happiness and unhappiness, or preference satisfaction or 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 originally driving 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, for instance, a world containing Hotel Hilbert (Example 5). Hotel Hilbert has infinitely many rooms, each of which has one occupant. Each occupant has a level of well-being corresponding to either zero or one unit of value. Let us suppose that, in order to be able to appeal to principles making use of the notion of an essential natural ordering, we take the locations of this world to be the rooms in Hotel Hilbert and the natural ordering to be given by the sequence of the room numbers:

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 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.[14] 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 the criterion given by SBI2 to a case like Example 5 therefore 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 have a notorious way of being counterintuitive to the human mind. Maybe we ought to regard it as a relatively minor concession for an aggregative theory to admit that the spatiotemporal pattern of distribution of goods and bads can make a moral difference in contexts where infinite values are involved. An ethical theory that admitted this could still preserve many other features associated with a welfarist outlook, such as ethical neutrality about the identity of persons. It could be maintained, for example, that no person has greater ethical status than any other person, or that personal identity lacks fundamental moral significance, and furthermore that in the finite case the value or a world is simply the sum of the values of its locations. This would be compatible with insisting that rearranging an infinite number of persons in space can alter the moral goodness of a world and thus make a moral difference.

An alternative to going down this route is to reject SBI2 and fall back on SBI1. Since SBI1 makes no use of the notion of an essential natural order, the rankings that it specifies are invariant under one-to-one permutations of values assigned to locations. But while this would ease some of the tensions with aggregative ethics resulting from the adoption of SBI2, it would also mean giving up the ability to handle a wide range of cases, including Examples 3-5. SBI1 fails to solve the problem of imperturbable infinities.

Even if we admit that spatiotemporal distribution can be morally significant and 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 of further strengthening SBI2 so that it can cover some cases in which the locations have an essential natural order that has more than one dimension, and some cases in which the two worlds to be compared do not have exactly the same locations. But even the strongest principle they formulate does not cover all such cases. Additionally, their strongest principle remains silent about cases where a single location has infinite value, and about cases of the type illustrated in Example 6, where the essential natural order has a more complex order-type.

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 $\omega + \omega^*$.[15] 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 of value) 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. So w10 is not ranked as better than w11.[16] And, of course, w11 is not ranked as better than w10 either.

The extensionist program, therefore, has not provided a general solution to the problem of imperturbable infinities 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 containing 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 that is not yet addressed by the principles we have explicitly formulated, we could simply add a clause expressing our intuitive judgment about each such new case. Even if we never manage to find an explicit principle that covers all possible cases about which we have definite intuitions, this failing could be seen an symptom of our cognitive limitations rather than as a fundamental problem with the underlying theoretical framework. Furthermore, one might think that even if the extensionist program does not succeed in addressing all possible cases, it could still amount to at least a partial solution of the problem of imperturbable infinities if it managed to cover a wide range of cases that included the cases we are most likely to actually consider in our moral reasoning.

It turns out, however, that these supportive considerations do not succeed in justifying the hope that the extensionist program could play a significant role in solving the problem of imperturbable infinities. The reasons for this pessimism become clearer once we consider the third big shortcoming of the extensionist approach.

The third shortcoming is that even if the extensionist program succeeded 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 than the other or that the two worlds are equally good. But the criterion would not tell *how much* better one world is than another.

Since we are not omniscient beings, we make our moral choices under uncertainty. When making decisions under uncertainty, 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 obtaining. More specifically, we need to 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 then to do one of the acts for which the expectation of value is maximal. (We postpone to later sections discussion of alternative decision procedures.) Now, in order to perform this operation we need a cardinal measure of the value of world. A mere ordinal ranking, telling us which worlds are better than which but not how much better, does not combine in the right way with probabilities and fails to enable us to calculate the expectation of value conditional on the various acts under consideration.

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 acts, *A* and *B*. Act *B* 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 give us 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 world will result from a given act.[17] 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 that contain a finite amount of value. Vicious gaps in our expected value calculations would then arise whenever we assigned a non-zero probability to the world being one of those for which no cardinal measure of its value was available. It would not in general be justifiable to ignore these gaps. Considerations concerning possibilities involving infinite positive or negative values can clearly be important. In order for maximizing aggregative ethics to yield acceptable moral guidance, more is therefore needed than the ordinal ranking of pairs of possible 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 a sufficiently rich set of ordinal preferences.[18] This result is a corner stone 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 on possible worlds (such as the one depicted in Example 7). But to assume such an ordinal ranking is to already help oneself to precisely what is problematic. By contrast to the case of individual prudential decision-making, we cannot appeal to a simple revealed preference account, for not everybody might be disposed to make the same evaluations of the ethical merits of various possible gambles. Such a ranking must come from an axiological theory. If that axiology is aggregative, we encounter the problem of imperturbable infinities.

In later sections we shall revisit to some of the issues broached here. Meanwhile, we can draw several preliminary conclusions: (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 bad; (3) for many classical aggregative theories, it is doubtful that they are compatible with the stipulation that the spatiotemporal distribution 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 imperturbable infinities.

4. Extending decision theory

Another idea for how to solve the problem of imperturbable infinities 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. Maybe if it doesn't matter what we do in the infinite case, we ought to focus on this finite case. Could the mere subjective possibility of the world being finite be a reed of sufficient strength to form the substance of a successful rescue operation for aggregative ethics?

To attempt this method of rescue, we must give up the claim that a right act is one that maximizes the expected value of the world. The expected value of a canonically infinite world is undefined in the standard model of arithmetic. In a canonically infinite world, there is an infinite amount of good and an infinite amount bad. Even if the probability that we are in a canonically infinite world is small, the expected good is infinite, since an infinite quantity multiplied by some small positive real number is infinite; and the expected bad is likewise infinite. The expected value equals the expected good minus the expected bad, but subtraction between two cardinal numbers of the same infinite size is not a well-defined arithmetical operation. The expected value of the world (conditional on our performing any feasible act) is consequently undefined, given that there is a finite probability that the world is canonically infinite. No act could be said to maximize the expected value of the world; hence, no act would be morally right. Moreover, if an act is wrong whenever there is another feasible act giving a higher conditional expected value of the world, then no act would be wrong either.

Let us therefore consider how we might proceed if we step back from the claim that a right act is one that maximizes expected value. We then need some other way of deciding what we ought to do. Here we shall explore the "finite probability of a finite world" approach, which uses the standard model of arithmetic but seeks to get traction on the problem of imperturbable infinities by considering the possibility – even if it is *only* a possibility – that the world is finite. We shall argue that this approach encounters difficulties that make it unpalatable, at least as a stand-alone solution.

Suppose there are two feasible acts: A^+, which is intuitively good (e.g. feeding the starving), and A^-, which is intuitively bad (e.g. committing genocide). Next, consider two possibilities separately: 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 by the above reasoning it does not matter ethically what we do. Neither of A^+ and A^- is morally preferable to the other, conditional on S_{inf}. But if S_{fin} obtains then our conduct does make an ethically significant difference. Conditional on S_{fin}, A^+ is strictly better than A^-. Therefore, A^+ dominates A^-: in no possible state is A^+ worse than A^- and in some possible state is A^+ strictly better than A^-. One could argue that this constitutes a moral reason for doing A^+ rather than A^-.

The dominance principle underpinning this reasoning boils down to the 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 this idea to the following decision rule.

> *Extended Decision Rule* (EDR)
> Let $P(\infty^+ \mid A)$ be a subjective probability assigned by the decision-maker 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 decision-maker at the time of the decision; we assume that this set is finite.[19]
>
> 1. For each action A_i in Ω, 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 that is the right act to perform and other acts are wrong.
> 2. If there is more than one act in Ω^* then consider, for each such act A_i, 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."[20] EDR generalizes this by stating one should rather maximize the difference between the probability of securing an infinite reward and the probability of incurring an infinite punishment. Second, EDR allows finite values to serve as tiebreakers when considerations about infinite values cancel each other out. EDR can be further generalized to cover cases where values of different infinite orders are at stake. This could be done by adding steps to the decision procedure such that values of greater infinite cardinality are always given lexicographic priority over lower-level values, which are used as tiebreakers in case of a draw at the higher levels.[21]

Does EDR solve the problem of imperturbable infinities? The hope would be that, with EDR, aggregative ethics can avoid collapse 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 as far as infinite values are concerned, EDR tells us direct 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 extension rather than a radical revision of traditional positions.

One feature of EDR that is attractive at first sight but on closer inspection turns out to be quite disturbing is the strict priority EDR gives to maximizing the probability that the world will contain some infinite good and to 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 huge. If there is an act such that one believed, 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 "side-effect" of laying to waste a million human species in a galactic nuclear holocaust. This act would be would be classified as morally right, despite it being practically certain that the horrendous sacrifice it would entail would achieve no good.

If EDR were accepted, speculations about infinite scenarios, however unlikely and farfetched, would come to dominate our ethical deliberations. We might become obsessively concerned with bizarre possibilities in which, e.g., some kind of deity exists which will direct its infinite powers to good or bad ends depending on what we do. No matter how fantastical any given such scenario would be, if it is a logically coherent possibility it should presumably be assigned a non-zero, finite probability,[22] and this would be enough to enable it to trump any consideration regarding merely finite but far more realistic values.

This implication is hard to accept, but some aggregationists might be willing to bite the bullet. They might argue that it is only our human fuzzy-mindedness – our inability to intuit just how good an infinite good would be – that makes the consequence seem wrong. (This is, presumably, what followers of Pascal would argue in response to an analogous objection to Pascal's Wager, although they might also maintain that the God-hypothesis is not so fantastically improbable.)

We must admit that our unschooled intuitions about infinitudes are often unreliable. Yet even after critically reflecting upon this implication, the impression persists that what we are being asked to accept borders on the insane. Aggregationists could attempt to blunt this impression by maintaining that the practical upshot of accepting their position is not as radical as we might fear. One way of doing this would be by arguing that considerations about the infinite possibilities cancel each other out. For each bizarre scenario in which our doing something which is intuitively wrong would lead to the realization of an infinite good, we can imagine an equally probable and opposite scenario in which our doing the same thing would lead to the realization of an infinite bad. Only if we have some discriminating information about infinite possibilities would taking them into account impact our deliberations. But (the argument would go) in fact we lack such information; hence we need to focus only on the finite possibilities,

about which we do have relevant information, and use these to determine what we ought to do. On the basis of these finite possibilities we can then safely conclude that gratuitous genocide is wrong and that feeding the starving is a much more promising candidate for being morally right.

Yet this dismissal of the worry is too swift. It relies on the claim that the various infinite scenarios one could concoct cancel each other out *exactly*. The probability increase of an infinite good that an act causes in one scenario must be precisely matched by a probability decrease of an infinite good in another scenario, or else offset by an increase in the probability of an infinite bad. But is that really how things would play out? The cancellation would have to be perfectly accurate, down to the seventeenth decimal place and beyond. While it might seem sensible to say that we have no discriminating information as to which infinite scenario might materialize conditional on us undertaking a given act, this is true only if we speak roughly. The subjective 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 farfetched) 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 be 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.

The practical upshot of EDR, far from conveniently reducing to our naïve mode of reasoning, in which we mostly ignore infinite possibilities, might instead be revolutionary. An aggregationist taking EDR seriously may well conclude that she should ignore finite values altogether and focus all her attention on figuring out which way the balance of infinite reasons tilts. She may turn with fervor to theology, or to theoretical physics or metaphysics, or to speculations about the future of technology – these being among the disciplines that have the strongest bearing on the less improbable ways in which infinite values could be subjectively correlated with our actions. The shocking implication is not that one or more of these fields deserve serious attention (which is true for other reasons) but that they deserve to monopolize our attention. The view that any consideration relating only to finite values should be completely ignored in our ethical thinking could be fairly characterized as fanatical.[23]

The defender of aggregative ethics could retort that the practical implications of accepting the injunction can be tempered provided one holds some suitable stabilizing view about how reasoning about the infinite values plays out. Suppose, for example, that one is convinced that the by far most likely scenario in which infinite values are probabilistically correlated with our actions is one involving something like a Christian God but who operates according to a principle of collective responsibility. In this scenario, homo sapiens is the only intelligent species and the universe is finite; and if we collectively exceed a certain threshold of (traditionally understood) moral merit, God will reward us all with infinitely long lives in Heaven whereas if we fail to do this then He will soon bring about an apocalypse and end the world. Given this view, EDR entails that we ought to act in accordance with traditionally morality and try to encourage others to

do likewise. Since by assumption this scenario is far more likely than other scenarios involving infinite values, there is no disconcerting implication that we ought to expend an inordinate amount of effort researching infinite values. The wild fanaticism we feared would follow form the adoption of EDR is here tamed by the assumption that the acts most likely to promote infinite goods are identical (or very similar) to the acts recommended by common sense morality.

It is not only theological hypotheses that could serve such a stabilizing role for aggregationists. Another example is the hypothesis that the most likely way by far for our acts to correlate with infinite outcomes is in a scenario like the following: Our descendants one day discover some new physics that enables them to develop technologies that make it possible to create an infinite number of people.[24] If our current behavior has some probabilistic effect, however tiny, on how our descendants will act, we would then according to EDR have a reason to act in such ways as would maximize the chances that we will have descendants who will develop such infinite powers and use them for good ends. It is not obvious which of our feasible acts have this property, but it seems plausible that this line of reasoning would lend support to many of our common sense moral convictions. For instance, it would seem more likely that ending world hunger would increase, and that committing 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, than that the opposite should be true. More generally, working towards a morally decent society might generally be a good way to advance the prospect of the eventual technological realization of infinite goods on the odd chance that such should be possible. The relevant magnitude here is not the absolute probability of success in such an endeavor but rather the relative probability of success compared to that associated with other projects we could undertake that might conceivably promote the creation of infinite goods and forestall the emergence of infinite bads. This defense, therefore, does not rely on the assumption that it is probable that our descendants could develop infinitely powerful technologies, or that it is probable that we could determine whether they will do so or that we could influence their uses of them.

All things considered, then, it seems that aggregationists accepting EDR could avoid unacceptable implications about what ought to be done, provided that they also hold some suitable stabilizing view according to which the acts most likely to maximize the difference $P(\infty^+ \mid A_i) - P(\infty^- \mid A_i)$ are not too radically different from the acts we normally think of as morally right. Moreover, it seems plausible that some such stabilizing view is correct. On balance, we would presumably judge that it is more likely that the difference is maximized by acts that, say, promote democratic institutions, save lives, and improve the human condition, than by acts of random murder or wanton destruction.

On one point there likely would be a divergence from traditional moral priorities. Given EDR, we should assign greater importance than we would otherwise do to such research as has even a remote prospect of informing us about the ways in which our acts might correlate with infinite values. This is because so little is currently known about which acts are best from the infinite perspective, and we can hope that further attention to

this issue from such fields as cosmology, theology, philosophy, and technological futurism will throw up some useful ideas. Even the slightest illumination of this topic would (given EDR) be of enormous value, since it would help us refine our understanding of what our practical aims should be. Yet this divergence seems acceptable. The aggregationist would not be committed to the view that we should spend all or most of our resources chasing after enlightenment about infinite prospect. She could hold that we already have a sufficiently good idea about what the aims should be that it is more cost-effective to work directly towards realization of those aims than to invest enormous effort into confirming or to fine-tuning them.

Not everybody, however, will be put to ease about the implications of EDR by these subtle considerations. Even if the above reasoning is correct, the feeling might remain that this kind of account of why it would be wrong to commit genocide, is too brittle. It might also appear that the whole construction, relying crucially as it does on the balancing of tiny subjective probabilities of infinite values, is itself monstrously inhuman and rationalistic.

Can this disquiet be more precisely articulated? Let us consider a few ways in which one could try to do that.

(a) One thought starts from the observation that there may be hardly a single individual in the history of the human species who based her moral practice on considerations like those above. Many may have done the right thing, but they all did so without having any inkling about the true ground for it being the right thing. Is it a defect in an ethical theory that it implies a vast gap between what people believed were the reasons and what were the real grounds for the rightness and wrongness of what they did? But the bar in this regard must not be set too high. 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 meta-ethical theory. There may not be anything especially objectionable about the aggregationist's gap between the true and the commonly supposed grounds for what we ought to do, so long as the actual prescriptions the aggregationist makes do not diverge too radically from what our pre-theoretic intuitions recommend. Moreover, at least the *proximate* grounds for the morality of particular acts would generally have been recognized by most people. For example, genocide is wrong because it unjustly inflicts great harms on large numbers of people, it leads to strife and war and large-scale destruction of resources, it forfeits opportunities for cooperation, trade, and friendship, and so forth. This is well known. It is only the ultimate ground that is esoteric, i.e. the fact that these consequences of genocide would decrease the difference $P(\infty^+ \mid A_i) - P(\infty^- \mid A_i)$.

(b) As for the perception that aggregationism is cold or impersonal, this is a common ground for objecting to it in the finite case, too. EDR might intensify this perception slightly but does not seem to make a drastic difference. Psychologically, it seems not much harder to internalize the idea that we should aim to improve the expected

well-being of an infinite population than that we should do this for a six-billion human collective.

(c) One may consider the hypothetical case in which we know (with absolute certainty) both that the world is canonically infinite and that out feasible acts are correlated only with finite values. Since conditional on these assumptions, all feasible acts are morally equivalent according to EDR, we would in this hypothetical case have to conclude that there would be nothing wrong in committing atrocities. This would be an unacceptable consequence. The aggregationist, however, could retort that it could never be rational for us to assign exactly zero probability to the assumptions not holding, so the supposed situation is impossible. Alternatively, the aggregationist could argue that even if there is some possible world in which we could rationally assign zero (or infinitesimal) probability to the relevant propositions, this counterfactual possibility is so remote that it is not damning if our ethics say counterintuitive things about it. This latter response depends on accepting something less than the most stringent methodological success criterion for ethical theories (see section 2).

(d) Suppose we get increasingly strong evidence that the world is canonically infinite and that the already week probabilistic link between our acts and the world's expected amount of goodness and badness becomes even more attenuated. It might seem strange that whether any act we could do would be morally wrong would depend on there still being a tiny subjective probability that the universe is finite after all or that our acts correlate, however weakly, with infinite values. Could we really accept that the wrongness of genocide and other atrocities dangles on such a thin thread? If we cannot accept that, then EDR does not solve the problem of imperturbable infinities.

In conclusion, we can say that EDR offers protection against the problem of imperturbable infinities provided the aggregationist is willing to go out on a limb by invoking non-trivial empirical assumptions about what rational people must believe about the likelihood of various infinite scenarios. The feeling that we are skating on thin ice persists.

Let us move on to consider other possible responses to the challenge of imperturbable infinities. In the concluding section we shall revisit EDR and consider whether it might work better as an adjunct to some other remedy.

5. Discounting to the rescue?

If we discount values on the basis of how spatiotemporally distant they are from the decision-making person at the time of the decision, we might be able to avoid having infinite aggregates of value enter into our calculation.

Temporal discounting is standardly used in economic analysis. Many different discount formulas have been proposed, the most basic and frequently used one being the exponential form:

$$U(C_t) = \left(\frac{1}{1+r}\right)^t u_t(C_t),$$

where $U(C_t)$ is the current utility-equivalent of the utility $u_t(C_t)$ derived at a time t years from now from consumption of C_t at that time, and r is the discount rate (often assigned a magnitude of about 5%). In order for discounting to have even a starting chance of solving the problem of imperturbable infinities, it must involve spatial as well as temporal discounting. One straightforward way of generalizing this to spatiotemporal discounting is would be by setting:

$$U(C_{t\bar{d}}) = \left(\frac{1}{1+r}\right)^t \left(\frac{1}{1+s}\right)^{|\bar{d}|} U_{t,\bar{d}}(C_{t\bar{d}}),$$

where $|\bar{d}|$ is the distance from the decision-maker (at the time of the decision) to a location $\bar{d}$ where the consumption of $C_{t\bar{d}}$ takes place, and s is a spatial discount rate. If the constant $|\bar{d}|$ is units of light years and s is on the order of the temporal discount rate, then the spatial discount rate would be negligible for events taking place on Earth or in the solar system.

It is however commonly thought that even the temporal discount factor does not represent a fundamental ethical truth to the effect that we should judge that events that are further into the future are *ipso facto* of less moral significance. Most moral philosophers probably hold that the a temporal discount factor serves merely as a proxy for a host of correlated factors (such as economic growth, unpredictability, risk of dying, weaker personal ties to people who will live in the far future, etc.) that explain why it generally makes sense to pay less for a commodity that will be consumed later. An ethically fundamental *spatial* discount factor is even less popular. 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.[25]

Discounting would entail a radical departure from some of the core ideas of aggregative ethics. It would, however, be able to preserve at least one element of aggregative ethics, namely agent-neutrality (also known as "anonymity"). This is the idea that the identity of locations is of no fundamental ethical significance. One sense of agent-neutrality could also be maintained in the spatiotemporal discounting approach, provided that we specify that an act that is morally right for an agent is one that

maximizes the expectation of discounted value, where the spatiotemporal distance is calculated using the agent in question (at the time of the decision) as the origin. This would make expected value agent-relative (and the expected ethical value of an act would not need to be the same even for agents who have the same probability function), and there could therefore be cases in which different agents would deliver different moral judgments on the same act. But there would still be agent-neutrality in the sense that the form of the moral reasoning would be the same for everybody. To relativize discounted value to agents-at-a-time seems less unattractive than to arbitrarily designate one spacetime point as the origin to be used by everybody in calculating an act's expected discounted value.

Even if we accepted exponential spatiotemporal discounting, this would still not solve the problem of imperturbable infinities if it is possible for one location to have arbitrarily large finite value. For we could then describe a case in which the values of locations further removed from us increase at a faster rate than the rate at which they are discounted, so that their sum diverges. This holds in general for any discounting scheme that assigns a finite ethical weight to (i.e. does not infinitely discount) every location that is finitely removed from us. Postulating an arbitrary threshold at a finite distance from us such that locations outside that distance were assigned no ethical significance does not seem like an attractive option. (And even if we supposed that there were such an infinite ethics-free zone surrounding a finite bubble of ethical concern containing ourselves, this would still not expunge the possibility of imperturbable infinities if it is possible for a location to have infinite value.[26])

In the case of spatiotemporal discounting we could not use the defense suggested for the aggregationist approach discussed in section 3, namely that infinitudes are mysterious and that we should perhaps not be too perturbed by the discovery that our ethical theories may need to be extended or revised in order to be able to handle infinite worlds; for spatiotemporal discounting would apply also to finite worlds. One might respond to this by stipulating that discounting should only take place in worlds that contain an infinite amount of (non-discounted) value. But this would have the perverse consequence that some worlds with a finite amount of good (and no bad) that would be ranked as better than some worlds with infinite amounts of good (and no bad). To see this, consider any world with an infinite amount of non-discounted good and let x be the discounted value of that world. By assumption, x is finite. Then simply pick a world containing only a finite amount of good that is greater than x. Since the value in this world would not be discounted, it would count as better than the world with the infinite amount of good.

Another way of discounting would be to discount not on the basis of spatiotemporal distance but instead to discount on the basis of how much value a world contains at other locations. The amount of contribution to a world's overall value that the content of a location would make would on this view depend on the content of other locations. Value would thus 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, this would be the case if we adopted the following form:

$$V = e^{-U^-} - e^{-U^+},$$

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 adding them together. Apart from the fact that this diverges significantly from aggregative ethics as originally understood, one might think that, since it guarantees that the there is a finite upper bound on the value of any possible world, it would solve the problem of imperturbable infinities. This is not so. Although the infinites are suppressed in the final assessment of the value of a world, they are still present in what we have here termed the utility of locations. The effect of this is to make the value of a canonically infinite world imperturbable despite being finite. If there is already an infinite amount of positive and negative utility in the world, then we cannot change it (any more than we can change an infinite amount of value) and so cannot change the finite value assigned to the world. If the moral rule is that a right act is one that maximizes the expected value of the world, then this amounts to complete moral indifference between all feasible acts, the very conclusion we hoped to avoid.[27]

The prospects for aggregative ethics of finding shelter from the problem of imperturbable infinities from the use of discounting are bleak.

6. Causal decision theory, or evaluating changes instead of worlds

If aggregative ethics runs into trouble in canonically infinite worlds because the finiteness of our acts is washed away in the ocean of infinite goods and bads that exist independently of what we do, then maybe the remedy is to consider the effects of our actions in isolation from the rest of the world. Rather than evaluating whole worlds, we can evaluate the changes that we can bring about. We could define a right act as one that maximizes the expected value of these changes. This approach is related to, but (as we shall see) distinct from, causal decision theory.

Causal decision theory was developed to correct some perceived defects in traditional evidential decision theory. According to evidential decision theory, one ought to do one of those feasible acts that has the greatest expected value, i.e. one of those feasible acts, conditional on whose performance, the expected value of the world is at least as great as for any other feasible act. Formally, a right act A is one that maximizes:

$$U(A) = \sum_{w \in W} U(w)P(w \mid A) \qquad \text{(Evidential Decision Theory)}$$

Here, the sum ranges over all possible worlds w which have greater than zero conditional probability on some action A; $U(w)$ is the utility of world w, where this is taken to include the utility of the act itself if is performed in w; and where $P(w \mid A)$ is the agent's subjective conditional probability of w given A. (In the case of ethical rather than prudential decision-making, $U(w)$ would be replaced by $V(w)$, the ethical value of w.) Evidential decision theory claims that one should maximize the expected "news-value" of one's decision, i.e. that one should act in such a way that news that one has selected to act thus is at least as good as would have been the news that one had selected one of the alternative acts.

Causal decision theory, by contrast, claims that one should do one of the acts that have the greatest expected causal efficacy in bringing about good results. It is often claimed that causal decision theory is equivalent to evidential decision theory except in cases where one believes that which action one chooses to perform alters the expected value of the world in non-causal ways. The paradigm example of such a case is the Newcomb problem (described in this footnote[28]). In the Newcomb problem, evidential decision theory recommends taking only the opaque box, since one's expected utility given that one takes only that box is nearly $1,000,000 whereas the expected utility of taking both boxes is merely $1,000. Causal decision theory recommends taking both boxes, since one knows that one's present choice has no causal influence on the content of the boxes (which was fixed in the past), so whatever their content may be one gets either as much or more by taking both boxes as by taking only one (depending on whether the predictor has made a mistake). Followers of evidential decision theory would get richer than followers of causal decision theory if the world contained a lot of Newcomb-like problems, but causal decision theorists retort that this is because such a world would reward irrationality.

Evidential and causal decision theory come apart in such cases because there is a correlation between one's beliefs about what the world is like and one's choice of act that is not mediated by the causal effects of the act. In the Newcomb problem, this non-causal link is the belief that the predictor's move, and thus the utility-consequences of one's available choices, is probabilistically correlated with one's choice even though that choice is known to have no causal effect on the predictor. By contrast, in "normal" cases, where the only relevant information given by the proposition that one will perform a particular act is that the causal consequences that that act would have will now take place, it is commonly thought that the evidential and the causal approach coincide.

It might consequently be thought that causal decision theory could help us deal with the case of a canonically infinite world, at least if we assume that there is some known finite upper bound on the size of the causal consequences that our acts might have. The idea would be that causal decision theory would tell us focus on our causal consequences, which in this case are guaranteed to be finite. Unfortunately, causal decision theory is generally formulated in such a way as to eliminate its capacity to deal with such cases. For example, David Lewis argues for the following general form of causal decision theories:

$$U(A) = \sum_K P(K)V(A \& K) \qquad \text{(Causal Decision Theory)}$$

The sum ranges over a partition of the space of possible states into "dependency hypotheses" (which are relative to an agent at a time). A dependency hypothesis K is a "maximally specific proposition about how the things he cares about do and do not depend causally on his present actions".[29] $V(A\&K)$ is the value of the proposition that act A is performed and that K holds. A and K together should specify exactly how the world is with regard to all the things one cares about, thus $V(A\&K)$ should have a definite value as opposed to merely an expectation value. $P(K)$ is simply one's current subjective probability that K holds.

In the ethical application, the dependency hypotheses would instead specify how the things that are of ethical value depend causally on the agent's actions, and $V(A \& K)$ would be the ethical value of the truth of $A \& K$. This formulation fails to help the aggregative ethicist to deal with the canonically infinite case, since the term $V(A \& K)$ is then undefined. (If the aggregative value of $A \& K$ is decomposed into its positive and negative elements, both parts will be infinite and the subtraction of the negative part from the positive part will be an undefined arithmetic operation.)

So a causal decision theory does not help with the problem of imperturbable infinities if the evaluation of outcomes is calculated on the basis of the value of whole worlds rather than on the basis of the value of the outcomes themselves. The obvious way to remove this stumbling block is to postulate that the evaluation should instead be restricted to the causal effects that an act may produce.

Consider the following situation, representing a line, infinite in both directions, of alternatingly happy and unhappy people. You have the choice between making the person p_3 one unit more happy than he would otherwise be (w14) or leaving everything as it is (w15) (Example 8).

w14:	…, 1, -1, 1, -1, 2, -1, 1, -1, 1, …
w15:	…, 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 8

Since in this case you can causally affect the value of only one location (p_3), a rule that would take only the value at this location into account could easily deliver the verdict that you ought to choose (w14).

The proposal, then, is to modify aggregate ethics by limiting consideration to a subset of all locations, those within your sphere of causal influence. A morally right act could then be defined as one that maximizes the expected aggregate of value of the locations within this sphere (possibly after those values have been weighted in some way according to the degree of your ability to influence them).

How should we flesh out this notion of "sphere of influence"? Let us consider some of the options.

(a) We could arbitrarily pick some very large spacetime region, say a million light years wide and a million years deep, and stipulate that only locations that are within this region should be considered in our ethical deliberations. – This has the theoretical defect of being plainly arbitrary. Furthermore, there is no reason in principle why we could not find ourselves in a situation where our acts would have consequences for what happens at a distance greater than one million light years from our current location. For example, the founding of a space-colonizing civilization that might eventually spread throughout a large part of the observable universe would be one kind of action that could well have such consequences. It would be wrong to suppose that those consequences could not possibly have any moral significance. We should reject (a).

(b) Somewhat less arbitrarily, we could define our sphere of influence to coincide with our future light cone. According to our best current physical theories, nobody can influence events outside their future light cone. However, this option would at best satisfy only a relatively modest methodological standard (section 2), since it assumes that no causal influence could propagate faster than light. It is undesirable for our fundamental moral theory to depend on the specifics of current physical theories. If, contrary to what our physical theories lead 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. Moreover, we can reasonably assign a non-zero probability to relativity theory being false and to superluminal influencing being physically possible. Yet (b), we should not factor in such possibilities in our deliberations even if we thought superluminal influencing to be quite likely; this, too, seems wrong. Finally, 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 spacetime, or in a steady-state cosmology.[30] The problem of imperturbable infinities would thus remain even if (b) were accepted.

(c) To rely less on our current understanding of physics, we could simply define our sphere of influence to contain all and only those locations that we can causally affect. However, the possibility that we might causally influence an infinite number of locations is a problem for (c) too. In some ways this problem becomes even more acute, since (c) is committed to including in the evaluation all locations that we might causally influence, not only those within our future light cone. The larger we define our sphere of influence to be, the greater the risk that it might contain imperturbable infinities. Even if in fact our sphere of influence contains only finite amounts of good and bad, reasonable agents might nevertheless assign a finite, non-zero probability to it containing infinite amounts.

(d) We could define a weighting over all locations whose strength would correspond to the degree of causal control with have over the value at that location. Our sphere of influence would consist of all locations for which this weighting is non-zero. In many of the scenarios in which we could have some causal influence over a canonically infinite region, our influence on all but a finite part (usually quite a *small* finite part) would be tiny, and on this count option (d) is superior to (c). For example, if in the cyclic time scenario alluded to above, we would typically have very little systematic influence

over remote regions in spacetime, and the same holds for many other of the physically least unrealistic scenarios in which we could affect an infinite number of locations. One consideration supporting this general presumption is that – assuming we are not special – there would in these cases typically be an infinite number of people like us who would have similar powers over the same range of locations. One would then expect each person to have only an infinitesimal degree of control over the overall outcome, just as most individuals in a large democratic country have only a very small degree of control over national policies. Therefore, if one postulates that agents should weigh values with a parameter reflecting the degree to which the agent can causally influence them, one will find in many cases where we can affect an infinite number of locations of value that this weighted measure of value will nevertheless be finite.

Each of the above alternatives comes in two versions. We could either take the definition rigidly so that for everybody and for always, the sphere of influence is one and the same; or else we could relativize the definition so that each agent at each time has her own sphere of influence, defined in terms of her location and her causal powers. The rigid version is unattractive. In a canonically infinite world, most people live outside any finite sphere of influence. If we pick a sphere that contains us, not only would this stipulation smack of partisanship but it would also imply that most people's acts are all be morally indifferent. An exact replica of Hitler, doing the sorts of things that Hitler did, would be acting wrongly even if the planet where he lives happens to be located outside our sphere of influence. It therefore seems best to relativize the definition of the sphere of influence. We would say that a morally right act, for an agent at a time, is one that is maximally good relative to that agent's sphere of influence at that time.

How does the method suggested in (d) relate to causal decision theory? It is clearly not equivalent to the general form given by Lewis's equation (reproduced above), since at least one of the terms referred to in that formula is undefined in canonically infinite worlds whereas option (d) yields a definite answer in many canonically infinite worlds. Notwithstanding, (d) is definitely compatible with the spirit of causal decision theory. It could also be made compatible with evidential decision theory, although there is arguably less of a natural fit in this case. We may schematically express option (d) as follows:

$$U(A) = \sum_l \left(\sum_v \overline{C}(A \mapsto V_l = v) \cdot v \right) \qquad \text{(D)}$$

where the first sum ranges over all possible locations l that whose value might be affected by act A, the second sum ranges over all values v that that location might have, and $\overline{C}(A \mapsto V_l = v)$ is a measure of the agents expectation of the "degree" to which A would be causally efficacious in giving value v to location V_l. (For possible locations l which are not actual, we would set $V_l = 0$.) Depending on how $\overline{C}(A \mapsto V_l = v)$ is interpreted, (d) could also be made compatible with evidential decision theory. Consider again the Newcomb problem. The relevant location here can be taken to be the decision-making

agent, and the value of this location can be taken to be the payoff received in the Newcomb problem. Whatever money the agent gets, she will herself have decisively causally contributed to her getting it since she only gets the cash that is in the box(es) she takes. Whether she should take both or only one box then depends whether *U*(*Both*) is larger than *U*(*Opaque*), where

$$U(Both) = \overline{C}(Both \mapsto V = \$1{,}001{,}000) \cdot \$1{,}001{,}000 + \overline{C}(Both \mapsto V = \$1{,}000) \cdot \$1{,}000$$

$$U(Opaque) = \overline{C}(Opaque \mapsto \$1{,}000{,}000) \cdot \$1{,}000{,}000$$

The outcome of this comparison depends on the value assigned to $\overline{C}(A \mapsto V = v)$, which in turn depends on whether this factor is construed in accordance with the spirit of causal or evidential decision theory.

Would adopting an approach along these lines of (d) entail giving up an essential part of traditional aggregative ethics? Can an aggregationist embrace the idea that we ought to "do good" even if we know that doing so would not make the world better? Some of the classical formulations of aggregative ethics do say that we ought try to do as much good as possible rather than that we ought try to make the world as good as possible. Mill, for example, defined utilitarianism as:

> The creed which accepts as the foundation of morals, Utility, or the Greatest Happiness Principle, holds that actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness.[31]

In this passage, it is natural to give "promoting" and "producing" a causal reading. Our objective here not being exegetical, we will simply note that approach (d) is incompatible with some forms of aggregationism and compatible with at least some theories that have traditionally been viewed as aggregationist.

In the finite case, which is probably what Mill had in mind when he wrote the above sentence, what we should do if we ought to do as much good as possible amounts to the same thing as what we should do if we ought to make the world as good as possible. That there should be this equivalence in the finite case could be imposed as a constraint on how $\overline{C}(A \mapsto V = v)$ may be defined. It is in the infinite case that the two approaches (maximizing the expected value of the world vs. the expected value of the chances we make in the world) can come apart, because infinite worlds may contain both infinite positive and infinite negative value such that the value of the world as a whole is undefined or imperturbable while the portion of the world that falls within our sphere of influence might still have a bounded finite value and thus be (both evidentially and causally) sensitive to how we choose to act.

Now if we do accept that the approach represented by (d) is compatible with aggregationism, does it then provide a solution to the problem of imperturbable infinities? Unfortunately, the answer is negative. First, it fails if our acts might have

causal consequences that extend over an infinite number of locations (unless the degree of our influence over more remote locations converges to zero sufficiently rapidly that the overall weighted influence of our acts is finite). Second, it fails if a single location might have an infinite value over which we have some non-infinitesimal degree of causal influence.[32] We must bear in mind that all it takes for either of these contingencies to threaten ruin to aggregative ethics is that the decision-maker reasonably assigns a finite, non-zero probability to their occurrence. The conclusion, therefore, is that (d) does not by itself solve the problem of imperturbable infinities.

7. Ignoring small probabilities or remote possibilities?

How much should we worry about farfetched scenarios in which our sphere of influence is infinite? Could we solve the problem of imperturbable infinities by combining the approach explored in section 7 with the postulate that, when determining what is right to do, we should simply ignore very unlikely possibilities?[33]

Whatever else might be said of this suggestion, it certainly lacks the virtue of theoretical beauty. As a piece of pragmatic advice, the idea that we should ignore small probabilities is no doubt 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 highly unlikely outcomes must be taken into account when the magnitude of the value we assign to them is sufficiently large. In deciding whether to invest in a safety feature for a nuclear power plant, we need to consider a range of very improbable scenarios. 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. When the value in question is infinite, as in many of the cases considered in this paper, their low (finite) probability is not in general a justification for setting them aside.

The different possible methodological standards discussed in section 2 do not map directly onto the point at issue here. Even if we think that it is good enough for an ethical theory to hold in the range of cases that we consider at least fairly realistic or plausible, this not does entitle us to ignore small probabilities in general. For the challenge here concerns the actual case. It is certainly not good enough for an ethical theory to fail in what we take to be the actual case. In the actual case, it seems that we have reason to assign a finite probability to many scenarios involving infinite values, and arguably also to scenarios in which we might be able to exert a finite degree of causal control over infinite values. By to any plausible methodological standard, ethical theories must deal with this actual case, in which reasonable people assign finite probabilities to infinite scenarios.

If someone is nevertheless tempted by the idea that we should ignore small probabilities in the context of ethical decision-making, the question then arises of just how small a probability would have to be in order for it to be negligible. If the threshold

is set too high, it will have the crazy implication that low-probability events such as nuclear disasters should be ignored in ethical reasoning. If the threshold probability is low enough to include events of this type in moral decision-making, then it will also be low enough to include scenarios involving infinite values. Reasonable agents who are well-informed about current cosmology assign a greater probability to the world being canonically infinite than to meltdown happening in any particular nuclear power plant. There is nowhere to place the probability threshold such that it would prevent all infinite scenarios without also blocking common sense concerns.

At best, the low-probability cut-off could serve as an adjunct to some other treatment for the imperturbable infinities problem. For instance, if we adopt the rule (D) discussed in section 7, the probability of a scenario in which infinite values enter into our calculations is diminished, although it is dubious whether the probability would shrink sufficiently that it can be extirpated without harm to the healthy tissue of ordinary applied risk analysis. (The theoretical implausibility and ugliness of this method would also remain.)

There is another way in which we could remove infinite values from consideration. Rather than seeking to avoid problematic infinites by introducing a low-probability cut-off, we could cut to the chase and simply stipulate that all possibilities involving infinite values should be ignored, independently of their probability. More specifically, we could formulate an ethical theory that postulates that a right act for an agent is one which the agent reasonably believes maximizes the expected value of the world, but where this expectation is calculated by omitting all possible worlds that contain infinite value.

Such a theory would be theoretically ugly. Moreover, it would constitute a departure from maximizing aggregative ethics since it implies that we ought to do something other than maximize aggregative value. It would also fail to meet the most rigorous methodological standard, because there are possible situations about which it would deliver clearly incorrect recommendations. For example, in situations where we are confident that infinite values depend directly on our actions, it would definitely be wrong to completely ignore those infinite values in deciding what to do.

Nevertheless, we may ask how well the theory could fit some lower methodological standard. Would this theory recommend approximately the same acts that we intuitively consider morally right, and do so in a sufficiently wide range of the cases we are likely actually to confront to be satisfactory at some less-than-maximal level of methodological stringency?

A proponent of the theory could argue that since our pre-theoretic moral intuitions were formed in contexts that did not include confrontation with infinite values, whether in thought or in practice, it would be unsurprising if these intuitions could be captured fairly well by a theory that likewise brackets off infinite values from consideration, even if the theory does so only by adding an ungraceful "epicycle". So long as we restrict ourselves to our pre-theoretic intuitions about what we should do – as opposed to intuitions about the structure of the theory – it seems plausible that such a theory may succeed in capturing these intuitions relatively well. (We are setting aside here any gaps

between theory and moral intuition that do not stem specifically from the possibility of infinite values but concern other aspects of aggregative consequentialism.) Divergences between such a theory and intuition would arise in cases where infinite values are sufficiently directly linked to our acts that what we ought to do considering these infinite values is something different from what we ought to do if we pretend that these infinite values could not exist. How likely it is that we will encounter such cases is an open question, but we should not be too confident they will not arise. Consider the following thought experiment:

> *The Funding Body*
> You are in charge of allocating research money to fundamental science. In front of you are two applications. One is from a team of physicists who want to explore a theory that implies that the world is canonically infinite. The other is from a team wishing to explore a different 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 are likely to achieve more good than if you fund the exploration of a theory that turns out to be false. Suppose that on the basis of all ordinary considerations, you judge the first application to be slightly stronger than the second one. Should you now reason that you ought nevertheless to fund the second proposal because you should ignore the possibilities in which there are infinite values (the possibilities under which the first proposal would deliver most benefits) and instead consider *only* possibilities in which there are only finite value (where the second proposal fares better)?

If the answer to the last question is "no", then the claim that we should ignore possibilities involving infinite values in our moral reasoning fails even a low methodological standard: it gives the wrong verdicts not just in farfetched counterfactual cases but also in some cases, like the Funding Body, that are quite likely to arise.

Despite these shortcomings of the proposal that we should ignore infinite outcomes, we shall revisit it in the concluding section after we have explored some other ideas with which it could be combined in such a way as may enable it to overcome some of the difficulties outlined above.

8. Passing the buck from ethics to decision theory

In light of the problem of imperturbable infinites, one move would be to scale back the ambition of aggregative ethics. Maybe such an ethics should content itself with specifying an ordinal ranking of the goodness of worlds and defining a right act as one that *in facts* leads to as good a world as any other feasible act does. This move would presuppose the success of the extensionist program discussed in section 3. Ethics would lay down the success criteria for an act being morally right but 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 been solved for all infinite cases should not count 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 given case should have a finite upper and lower bound (unless the decision problem happens to be one of the special cases where a plausible method for incorporating infinite values or an unbounded range of options exists). Such a requirement need not even be externally imposed but emerges naturally in subjectivist decision theory, which is founded on the concept of revealed preferences. If an individual is unwilling to gamble everything on a tiny (but finite) chance of getting the opportunity to spend an infinite amount of time in heaven, then that means that that individual does not have an infinitely strong preference for heaven. Subjectivist 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 disconnected from the evaluation of the plausibility of the theory. If some ethical theories refuse to play ball by making impossible demands on decision theory, that counts against those theories and in favor of other ethical theories that can be integrated in workable ways with decision theory.

What if the proponent of aggregative ethics went further and invoked not only an actualist ethics but also an actualist decision theory saying simply that we ought to decide to do the act that in fact would have the best moral results (or one of those acts, in case of a tie)? 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 plainly a pseudo-solution! It is easy enough to stipulate that we should decide to do an act that will in fact have the best moral consequences, but this injunction is non-actionable if we lack an effective way of figuring out which of the feasible acts is best in this actualist sense. At one place or another, the subjectively possible outcomes (even ones that would not actually obtain) 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 what consequences their acts would have. The problem of imperturbable infinities originates 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.

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 that someone fires a bullet into an innocent man's chest with the intent of murder. As it happens, the man survives and, fortuitously, the bullet hits 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. Those who take this line would see considerations concerning the probabilities of different possible outcomes as an integral part of moral reasoning.[34]

9. Rule-consequentialism and aggregate actions

In this section, we shall examine a group of approaches that attempt 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 connected. These larger units might either be *rules*, whose general acceptance can have wide-ranging consequences which we 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 (an idea to which we shall return shortly).

Consider first how rule-consequentialists might cope with some of the situations that cause ethical paralysis for act-consequentialists. The problem of imperturbable infinities, in its simplest form, 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 goodness even of a canonically infinite world. Rule-consequentialism, in its most rudimentary form, is the idea that an act is morally right if is recommended by the set of moral rules whose "general acceptance" would have the best consequences. If the population is infinite, as it is in canonically infinite worlds, then the general acceptance of a rule could easily have infinite consequences. These infinite consequences might then serve as a basis for evaluating the moral rightness of individual acts.

To make progress with this idea, we need to mate it with some principle for comparing the value of different canonically infinite worlds. Some principles for doing that were explored in earlier sections of this paper. Let us here consider another approach, the idea that we should identify the value of a canonically infinite world with its "average value", its value-density. To apply this idea, we need to assume that the cosmos is, on 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 spatiotemporal 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 as an index of the value of such a world, for the purposes of comparing it with other homogeneous canonically infinite worlds of the same order-type.

Unless one accepts a non-aggregative view such as average utilitarianism, one would not want in general to identify the value of a world with its value-density. For one might well judge that any world that contains an infinite amount of good and has a positive value-density is better than any finite world. One would then place canonically infinite worlds with positive value-density as lexicographically above all finite-value worlds (and canonically infinite worlds with negative value-density as lexicographically below all finite-value worlds). But within a class of homogenous canonically infinite worlds of the same-order type, value-density could be multiplied with probability to feed into the format of an expected-utility calculation, with value-density taking the place of utility.

Before discussing the limitations of this strategy, let us first consider an alternative approach that has many of the same advantages but which might be more congenial to act-consequentialists (and which might also potentially be of some independent interest).

Consider the case where there are an infinite number of exact copies of you spread throughout an infinite cosmos. (Note that this case is not farfetched. It is in fact empirically plausible, especially if the world is canonically infinite.[35]) Now, suppose that we conceive of "you" in a broader sense than usual – as not just this particular body but instead as the aggregate of all physical copies of you throughout the cosmos. Let us call this distributed aggregate entity "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 value-density of the world. 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. If the goodness of canonically infinite worlds is defined in terms of value-density, as suggested above, then the aggregate acts (ACTS) that are feasible for YOU could well be separated into right and wrong ACTS in a way that accords with our pre-theoretic intuitions – e.g. kind ACTS, which would make an infinite number of people happier, would increase the value-density of the world by some finite amount, and would thus tend to be rated as morally good. We can fit this idea into the standard decision theoretic framework by postulating that an act is morally right for you if and only if the corresponding ACT is right for YOU, which, in turn, is the case if and only if that ACT is one of the feasible ACTS for YOU that maximizes the expected value-density of the world. (For causal decision theorists, the last step would instead be: an ACT is right for YOU if and only if it is one of the feasible ACTS for YOU that has the greatest expected causal efficacy in raising the value-density of the world.)

Variations of this basic idea are possible. For example, 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".[36] How similar an instantiation would have to be to count as an instantiation of "the same decision process" as the one you are currently implementing is an issue that would need further specification, but one could argue that, e.g., if one atom in your finger, or in your brain, had been in a slightly different location that it was, but in a way that would not have had any meaningful effect on the reasoning you are conducting or the decision you end up making, then the decision process that you would have implemented would have been the same as the one you actually did implement. Thus one could aggregate a host of instantiations of decision-making throughout spacetime that are, by some suitable criterion, instantiations of the same decision process as your current one; and we can call the aggregate of the decisions made by these decision-processes YOUR DECISION. All the decision-processes whose decisions are included in YOUR DECISION would have to result in "the same decision" being made, or else they would not qualify as being the same decision-process. (A concrete example: one DECISION might be an aggregate of an infinite number of very similar implementations of some decision-process resulting in an infinite number of decisions to save a drowning child – where of course a different drowning child would be the object of each of these decisions, but in each case a child that was drowning near the corresponding decision-maker.) Aggregative ethics can then be formulated as saying that a morally right decision for you to make is one such that the DECISION of which this decision is a part has the greatest expected value (or expected causal utility) out of all the DECISIONS that one could have instantiated.

This approach, whether cast in terms of ACTS or DECISIONS, bears a resemblance to rule-consequentialism. It could in fact be described as a form of rule-consequentialism where the relevant rules are those that would deliver the best consequences if they were accepted – not generally by everybody or almost everybody as in most traditional version of rule-consequentialism – but by every person or every decision-process that is qualitatively identical (or sufficiently similar) to you or your own current decision-process. But this aggregate act approach might be more acceptable to act-consequentialists than rule-consequentialism is, because it is immune to one common objection against rule-consequentialism. An act-consequentialist is prone to question why she ought to do *A* rather than *B*, if *A* has better expected consequences than *B* while *B* happens to be in accord with a rule whose general acceptance would have better consequences and would thus be recommended by rule-consequentialism. This gap cannot arise in the aggregate act approach if we require that the expected goodness must be the same for all acts that form a part of a given ACT. In the finite case, an ACT will then be right according to the aggregate act approach if and only if all its constituent acts are right by act-consequentialist lights. In the finite case it therefore makes no difference whether adopts the act-consequentialist or the aggregate act approach. In the infinite case, however, it does make a difference. This is because the consequences of an individual act may be finite and hence impotent to affect a world's overall value whereas the

consequences of the corresponding ACT may be infinite and capable of changing the world's value. The aggregate act approach can consequently succeed in delivering plausible moral advice in some cases where traditional act-consequentialism either falls silent or absurdly implies that it is morally indifferent what we do.

One positive argument that could be adduced in favor of the aggregate act approach is that it can draw support from evidential decision theory. What one instance of a decision-process decides is relevant evidence about what other qualitatively identical instantiations of the same decision-process decide. The expected evidential value of saving a drowning child includes not just the value of this particular child being rescued but also the expected value deriving from the evidential linkage between the decision of this instantiation of the decision-process deciding to save this child and other analogous decisions by qualitatively identical instantiations of the same decision-process throughout the cosmos making analogous decisions to same drowning children in similar situations. Your choosing to save this child gives you evidence that your copies everywhere will tend to make the same choice regarding the children they see drowning. The expected value of the world conditional on your saving this particular child can therefore exceed the expected value of the world conditional on your not saving this child by an amount much greater than the value of this one child's life. Evidential moral decision theory entails that you should take into account these indirect evidential linkages and that they provide a reason for saving the child, a reason that is strong enough that it could operate even in worlds known to be canonically infinite.

Causal decision theorists will of course be unmoved by this argument, but they might have reason to support the aggregate act approach nonetheless if it could solve the problem of imperturbable infinites. The aggregate act approach would then be given a causal construal (along the line suggested above) and rather than being sold as an ethical application of a method of prudential decision-making it would instead be presented as a distinctly moral consideration which might be acceptable to some causal decision theorists in much the same way as ethical rule-consequentialism may be acceptable to some causal decision theorists despite there being no suggestion that rule-consequentialism somehow receives special support from causal decision theory.

The rule-consequentialism and the aggregate act approach both, however, have severe limitations when it comes to tackling the problem of imperturbable infinites. The value-density method entails giving ethical significance to the spatiotemporal distribution of values. Even if this consequence is accepted, other difficulties remain. The value-density method cannot be applied at all to worlds for which a value-density is not defined, such as the canonically infinite but inhomogeneous worlds of Example 9.

w16: 1, 2, 3, 4, 5, 6, 7, 8, …
w17: 1, -1, -1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1, -1, -1, 1, …

Example 9

The value-density approach also falters if single locations can have infinite value, and additional problems arise if we allow the possibility of values of higher orders of infinity.

Moreover, the value-density method cannot be straightforwardly applied to situations where worlds resulting from our acts may be of different order-types, such as in Example 10.

w18: 8, 8, 8, 8, 8, 8
w19: 7, 7, 7, 7, 7, 7, …
Example 10

Here, w18 has a higher value-density than w19 but only a finite number of locations. World w19, which has an infinite number of locations, can plausibly be regarded as much better than w18. If worlds can contain more than $\aleph_0$ locations, they may similarly be regarded as better even if they have a somewhat lower value-density.

Given that the value-density approach is already committed to giving ethical significance to the spatiotemporal distribution of value, a proponent of this approach might want to take the further step and maintain the order-type of infinite worlds that have the same cardinality could also be ethically significant. Thus, one might hold that w20 is better than w18 despite having lower value-density (Example 11).[37]

w18: 8, 8, 8, 8, 8, 8, …
w20: 7, 7, 7, 7, 7, 7, …, 7, 7, 7, 7, 7, 7, …
Example 11

Yet another problem is that if a canonically infinite world contains only finitely many exact or similar copies of you (each of which can affect only a finite amount of good or bad) then even the aggregate acts or decisions taken by this set of copies would be incapable of affecting the total value of the world. (A rule-consequentialist could encounter the same problem. This could happen if, e.g., there are only finitely many agents in the world but still an infinite amount of good or bad such as pleasure or pain in an infinite population of beings who are not agents.)

To this latter objection it could be responded that an agent could never reasonably assign a zero or infinitesimal probability to there being an infinite number of exact of similar copies of herself and that this would be enough to avoid ethical paralysis. An act maximizing the expected value of the world would simply be one that maximizes the value under the eventuality (whatever its precise probability) that there are infinitely many copies of the agent. However, while this reply has a point, the same line of reasoning would also tend to show that an agent may not be able to reasonably assign zero probability to there being infinite value at a single location or to any of the other abovementioned problematic possibilities. Consequently, there may not be *any* worlds were the value-density approach works. Even where the conditions are optimal for this approach, an agent could not reasonably be absolutely certain that they are, and so the possibility would always remain that the world is one of those for which the value-density approach breaks down. This would make the expression for the expectation value

of any act contain undefined terms, ruining the idea that a morally right act could be defined as one that maximizes this expectation.

The value-density approach would therefore have to be amended to take account of the size and order-type of a world, in addition to its value-density. This brings us back to the challenges facing the extensionist program that we discussed in section 3, and in particular to the problem of providing a function that represents the values of worlds in terms of quantities that can be combined in some meaningful way with probabilities to connect the axiology to decision theory.

If we try to meet this challenge by giving higher-order infinities lexicographic priority over lower-order infinities and finite values, then the resulting decision principles come to resemble the principles discussed in section 4, except that either aggregate acts or rules take the place of individual acts as the focus of evaluation. This would make the rule-consequentialist or aggregate act approach dependent on a stabilizing empirical assumption in much the same way as the original individual act approach (as we saw in section 4) – and with much the same disconcerting implications.

However, the implications are not *exactly* the same. The rule-consequentialist or aggregate act approach might have some advantage because the stabilizing empirical assumption on which they would depend may be somewhat more robust than the stabilizing empirical assumption needed by the individual act approach. A stabilizing empirical assumption, remember, is an assumption to the effect that the acts that are best from the perspective of maximizing lexicographically higher infinite values are similar – in most circumstances we are likely to encounter – to the acts that are best from the ordinary perspective which considers only finite values.

What kind of stabilizing empirical assumption is needed by the individual act approach? On this approach, the most likely scenarios in which infinite values come into play are relatively farfetched. These scenarios typically imply either that infinite benefits or harms are bestowed on an finite set of people or else that individually finite benefits or harms are distributed over an infinite set of people but that the decision-making agent is special in that the outcome for all these people depends especially on his or her acts. Scenarios of this kind diverge quite radically from our empirical understanding of the world. We may therefore lack confidence that the acts that would be best in such fanciful scenarios would also be best (or even acceptable) according to common sense morality. The empirical assumption that asserts that such harmony exists is therefore open to serious doubt.

Contrast this with the stabilizing empirical assumption needed by the rule-consequentialist or the aggregate act approach. The most likely way in which infinite values come into play on this kind of approach is less farfetched. It simply involves an infinite number of agents each affecting a finite amount of good or bad. This could well the case in the actual world. We may therefore assume with more confidence that what the theory says would be the best thing to do in this kind of scenario is similar to what common sense morality dictates. For example, if these infinitely many agents act kindly, in a world like the actual one, then each of them will tend to do a little good, and the aggregate effect is likely to be better than if they all act viciously.

There is a complication. The proponent of rule-consequentialism or the aggregative act approach should not take too much comfort in the remarks above. Even if we could be confident that their theory could handle the simplest infinite possibility, in all but some rather farfetched scenarios, this is of no avail if reasonable agents can, in the actual case, assign a finite non-zero probability to the world being such that higher-order infinite values are at stake. If an agent assigns a finite non-zero probability higher-order infinite values being at stake, then the scenarios in which these values are affected would completely dominate her ethical deliberations, given that such values take lexicographic priority over finite and lower-order infinite values.

The key question is therefore not whether rule-consequentialism or the aggregate act approach does well with the lowest-order infinite values, but rather whether they do well with arbitrarily higher-order infinite values. Is it the case that, for each level of infinite value, the acts that these approaches recommend as the best way to maximize values of that level are similar to the acts recommended by common sense morality? If our concern is whether these approaches have an advantage in this respect over the individual act approach, then we must compare (A) the least improbable scenarios in which according to rule-consequentialism or the aggregate act approach higher-order infinite values are at stake with (B) the least improbable scenarios in which according to the individual act approach such values are at stake. To the extent that A-scenarios are more similar than are B-scenarios to what we take to be the actual world, we might then reason (tenuously) that adopting the rule-consequentialist or the aggregate act approach confers some advantage in terms of being able to make to with a more robust stabilizing empirical assumption. If at each level of infinite value, the most plausible way for a value of that level to be at stake is by virtue of there being a corresponding infinite order of finite beings, similar to ourselves, whose aggregate activities or hypothetical rule-following would influence the probability of that value existing, then it seems that the rule-consequentialist or the aggregative act approach could get away with making a less vulnerable stabilizing empirical assumption than could the individual act approach. In so far as this is the case, they are, in this respect at least, theoretically better supported than the individual act approach.

10. Using non-standard analysis

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.[38]

There exists a well-developed mathematical theory of the so-called hyperreal numbers, numbers that can be infinitely large or infinitesimally small. The 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 a hyperreal framework so it can process these values. The task for aggregative ethics is more complicated, 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, and it has perhaps never even been attempted. However, let us explore this option and see how far it can take us.[39]

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.) The hyperreals thus lie 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 it look like they could form the basis of a promising framework for analyzing ethical problems involving infinite goods.

While we lack the space for a thorough introduction to nonstandard analysis, it might nevertheless be useful to provide a thumbnail sketch of the nature of these hyperreal numbers before turning to the question whether introducing them solves the problem of imperturbable infinities.[40]

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 $a + (-a) = z$ and $a * (a^{-1}) = e$

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 the identically. For instance, in ${}^*R$ there exist 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.

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:

$(1, 2, 3, \ldots)$

and of an infinitesimal hyperreal, its inverse:

$(1/2, 1/3, 1/4, \ldots)$

The product of these two numbers equals the (finite) number $(1, 1, 1, \ldots)$. 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 slightly 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 Leibnitz than the "epsilon-delta limit" approach that is the common fare in introductory university 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 now consider how the introduction of the hyperreals might help solve the problem of imperturbable infinities.

For a start, we need a way to map worlds containing some distribution of value at its various locations 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 can 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 to map 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 the real in the preceding place in the hyperreal's sequence.[41] If the local values in a has a one-dimensional essential natural order which is infinite in one direction, $(v_1, v_2, v_3, v_4, \ldots)$, it's value will thus be represented by the hyperreal $(v_1, v_2+v_1, v_3+v_2+v_1, v_4+v_3+v_2+v_1, \ldots)$.

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, …) = $\omega * 2$
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, …) = $\omega + 2$

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, …) = $\omega + 2$

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

w21: 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*) the following world:

w22: 1, 4, 1, 1, 1, 1, 1, 1, …

Since w22 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 smaller than 1/3.

This "finite-sum" version of the hyperreal approach thus has some things going for it. To deal with cases in which the locations do not have the order-type of the natural numbers, we can employ the method described in the previous section, taking each place in the hyperreal sequence to be the sum of the real at the preceding place and the value of a constant-volume expansion around the spacetime point occupied by the decision-maker. This will cover both cases where the past is infinite as well as the future and cases involving more than one dimension.

Even with this extension, however, some kinds of case are not yet covered by the present approach. These include cases where individual locations have infinite value (whether $\aleph_0$ or even higher cardinality). Worlds with the order-type of w20,

w20: 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 (i.e., in this case, (ω*7) + (ω*7) = ω*14), but if we consider a world like w23, where the decision-maker is located in the first segment of ones,

w23: 1, 1, 1, 1, 1, 1, …, …, -2, -1, 0, 1, 2, 3, …

we run into trouble because we then lack a preferred location for the expansion that we want to define the value of the second segment. For instance, if we start expanding in both directions from the location which has the value –1, we get the hyperreal value (for second segment of the world):

(-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

Combining this hyperreal approach with the causal approach would be helpful here, since – if we assume no action at distance, or at least not an infinite distance – then the causal approach directs the agent positioned in the initial segment of ones to ignore the entire second segment in their ethical deliberations. More generally, casual effects emanating from an agent and propagating in a continuous fashion would define a connected spacetime region for which the constant-volume expansion method could then be used to sequential ordering of local values such as could be mapped to a hyperreal in the manner described above.

Before concluding this section, we should make two observations about this proposal to use hyperreals in conjunction with the causal approach. First, it shares with several of the other proposals that we have examined in this paper the feature that considerations about infinite values always trump considerations about finite values (unless the relevant probabilistic differences are infinitesimal). This, as we have seen, leads to potentially disturbing consequences.

The second observation concerns the use of a non-principal ultrafilter in nonstandard analysis; we noted that the choice of such a filter is arbitrary. For the mathematical uses of nonstandard analysis, this multiple insatiability causes no problem. However, it might be thought of as undesirable to have such arbitrariness in the foundations of our 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. A case where this happens is illustrated in Example 12.

w24: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
w25: 1, -2, 1, 1, -2, 1, 1, -2, 1, 1, -2, 1, 1, …

Example 12

These two worlds are assigned the following hyperreal values:

Value(w24): (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …)
Value(w25): (1, -1, 0, 1, -1, 0, 1, -1, 0, 1, -1, 0, …)

w24 is identical to w25 in an infinite number of locations, greater than w25 in an infinite number of locations, and smaller than w25 in an infinite number of locations. By selecting a suitable ultrafilter, we can thus get either Value(w24) = Value(w25), or Value(w24) > Value(w25), or Value(w24) > Value(w25). A proponent of the hyperreal approach could argue that this indeterminacy is actually a subtle virtue because it matches a similar indeterminacy in our intuitions about the relative merits of w24 and w25. Alternatively, she could issue a promissory note that additional constraints could be specified that would remove the indeterminacy.

11. Combination therapies?

The problem of imperturbable infinities threatens to blow a large fleet of popular ethical theories out of the water. This under-appreciated problem is potentially more serious than traditional objections against aggregative ethical theories because it threatens to show that these theories fail even by their own criteria by implying that it is ethically indifferent what we do. This would be a reductio by almost everyone's standards.

The problem of imperturbable infinites challenges proponents of maximizing aggregative ethics to defeat the *prima facie* case their theory suffers from complete paralysis. Aggregative ethics implies that canonically infinite worlds contain an infinite amount of positive and negative value. Cosmology tells us that the actual world might well be canonically infinite. It is certainly the case that reasonable agents should assign a finite non-zero probability to the world being canonically infinite. If maximizing aggregative ethics, coupled with this fact, implies that it is morally indifferent what we do, then this kind of ethics should be rejected. The problem of imperturbable infinities cannot easily be brushed aside, but must be solved, or maximizing aggregative ethics is sunk. Going down with it would be a much larger class of ethical theories that contain some crucial maximizing aggregative component.

Among the various approaches surveyed, none provided a completely satisfactory solution. The most promising methods of avoiding the reductio involved modifying aggregative ethics in some significant way and making problematic empirical assumptions. We argued that the approaches using *discounting*, a *low-probability cutoff*, or adopting the *actualist stance* are dead ends. We also considered what we called the *extensionist program*. What noted that, unless combined with other countermeasures, it is of extremely limited use. The ordinal rankings it delivers in some cases (and which it aims to deliver in more cases) are quite useless for the kind of probabilistic decision theory needed by human agents who are always operating under conditions of uncertainty.

We therefore explored how the extensionist program in axiology could be combined with *extending decision theory* by introducing a rule like EDR, to specify how we ought to decide when infinite outcomes are at stake. We argued that adopters of EDR would need to rely on some *stabilizing empirical assumption* in order to avoid being confronted with radical and potentially unacceptable implications about what people ought to do. It would need to be assumed that the acts most likely to promote infinite goods are also morally acceptable from our ordinary finite perspective. The reason why this assumption would be necessary is that, according to EDR, considerations of infinite values take lexicographic precedence over considerations of finite values in every case where a finite non-zero subjective probability is reasonably assigned to such infinite values being linked to our actions. Further, users of EDR would need to assume that the acts that are more likely to promote higher-order infinite goods are similar to those that tend to promote lower-order infinite goods. Without the empirical assumption that such a convenient harmony holds between the higher-infinite levels of consideration and the levels of considerations that we naively take to determine morally acceptable conduct, EDR would threaten to ride roughshod over our pre-theoretic moral intuitions by

recommending behavior that most people would regard as fanatical if not deranged. To satisfy a reasonable methodological standard, this empirical assumption would need to hold for (at least) the cases we are most likely to actually encounter.

The *rule-consequentialist* and the *aggregate act* approaches offer interesting amplification devices which could enlarge the objects of ethical evaluation so that they become comparable to the infinites we would need to perturb in order to make a difference to the value of a canonically infinite world. The aggregate act approach is supported by evidential decision theory. Yet even those who reject evidential decision theory in favor of causal decision theory for prudential rationality could embrace either rule-consequentialism or the aggregate act approach as a specifically moral construct.

We introduced the concept of *value-density* and noted its potential symbiosis with rule-consequentialism and the aggregate act approach. World evaluations based on value-density would enable a certain class of infinite-valued worlds to be evaluated in a way that could naturally be combined with probabilities within a maximizing consequentialist framework. If agents' beliefs were conditionalized on the actual world being in this class, then the value-density approach would deliver useful and (at least by aggregationist consequentialist lights) plausible moral advice.

We noted, however, that there are cases to which the value-density approach does not apply. Example 9 illustrated two such cases, where the value-density of a world is undefined because the value-distribution is inhomogeneous. Other cases include worlds that contain infinite values at single locations, or contain more than $\aleph_0$ locations of value, or whose order-type otherwise differs from ω.

An obvious way for aggregationists to try to tackle such cases is via a principle like EDR. According to such a principle, considerations about finite worlds would be trumped by considerations about worlds where infinite values are at stake. For example, if an act had any effect, however tiny, on the expectation of value-density in a canonically infinite world of order-type ω, and if a finite non-zero probability is reasonably assigned to such a world being the actual one, then any consideration about the consequences of this act in finite worlds – however catastrophic these consequences would be – would be completely trumped. And the considerations about how the action plays out in worlds with order-type ω would, in turn, be trumped by considerations about how it plays out in worlds where higher-order infinite values are at stake. In this scheme, the concept of expected value-density is demoted to the role of a tiebreaker for cases where two acts are exactly on a par as regards their expected effects on all possible higher-order values (or all such possible values which an agent reasonably assigns a finite non-zero probability to being at stake).

Such a modified version of EDR would, like the original, bring with it the need for making a stabilizing empirical assumption. The needed assumption is that the acts favored by considerations about higher-order infinite values do not diverge too radically from the acts favored by common sense morality, at least not in the cases we are most likely to encounter. Yet cases that are not very farfetched and in which there is at least a modest degree of such divergence can be described. Consider for example the following

variation of the Funding Body thought experiment, in which the rule-consequentialist and the aggregate act approaches have peculiar consequences.

> *The Funding Body (reversed)*
> As before, you are tasked with selecting between two competing physics research proposals: either a project that will explore a theory implying that the world is canonically infinite, or a project that will explore a theory implying that the world is finite. The two projects are roughly balanced by all ordinary criteria but this time the second project has the slight upper hand. You believe that the exploration of a true theory is likely to do more good than the exploration of a false theory. The unusual consideration now, which purportedly should tip the scales in favor of the first project, is that if the theory which the second project would explore is correct, then the world is finite, and because any increase in the value-density of a finite world is lexicographically trumped by an increase in the value-density of a canonically infinite world, it would therefore be better to fund the first project, since doing so holds out the prospect of the benefit of having explored a true theory obtaining in a world that is canonically infinite.

Were this consequence utterly unacceptable, the rule-consequentialist and the aggregate act approaches would have to be rejected. However, the consequence in this reversed Funding Body is less clearly wrong than the conclusion of the original Funding Body (in section 7). In the reversed version, there is at least an intelligible rationale for why one might suppose that the project least favored by ordinary considerations should nevertheless be funded, namely that the general acceptance of the rule saying that this should be done (or the aggregate act of thus funding) would produce a greater expected value than the alternative. The strangeness of this particular consequence is hence not a decisive reason for rejecting either the rule-consequentialist or the aggregate act approach. But other cases might exist in which these approaches would have more outrageous moral implications. It is not clear how confident we should be in the empirical assumption that all such cases are sufficiently farfetched to enable these approaches to meet some reasonable methodological standard.

The *causal approach* (section 6) might be thought to avoid the strange consequences of both the original and the reversed Funding Body thought experiment. This approach tells us to focus on the causal consequences of our own acts, i.e. the (expected) changes we might bring about in the world. If, in the context of the Funding Body thought experiments, these expected changes are guaranteed to be finite, whether the world is finite or canonically infinite, then this form of aggregationism will give us no perverse reason to favor one of the funding proposals over the other.

We noted that casual decision theory, in its standard rendition, is unusable for any agent that assigns a finite non-zero probability to the world being canonically infinite. This is because causal decision theory is standardly formulated in a way involves terms referring to the value of an entire world. We can overcome this problem by recasting the

theory in the format of schema (D). Schema (D) can be fleshed out in accordance with either causal or evidential decision theory.

Yet even if we use schema (D) and limit our evaluative scope to the causal consequences of our actions, and we weight these consequences by the degree to which we have causal control over them, it is still possible for infinite values to leak into and potentially short-circuit our moral decision-procedures. For it seems that a reasonable human agent should assign a tiny but finite probability to the hypothesis that her acts may causally control infinite values. Some theological speculations, futuristic technological scenarios, miscellaneous metaphysical hypotheses, and presumably many other specific scenarios that our philosophy hasn't yet dreamt of would have such implications, and we lack justification for categorically ruling out that some such possibility might be true. Here again, therefore, the need for making a problematic stabilizing empirical assumption arises.

One option is to introduce an ad hoc *cut-off* stipulating that certain possibilities should simply be ignored in ethical deliberation. We argued in section 7 that if one takes this route, it is better to point the scalpel directly to unmanageable infinite possibilities rather than hoping to eliminate such possibilities indirectly by cutting out low-probability hypotheses.

We can reduce the scope of the required cut-off by combining it with the causal approach. By doing this, we can spare some infinite possibilities that would otherwise have to be removed from consideration, namely those possibilities in which the world contains infinite values but where we can only causally affect finite values. This has the advantage of sparing those possible worlds that current cosmology indicates are the empirically most plausible candidates for being the actual world.

Introducing a cut-off into our ethical theory means that it can no longer satisfy the highest methodological standard, because there will be possible cases about which the scarred theory delivers erroneous verdicts. For example, if we postulate that we should maximize the expectation of the value that we causally bring about, but that in doing so we should ignore any scenarios in which we have infinite causal power, then our theory will make incorrect claims about those possible cases in which we are reasonably convinced that we do indeed possess infinite causal powers. For in such cases, we should plainly take our potential infinite causal consequences into account. The cut-off strategy (even when coupled with the causal approach) can introduce awkward gaps between what is intuitively right and what the ethical theory implies is right not only in cases where we are convinced that our actions might have infinite causal consequences, but also in cases where we assign this merely some non-trivial probability. (A case of this sort could be illustrated by a "causal" version of the Funding Body thought experiment.[42])

Finally, we examined the hyperreal option. Using a framework that includes hyperreal numbers seems very promising, provided use an appropriate method select which hyperreal is to represent a given distribution of local values. Such a method was sketched in section 10. This approach has several advantages: it needs no special rules for handling infinite cases; it can handle diverging sequences of local value (by contrast to e.g. the value-density approach); it doesn't presuppose a rule-consequentialist or

aggregate act approach (and thus works even in worlds where there are only finitely many "sufficiently similar" copies of you); and it employs a neat, well-developed, independently useful mathematical formalism. As a solution to the problem of imperturbable infinities, however, the hyperreal approach still has some shortcomings: it involves placing ethical significance on the spatiotemporal distribution of local values; it makes arbitrary determinations of the relative merit of many world-pairs based on the selection of an ultrafilter; and it fails to rank certain kinds of worlds that have infinite local values or a complicated order-type. The last of these shortcomings could be substantially alleviated by combining it with the causal approach.

12. Conclusion: At What Cost a Cure for Paralysis?

The problem of imperturbable infinities is intertwined with two related questions: How can making finite changes affect the value of a canonically infinite world? How can ethics remain sensible if absolute lexicographic priority is given to concerns involving infinite values?

The hyperreal provides the best available answer to the first question, but – as we saw – only a partial answer. Even if it could be supplemented with principles to eliminate (the unwanted parts) of the arbitrariness in the selection of ultrafilter and to extend it to cover complicated order-types, it comes at the cost of investing the spatiotemporal distribution of local value with ethical significance. This cost is common to all approaches that make fine value-discriminations between canonically infinite worlds.

Another general conclusion is that aggregative ethics must make stabilizing empirical assumptions. If we do not cut infinite possibilities from consideration but instead seek to integrate them into our ethical decision-making e.g. via the hyperreal framework, we need to make a stabilizing empirical assumption. This is the case even if we adopt the causal approach, for lacking grounds for assigning a zero or infinitesimal probability to our actions having infinite causal consequences, the prospect of such infinite consequences overwhelms any considerations involving merely finite value. We thus need to assume that the acts that would have the greatest chance of creating infinite goods or of averting infinite bads are similar to the acts that would seem justified from our everyday finite perspective. If there is too radical a divergence between these two kinds of act, either in the actual case or in cases we think of as close to the actual case, then an aggregative consequentialist theory that does not cut these potential infinite values out from consideration altogether would have unacceptable implications about what we ought to do, either in the actual case or in those close possibilities. Such a theory would fail by any reasonable methodological standard.

In section 4 and other places of this paper, we briefly considered how plausible it is that the needed sorts of empirical stabilizing assumptions hold. We found that it was not clear that they fail, but that substantial concerns remain that many agents might have reasonable beliefs such that, according to aggregative ethics taking this approach, it would be right for those agents to behave in ways most of us would regard as fanatical

and perhaps unethical. It is not evident that there is any substantial difference in the plausibility of the stabilizing empirical assumption needed by the causal approach and that needed by a non-causal approach or by a rule-consequentialist or aggregate act approach. Combining infinite values with maximizing consequentialism, like letting elephants into a porcelain shop, is inherently perilous.

If we *do* introduce a cut-off, either alone or in combination with another countermeasure (such as the causal approach), we can remove the problematic infinities from our ethical deliberations. But doing so gives rise to a new set of problems. To completely ignore worlds containing problematic infinite values can cause distortions in our deliberations. The distortions are most pronounced in cases where an agent is reasonably convinced that infinite values are directly at stake, but distortions can also manifest themselves more indirectly, e.g. in situations analogous to those described in the various Funding Body thought experiments. Again the need arises for making an empirical assumption, namely that we are unlikely to find ourselves in situations in which significant distortions would occur. For if we are likely to encounter moral problems in which using a cut-off produces moral recommendations that are clearly unacceptable, then this theory fails meet even a low methodological standard. The cut-off approach has the further demerits of being theoretically ugly and unjustified.

What might help, somewhat, are non-consequentialist *side-constraints*. We could add to the maximizing component of our aggregative ethics side-constraints limiting the circumstances in which we are to engage in maximization or restricting the means we are permitted to use to pursue the goal of maximization. An ethical theory constructed along these lines might say that we ought to do one of those feasible acts that (1) satisfy certain side-constraints (e.g. it must not involve unjustified killing, lying, cheating, stealing, etc.) and (2) is among the best acts according to a maximizing criterion of the acts satisfying (1). (There would be different possible choices of a maximizing criterion, as described in preceding sections.) The advantage of such a constrained-maximization aggregative ethics is that the side-constraints could serve to fortify the theory by making counterintuitive consequences of the maximizing component less likely. This would reduce the theory's reliance on a stabilizing empirical assumption. Unless a great many side-constraints were added, the theory would probably still need a stabilizing empirical assumption, but the aid of a deontological flying buttress would enable the theory to get by with a weaker empirical assumption than it could without such external stabilizers. It should be emphasized, however, that this course would take us away from aggregative consequentialism. Yet if the alternatives are sufficiently unpalatable, then a partial selling-out as might be the best option for salvaging at least some elements of aggregationism.

If the costs of holding on to aggregative ethics in the teeth of the problem of imperturbable infinities becomes too great, at some point even those originally favorable to such a theory may need to consider if more drastic rethinking is not called for: to give up aggregative ethics altogether. How eager or reluctant one should be to take this step depends, of course, on the merits of alternative ethical paradigms – a matter outside the scope of this paper.

Even if one rejects that aggregationism as a fundamental ethical theory, one may still find an important place for aggregationism in more limited contexts where it would be imbedded within some more encompassing ethical theory. To give just one example, one might hold that certain institutions ought to take a maximizing aggregative stance with regard to the interests of their constituencies; or, more weakly still, that an aggregative stance reflects one type of consideration that some institutions have a duty to incorporate into their decision-making. Using aggregation in this manner leads 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 each member. To the extent that utilitarian and other aggregationist ethical ideas find their way out into the world outside philosophy departments, it is usually in such a circumscribed capacity. Social choice theory, for instance, often find 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 that if a person is unwilling to make an arbitrarily large sacrifice for some chance, however slight, of obtaining a particular good, then that means that the person does not have an infinitely strong interest in that good. We also find echoes of utilitarianism or aggregationism in such policy tools as cost-effectiveness analysis, impact statements, and QUALY-based evaluations of health care policies. In these real-world applications, what is being used is some form of circumscribed aggregationism rather than the ambitious unlimited kind of aggregationism that we encounter in foundational ethics.[43]

References

1. Moore, G.E., *Principia Ethica* (Cambridge: Cambridge University Press, 1903).
2. Martin, J.L., *General Relativity*, 3 ed. (London: Prentice Hall, 1995).
3. Hawking, S.W. and W. Israel, eds. *General Relativity: An Einstein Centenary Survey*. 1979, Cambridge University Press: Cambridge.
4. Belot, G., J. Earman, and L. Ruetsche, "The Hawking Information Loss Paradox: The Anatomy of a Controversy," *British Journal for the Philosophy of Science* 50(2) (1999): pp. 189-229.
5. Broome, J., *Weighing Goods: Equality, Uncertainty and Time* (Oxford: Blackwell, 1991).
6. Vallentyne, P. and S. Kagan, "Infinite Value and Finitely Additive Value Theory," *Journal of Philosophy* 94(1) (1997): pp. 5-26.
7. Hamkins, J.D. and B. Montero, "With Infinite Utility, More Needn't be Better," *Australasian Journal of Philosophy* 78(2) 231-240.
8. von Neumann, J. and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944).

9. Sorenson, R., *Infinite Decision Theory*, in *Gambling on God: Essays on Pascal's Wager*, J. Jordan, Editor. 1994, Rowman & Littlefield: Savage, Maryland. pp. 139-159.
10. Schlesinger, G., *New Perspectives on Old-time Religion* (Oxford: Clarendon Press, 1988).
11. Lewis, D., "Causal Decision Theory," *Australasian Journal of Philosophy* 59(1) (1981): pp. 5-30.
12. Tipler, F., *The Physics of Immortality* (New York: Doubleday, 1994).
13. Linde, A., "Life After Inflation," *Physics Letters B* 211 (1988): pp. 29-31.
14. Dyson, F., "Time without end: physics and biology in an open universe," *Reviews of Modern Physics* 51(3) (1979): pp. 447-460.
15. Cirkovic, M. and N. Bostrom, "Cosmological Constant and the Final Anthropic Hypothesis," *Astrophysics and Space Science* 274(4) (2000): pp. 675-687.
16. Bostrom, N., "Astronomical Waste: The Opportunity Cost of Delayed Technological Development," *Utilitas* 15(3) (2003): pp. 308-314.
17. Parfit, D., *Reasons and Persons* (Oxford: Clarendon Press, 1984).
18. Mill, J.S., *Utilitarianism* (London: Parker, Son, and Bourn, 1863).
19. Adams, R.M., *Must God Create the Best?*, in *Ethics and Mental Retardation*, J. Moskop and L. Kopelman, Editors. 1984, Reidel Publishing Company: Dordrecht. pp. 127-140.
20. Gorovitz, S., *The St. Petersburg Puzzle*, in *Expected Utility Hypothesis and the Allais Paradox*, M. Allais and O. Hagen, Editors. 1979, Reidel: Dordrecht. pp. 259-270.
21. Weirich, P., "The St. Petersburg Gamble and Risk," *Theory and Decision* 17(2) (1984): pp. 193-202.
22. Broad, C.D., "The Doctrine of Consequences in Ethics," *International Journal of Ethics* 24(3) (1914): pp. 293-320.
23. Bostrom, N., "Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation," *Journal of Philosophy* 99(12) (2002): pp.
24. Sobel, H., *Logic and Theism: Arguments For and Against Beliefs in God* (Cambridge: Cambridge University Press, 2004).
25. Conway, J., *On Numbers and Games* (New York: Academic Press, 1976).
26. Robinson, A., *Non-standard Analysis* (Amsterdam: North-Holland, 1966).
27. MathForum, *Nonstandard Analysis and the Hyperreals*.

---

[1] 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 problem of imperturbable infinities thus threatens to arise if *either* the sum of the values of the parts is infinite *or* the value that the parts have as a whole is infinite [1].

[2] See e.g. [2].

[3] In fact, in an infinite universe, it seems that there will even be an infinite number of people spontaneously materializing in gas clouds or from black hole radiation, for even though such occurrences would be extremely improbable, their chance is, according to quantum physics, some finite number greater than zero, and they would therefore be expected to happen infinitely many times in a universe that contains an infinite number of gas clouds and black holes. 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. (See e.g. [3], p. 19: "[I]t is possible for a black hole to emit a television set or Charles Darwin" (p. 19). To avoid making a controversial claim about personal identity, Hawking and Israel ought to have weakened this to "… an exact replica of Charles Darwin". But see also [4].)

[4] See e.g. [5] and [6].

[5] 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.

[6] E.g. the series 1, -1, ½, -1/2, ¼, -1/4, …, $\pm(1/n)^2$, … 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.

[7] [6]. This paper, which builds on earlier works, represents the highest development of the extensionist program to date. References to the earlier literature can be found in that paper.

[8] But not uncontroversial. Hamkins and Montero have recently argued that it is incorrect; see [7]

[9] 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.

[10] Vallentyne and Kagan, p. 9.

[11] Again we are setting aside some technical complications unrelated to our present concerns.

[12] Vallentyne and Kagan propose some principles that are slightly stronger than SBI2, which we shall omit here in order to save space. In particular, they formulate a principle that can deal with cases where locations have more than one dimension and with some cases where the two worlds to be compared do not have exactly the same locations.

[13] 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 or metric.

[14] 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, $\aleph_0$.

[15] The order type $\omega + \omega^*$ 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 a 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.

[16] 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.

[17] Technically, one could define the so-called pair product between a probability and an ordinal value of a world to be simply the ordered pair of the probability and the ordinal. Each act would then be associated with a set of such ordered pairs. But this definition would, of course, leave completely unanswered the question how we are to determine what we ought to do on the basis of a list of our feasible acts together with their associated sets of ordered pairs.

[18] [8].

[19] 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. For some possible beings (such as God?) this might not be true, and additional problems are known to arise in such cases for decision theory in such cases (see e.g. [9]). The subjective probabilities referred to in EDR could be qualified as "reasonable" or "rational" credences if one wishes to maintain that an ideally motivated but imperfectly rational agent might fail to choose to do the right thing. 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. Defendants of aggregative ethics 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 many practical purposes albeit not perfectly right "strictly speaking". We will not pursue this issue, which also arises for aggregative ethics in the traditional finite case.

[20] [10], p. 154.

[21] Extending EDR to recognize infinite values of different cardinalities is straightforward. It is less clear how to extend 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 matter whether get these finer discriminations right, for reasons we will explore later in this section.

[22] David Lewis argues that we not only *should* not, but *cannot* assign a zero probability such possibilities. See [11], p. 14.

[23] Some have seen it as a problem for views such as utilitarianism that they are too demanding, quite apart from any consideration about infinite values, in that they seem to imply that we ought devote more or less 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 fanaticism.

[24] Speculative scenarios of this kind have been described; see e.g. [12-15]. For a parallel with the finite case, see [16].

[25] [17], p. 486.

[26] It is not as obvious that it is possible for a location to have infinite value as it is that is possible for a world to have an infinite number of locations whose aggregate value is infinite. For starters, there is no empirical evidence that any location has or can have an infinite quantity of the sort of thing to which we normally assign value. It is even difficult to imagine how it could be the case. Two kinds of scenarios spring to mind: supernatural ones (in which, for instance, an infinitely powerful and valuable divine mind resides at a particular spacetime point) – the coherence of which many may find doubtful; and hypothetical physical ones, involving perhaps infinitely many people crammed into a finite space or a single mind performing an infinite number of computations and experiencing an infinite amount of "subjective time" within a finite objective time interval (cmp. [12]). It might, however, be possible to re-describe such physical scenarios in ways that would distribute the infinite value over an infinite number of locations (such as moments of subjective time in the infinite mind's mental history). Then a modified version of discounting could be applied to these locations.

[27] A further problem with this approach is that it threatens to ruin our ordinary principles for decision-making under uncertainty, even in the finite case. It could lead, for example, to reckless gambles that have

a smallish risk of a horrifically large negative-utility outcome seem worthwhile if they have a somewhat larger probability of delivering a very modest positive-utility outcome.

[28] In the Newcomb problem, two boxes are put in front of you. One box is transparent and contains a thousand dollars. The other box is opaque and may contain either a million dollars or nothing. You have the option of taking only the opaque box (which would net you either $1,000,000 or $0) or taking both boxes (which might net you either $1,001,000 or $1,000). It seems obvious that you should take both boxes. However, there is a twist. You know that whether there is a million dollars or nothing in the opaque box depends on what a famous psychologist, the predictors, has put in there. You also know that the predictor has a very strong track record of correctly predicting what choices people confronted with the Newcomb problem will make, and that the predictor will have put the million dollars in the opaque box if and only if she predicted that you would take *only* the opaque box. So you know that if you take only the opaque box, you will probably get $1,000,000, whereas if you take both boxes, you will probably get only $1,000.

[29] [11], p. 11.

[30] On the other hand, 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 [15].

[31] [18], chapter 2.

[32] 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. (If we think of an omnipotent God as choosing between different creation acts in which different possible worlds would be realized, then if for every possible world there is a better one, God would face a similar predicament. Everything God could do would be wrong. This consequence has led some to revise the above criterion for right action; see e.g. [19]. Maybe instead we should say that any act by somebody in that position would be right, or that any act that is expected to lead to least "rather good" results would count as right, even if there were other acts that would have been better.)

[33] The idea that small probabilities don't count is advocated in [20]; it is criticized in [21].

[34] For a discussion of some related themes, see [22]

[35] See [23].

[36] This paragraph has especially benefited from discussion with Eliezer Yudkowsky.

[37] Yet further complications may ensue 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.

[38] [24], appendix 2.2.

[39] 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 [25].

[40] The *locus classicus* is [26]. For an accessible primer, see [27].

[41] I am indebted here to Toby Ord (personal communication).

[42] *The Funding Body (causal)*

> Same as before except now your task is to select between the following two projects: one that would explore a theory that implies that we could develop infinite causal powers and another that would explore a theory that implies that infinite causal powers are impossible. And the problematic reasoning: The cut-off postulate implies that you should ignore the possible worlds in which the first theory is true. But since the funding provides most benefits if it goes to a theory that turns out to be true, then ignoring all the cases where the first theory is true would tip the balance in favor of funding the second theory.

This casual version of Funding Body might be a little bit less damaging for the theory that uses a narrow cut-off and a casual stipulation than the original Funding Body is for the theory that uses a cut-off of wider scope and no causal stipulation, because we might judge that the causal version of Funding Body is empirically somewhat more farfetched and thus that it matters less if an ethical theory implies something

counterintuitive about the situation it describes. Note that the relevant consideration is not directly how farfetched a world is where we have infinite causal powers, but rather how farfetched a world is in which we encounter a decision-problem like the one in this Funding Body.

[43] For helpful discussions and comments, I am very grateful to Tyler Cowen, Robin Hanson, Mitch Porter, John Broome, Jeremy Butterfield, Brad Hooker, Daniel Isaacson, John Leslie, Toby Ord, Howard Sobel, David Wallace, Tim Williamson, Alex Wilkie, and Eliezer Yudkowsky.