The University of Edinburgh

School of Philosophy, Psychology and Language Sciences

Infinite Ethics and the Social Discount Rate

B188154

Supervisor: Wolfgang Schwarz

Date of Submission: April 2025

Word count: 7983

There is a concept which corrupts and upsets all others. I refer not to Evil, whose limited realm is that of ethics; I refer to the infinite.

Jorge Luis Borges

1. Introduction

It's a live possibility that the universe is infinitely large. Carroll (2020) gives a representative view:

> [T]his is an open question in cosmology . . . the possibility's on the table [that] the universe is infinite, there's an infinite number of observers of all different kinds, and there's a possibility . . . that the universe is finite.

An infinite universe presumably contains infinitely many conscious beings, an infinite amount of pleasure and pain, and indeed, infinitely many persons arbitrarily similar to oneself.

This poses puzzles for aggregation-based consequentialists. If the good consists in somehow aggregating value across individual locations, then there is a twofold problem:

1. Is there a way to aggregate value across infinitely many locations in order to say whether one world is better than another?

2. Following the above, we are also interested in evaluating actions for which we are uncertain what outcome will obtain. So, is it possible to evaluate lotteries over infinite outcomes?

If the universe were finite in spatial extent, we would get the same problems if time were infinite in duration, as is predicted by many widely accepted models of cosmology (Carlsmith, 2022, 101). In his paper 'Infinite Ethics', Nick Bostrom (2011, 3) calls the worry that an aggregationist would be completely unable to make moral decisions in an infinite world 'infinitarian paralysis', and he is deeply troubled by it:

> This should count as a reductio by everyone's standards. Infinitarian paralysis is not one of those moderately counterintuitive implications that all known moral theories have, but which are arguably forgivable . . . The problem of infinitarian paralysis must be solved, or else aggregative consequentialism must be rejected.

An obvious response is to restrict the domain of ethics to the regions we can causally influence. This is arguably to reject aggregative consequentialism entirely, although sometimes the term is still used to refer to more restrictive varieties. But, as Arntzenius (2014, 52) points out, restricting ethics to what we can influence is more of a gesture in the direction of a criticism, rather than a worked-out view. Various ways that agents could have infinite causal influence have also been suggested in which, it is argued, we should have nonzero

credence.[1] And as a theoretical matter, we might still be interested in whether consequentialist ethics is possible in worlds where infinite causality is possible.

A different response is to reject the existence of infinity. Perhaps you could, like Aristotle, claim that there are only 'potential' and no 'actual' infinities (Moore, 2018, 32). But premising your ethical theory on the non-existence of infinity is a shaky foundation. The problems arise if you're at least unsure whether infinity exists. And Bostrom (2011, 38) notes that, even *conditioning* on the absence of infinities itself gives bizarre conclusions – for example, refusing to fund any scientific projects premised on infinity-involving hypotheses.

A (simple) lottery is a probability distribution over a set of finite or countably many outcomes. In standard decision theory, lotteries are evaluated by calculating their expected value. It has been an enduring puzzle whether and how the expected value framework can be applied to cases involving infinite possible outcomes. Pascal's Wager is presumably the most well-known case. In a naive expected value model, the expected value of an infinitely good outcome with any nonzero probability is still infinite, but accepting this at face value leads to contradictions and paradoxes (Hájek, 2024). There are nonzero probabilities of many *different* infinite outcomes, and there is no accepted procedure for how to compare expected values between them.

There has been a strong reluctance to accept the conclusion that our moral decision-making should be completely dominated by tiny probabilities of

[1] (Carlsmith, 2022, 101) gives an overview.

infinite outcomes. This worry is known as moral fanaticism. Indeed, there has been a strong reluctance even to accept that our decision-making should be dominated by tiny probabilities of extremely large *finite* outcomes, as in Bostrom (2009)'s case of Pascal's Mugging. I hope it will become clear that infinity poses difficulties which are qualitatively different from those in the finite case.

Even if one were in principle willing to bite the bullet on moral fanaticism, since there is no generally accepted way to evaluate lotteries across infinite outcomes, it's not at all clear how one could do it.

One reaction to all of this is to say that we are rationally justified in discounting sufficiently tiny probabilities precisely to zero. This view has few explicit defenders in philosophy, although Monton (2019) makes a respectable effort. But even if this view were correct, it's doubtful that the probabilities involved are nearly so low that it would help. An infinite universe is taken sufficiently seriously that our credence in it should be much higher than what the probability-discounters have in mind. Arguably, our credence in infinite causality should also be high enough so as not to be discounted (although this is far from central to my arguments). A veritable zoo of alternative decision theories has been proposed to evaluate lotteries over infinite outcomes, and the field has reached no consensus.[2]

[2] A recent favourite is to replace expected value with stochastic dominance reasoning (Tarsney, 2020).

The problems arise for other varieties of consequentialists. Non-aggregative consequentialism claims that the moral value of actions is determined entirely by their consequences, but does not presuppose that value can be divided into individual locations, or aggregated across them one at a time. But there's the rub: we don't know how to make comparisons when infinite consequences might be involved. In Moorean (1903) views, the value of an 'organic unity' is the value of its parts, taken in isolation, plus the extra value that emerges from their unity. That runs into the same difficulties if either the sum of the individual values is infinite, or if the value that the parts have 'as a whole' is infinite (Bostrom, 2011, 51). Similar cases arise in hybrid theories, or theories with deontological side-constraints, that involve some form of aggregating consequences. Any account of ethics or axiology that considers value across infinitely many locations, or infinite value at a single location, is in trouble.[3]

I'll be restricting my focus to the countably infinite. Work examining the moral implications of Cantor's transfinite arithmetic and beyond is still in its infancy. Joe Carlsmith (2022, 125) has argued that ignoring orders of infinity is a recipe for the same kind of 'rude awakening' that (countably) infinite ethics provided.

Several solutions have been proposed for these problems. In this paper, I'll be defending a deeply unpopular view: that nobody has yet proposed an alternative any better than discounting. For consequentialism to make sense in infinite worlds, a defender of such a view must apply some kind of discount rate

[3] Carlsmith (2022, 129) and Askell (2018) argue that infinite ethics is a problem for 'everyone', even quite distant moral theorists like Kantians and virtue ethicists. But they reason from a controversial premise set, including the qualitativeness of $\succcurlyeq$ (see §3), which I have no wish to assume here.

across the locations of value. There is no way to be truly impartial between them.

When I say 'discounting', I mean the downrating of the ethical significance of some value merely because of the location it appears in. I take no stance on the issue of what an aggregative consequentialist should discount with respect *to*. The point is that, if they want to take their theory seriously in the infinite domain, they have to discount with respect to *something*. I argue that this is a reasonable null hypothesis.

My position is so unpopular that this version of it has not been defended in print before. Certain details relate to mathematical results proved more recently than any infinite ethics publications, and other elements derive from personal correspondence with the authors.

In what follows, I'll be using this notation: $w_1 \succeq w_2$ is the binary relation that means that world 1 is at least as good as world 2. $w_1 \succ w_2$ means that world 1 is strictly better, and $w_1 \sim w_2$ means that worlds 1 and 2 are equally good.

I'll start by elaborating on the concept of a basic location of value. I will then consider various desirable properties for ethical comparisons to have, and show why they're not jointly satisfiable. §4 and §5 review the economics of intergenerational equity, and discuss how my view relates to the earlier debate over utility maximisation given an infinite horizon. §6 and §7 compare my view to the main proposed alternatives, 'Expansionism' and the use of the hyperreal number system, and finds them wanting.

2. Basic locations of value

In aggregative ethics, candidates for the basic value-bearing locations of a world have included acts, persons, space-time regions, and experience-moments. I take no stand on which is most plausible. Note that intergenerational economics commonly considers *generations* as the basic locations. It's sometimes unclear how literally to take this: there is essentially no philosophical work developing the idea of generations as fundamental units of ethical concern, and it doesn't seem especially plausible (Askell, 2018, 12).

For the sake of completeness, I'll also note that there is one special case of aggregative consequentialism which dramatically simplifies infinite ethics: lexical priority views. These hold that there are some goods for which any nonzero amount is better than *any* amount of any other good. This considerably simplifies the aggregation problem across infinite worlds, but they are unpopular.

Aggregative consequentialism is still a somewhat general view, in that you can use whatever algorithm you want to aggregate across the locations. That algorithm could be summation, weighted averaging, or some other kind of social welfare function (§5). The locations may also include fundamentally incomparable goods.

Early in this literature, it was noticed that moral judgements are particularly sensitive to our assumptions about the basic locations in the infinite case. Cain (1995) contemplated which of these worlds would be better:

> **The Sphere of Suffering:** Infinitely many immortal people start off happy, and from some point a sphere begins expanding at a uniform rate. As soon as someone is inside the sphere, they spend the rest of their life suffering.
>
> **The Sphere of Happiness:** Infinitely many immortal people start off suffering, and from some point a sphere begins expanding at a uniform rate. As soon as someone is inside the sphere, they live the rest of their life in happiness.

It's a tricky case. In the Sphere of Suffering world, at any one time, there are infinitely many happy people, and only finitely many suffering people. But, for each individual, they will spend only a finite time happy, and an eternity suffering. If you consider persons to be the basic locations of value, then the Sphere of Happiness is greatly preferred: each individual location will get infinitely more utility, compared to in the Sphere of Suffering world. But if you consider time as the basic location, then you get the opposite recommendation.

The next characteristic of infinitely many locations which was widely noticed is that aggregating across them is necessarily order-dependent. In general, infinite worlds will contain infinite amounts of value and disvalue, with no well-defined sum. The value of an infinite world 'as a whole', if it is coherent to speak of such

a thing, is particular to the order that you aggregate in. To my knowledge, no theory of infinite ethics even *claims* to avoid order-dependency, although arguments have been made about which orderings are most plausibly the 'essential natural order' (§6). This order-dependency becomes particularly apparent when we consider how to formalise certain desirable properties for an ethical ranking.

3. Pareto, Anonymity, Completeness

Suppose that we had solved the problem of creating an ordinal ethical ranking over infinite worlds. It's natural to hope that we would be able to extend these ordinal preferences to cover preferences over lotteries. By the result of von Neumann and Morgenstern (1944), if an agent has an ordinal ranking over all possible lotteries, then, subject to certain axioms, that is mathematically equivalent to a cardinal utility function to be optimised. On the face of it, that would give us all we need to have a moral theory that works across infinite possible outcomes. Unfortunately, it is almost certainly impossible to create such a ranking, for reasons that will become clear.

I will introduce some jargon. Weak Pareto is the condition that, if every location is at least as well off in world 1 as in world 2, then $w_1 \succeq w_2$. Strong Pareto is the condition that, if every location is at least as well off, and there is *some* location that is strictly better off, in world 1 than in world 2, then $w_1 \succ w_2$.

The next concept is Finite Anonymity, which says that, if there is an equal amount of value at two locations, then we can permute those locations while

holding the value constant, and the world will still be equally good. Another way of thinking about Anonymity is that, if two worlds have the same distribution of value, then our judgment about which world is better should not depend on the particular identities of the locations in question (Askell, 2018, 20). Strong Anonymity says that, if there is a value-preserving bijection from the locations in w_1 to the locations in w_2, then $w_1 \sim w_2$. Strong Anonymity claims that Anonymity holds even under infinitely many permutations.[4]

The Anonymity and Pareto principles can only make comparisons across worlds with the same value-bearing locations. None of them allows us to compare a finite world with an infinite world, for example.

Our first result is this: Strong Pareto and Strong Anonymity directly conflict. Suppose we have an infinite world with utility levels <1, 0, 0, 1, 0, 0 … >, and we increase the utility of every third location by 1 to give <1, 0, 1, 1, 0, 1 … >. This new world is better by Strong Pareto. It should be a particularly easy case: infinitely many locations have been made better off, and none have been made worse off. But by Strong Anonymity, the worlds are equally good; there exists a bijection by which the utility in the new world and the experiences in the old can be exactly matched up. This conflict does not assume that value is numerically representable.

[4] Askell (sometimes) calls Anonymity 'equity', and much of the cited economics research calls it 'neutrality'. In the economics literature, the term Sensitivity generally refers to Strong Pareto.

Given this, most authors have been inclined to reject Strong Anonymity (Carlsmith, 2022, 110). Indeed, the central thesis of (Askell, 2018), by far the most comprehensive account of infinite ethics, is that any theory thereof should be compatible with agent-based Strong Pareto.[5]

Two further desiderata for an ethical ranking are:

Completeness: For all w_1 and w_2, either $w_1 \succeq w_2$ or $w_2 \succeq w_1$.

And:

Transitivity: If $w_1 \succeq w_2$, and $w_2 \succeq w_3$, then $w_1 \succeq w_3$.

Transitivity has long been seen as a particularly secure principle, although see (Askell, 2018, 182) for some dissent.

Completeness immediately strikes many of us as the weakest of these assumptions. The idea that our ranking of worlds is necessarily partial may not seem so bad, especially if we can show that the worlds which are fundamentally incomparable are particularly unrealistic. But the worry is that, if we loosen Completeness *at all*, then we will lose the ability to rank even simple worlds for which we have extremely strong intuitions that one is better than another. Amanda Askell (2018, 72) has shown that, if they reject Completeness, a wide

[5] Fixed- and Variable-Step Anonymity are attempts to rescue the intuition of Strong Anonymity, while avoiding the paradoxes of considering infinitely many locations being permuted all at once, but they involve many issues beyond our scope (Askell, 2018, 26).

and realistic class of worlds are fundamentally incomparable for aggregationists, if they assume that the ≥ relation must be qualitative. A qualitative relation, in David Lewis's use of the term, is one that depends only on the intrinsic properties or qualities of the relata, and not on their particular identities. A full consideration of Askell's proofs, and of whether ≥ must be qualitative, would take us too far afield. I nevertheless note the context that the worry that any loosening of Completeness leads to 'ubiquitous incomparability' is widely held in the field.

It is easy to confuse several related terms here. A universal domain moral theory is one that claims to 'say' something about every conceivable moral scenario. A totalising theory claims that it is the *only* framework in which moral outcomes are to be evaluated. There is nothing contradictory about having a universal domain totalising theory which is not Complete. Such a theory would not provide a total ranking over all possible worlds, but it would provide general guidance even in evaluating theoretically 'incomparable' outcomes. There will only be a bidirectional relationship between Completeness and universal domain for aggregative theories that evaluate worlds by considering them along a single measure. Completeness is a *much* stronger condition.

So, while the incomparability of infinite outcomes is occasionally invoked as a reason to reject that ethics has 'universal domain', it is only utilitarians and their ilk who need have this precise worry. But for utilitarianism, a rejection of Completeness might be fatal, because it means a rejection of the universal domain of ethics. And many authors take a universal domain to be part of the definition of utilitarianism (McLaughlin, 2022a).

In the following sections, these concepts and vocabulary will help us to see why ethical comparisons based on extending certain principles of rational choice have proved to be so unsatisfactory.

4. The social discount rate

The term 'social discount rate' doesn't have entirely consistent usage. In economics, the social discount rate accounts *both* for an intrinsic privileging of time periods which are closer to the present, *and* for how future people are likely to be richer – and so, given the diminishing marginal utility of wealth, less likely to benefit from a given unit of resources. An intrinsic privileging of the present is called 'time preference', and is often represented by the symbol δ.

The question of how the social discount rate r and time preference δ are related is highly non-trivial. Under one widely used set of assumptions, they are related by the celebrated Ramsey-Keynes rule (McLaughlin, 2022b, 22). Further adding to the confusion, an individual's utility, as the term is standardly used in economics, is *already* adjusted for their time preference. 1 util today is the same as 1 util next year, by definition. It makes some sense that this convention has not carried over to philosophy, given that the rationality (or not) of discounting one's own utility is itself disputed.

The foundational paper of intergenerational economics is Frank Ramsey's 'A Mathematical Theory of Savings' (1928). In it, Ramsey famously argued that the inclusion of δ was only for mathematical completeness, and that intrinsically

discounting future welfare was a 'practice which is ethically indefensible and arises merely from the weakness of the imagination' (1928, 543).

Ramsey's position was prominent until around 1965, after which the view that we should not discount future welfare was almost entirely abandoned within economics (Van Liedekerke & Lauwers, 1997, 160). An enduring difficulty has been that, if we set δ to zero, we lose the ability to compare utility streams with an infinite horizon, and/or models give us no well-defined results. For this and related reasons, in economics Koopmans (1960, 1965, 1972) argued that it was extremely difficult to axiomatise rational choice in a way that did not logically entail some form of discounting. The cultural divide persists to this day: While it's common for economists to set δ at around 2%, philosophers are almost unanimous in advocating a zero rate of time discounting (Ord, 2020, 253).

The issue of discounting has been intimately tied up with the debate over whether utilitarianism is too demanding. If we do not discount the welfare of the geographically far away, then prima facie, enormous moral demands are placed upon the utilitarian (Singer, 1972). And if we do not discount in time, then prima facie the utilitarian's decision-making will be dominated by their effects on future generations, even at the expense of current ones. After the puzzles for rational choice posed by zero discounting, a reluctance to accept the demandingness of certain ethical theories has probably been the second largest source of academic support for discounting.

In ‘Against the Social Discount Rate’, Parfit and Cowen (1992, 159) provide a range of arguments for a zero rate of time discounting, which have been widely accepted:

> Remoteness in time roughly correlates with a whole range of morally important facts. So does remoteness in space . . . 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 of *n* percent per yard. The temporal discount rate is, we believe, as little justified.

My contention is that the strongest arguments against the discount rate in the finite case rest on intuitions about how moral importance cannot possibly be sensitive to orderings: that ‘It cannot be argued that [the] forthcoming slice of time is worth less simply because [we] must wait for it’ (Cowen, 2018, 67). But we already saw in §2 that we have order-dependence in the infinite case. This is not disputed by any of the current theories. We further saw that most of us are inclined to reject Strong Anonymity over Strong Pareto, in which case at least *some* of the permutations of locations of value must be morally important.

Finally, the reason why Bostrom (2011, 25) in his original presentation did not consider (spatiotemporal) discounting as a serious contender to resolve infinite ethics was because:

> For any given discount factor, we can consider worlds that have, centered on the decision-maker, a sequence of locations whose values increase at a

> faster rate than the discount factor discounts them, so that the sum of discounted values is infinite. To avoid this, we would have to postulate that the discount rate at some point becomes infinite, creating an ethics-free zone at some finite distance from the decision-maker – making a travesty of aggregative consequentialism.

I am left puzzled by all of Bostrom's aspirations for what a consequentialist theory of ethics can do. On the most straightforward reading, a consequentialist in an infinite world who has no 'ethics-free zone' is an agent with truly infinite moral concern. I do not have a view on whether such an agent could exist, or whether we need to develop a decision theory for them. My project is more modest: to suggest modifications so that ethically comparing worlds and lotteries is still possible, given that infinities may be involved.

5. Impossibility results

The conflict between Strong Pareto and Strong Anonymity is the simplest of the impossibility theorems[6] of infinite ethics, but there are many others. Understanding them requires the context that a social welfare function (SWF) is a function which maps vectors of bounded real numbers, usually representing utilities, to the reals:

$$f: \mathbb{R}^N \to \mathbb{R}$$

[6] This was first demonstrated rigorously by van Liedekerke and Lauwers (1997).

The output of a cardinal SWF generally represents the overall 'goodness' of a state of affairs. The output of an ordinal SWF is a numeric index that defines a (weak) social welfare relation. This is a reflexive binary 'at least as good as' relation over the set of possible outcomes.[7] Social welfare functions typically obey Weak Pareto and Transitivity by definition, which I will take to be the case here.

Their relevance is this: In any aggregative moral theory in which value is numerically representable,[8] the question of whether worlds can be compared subject to certain constraints is *equivalent* to the existence of a social welfare function obeying those constraints.

These theorems generally emerged in the context of comparing utility streams given an infinite horizon, but they apply equally to comparing any set of numeric values appearing at infinitely many locations. To explain them, I'll need to introduce one more axiom:

> **Continuity:** For a social welfare function *f*, if the values at a set of locations *l* converge to *l**, then *f(l)* converges to *f(l*)*.

Lauwers (1995) proved that the only way to have a Complete Strongly Paretian SWF, of any kind, obeying Continuity, under the assumption that value is linearly additive, is to have a discount rate (Askell, 2018, 37). The basic intuition

[7] I am following the Bergson-Samuelson tradition. In the Arrow (1951) tradition, the output of an ordinal SWF is a single social welfare *ordering* over all possible outcomes.
[8] This paradigmatically includes, though is not limited to, classical utilitarianism. For simplicity, I'll just refer to 'utilitarianism' in this section.

is that, if we don't have discounting, then any Complete ethical ranking obeying Strong Pareto will sometimes jump around wildly given tiny changes in its (numeric) inputs.

Dubey (2011) further showed that the only way to construct an ordinal SWF comparing infinite locations which is Complete and Finitely Anonymous involves a non-constructive proof invoking the axiom of choice. Non-constructive proofs are existence proofs which don't tell us how to construct a given object. Mathematicians have generally made their peace with the axiom of choice, and with non-constructive proofs, but there are reasons why this has widely been considered troubling in a philosophical context (Askell, 2018, 31). First, if the proof is non-constructive, it's not clear how much comfort a utilitarian should take from the mere fact that an ethical comparison with certain characteristics *exists*, if we don't know anything about how to actually find it. Second, if a proof relies essentially on the axiom of choice, that generally reflects the necessity of making some set of choices that cannot be specified by any rule or axiom. The allegation that such choices are arbitrary is a reasonable one, especially when they are about abstract mathematical objects with no basis in our moral intuitions. The concern is not about the mathematical methods themselves; the concern is specifically about using them to make a comparison which is claimed to be of ethical significance. It was for these reasons that Easwaran (2021, 2014) argued that no comparison which depends in its essentials on the axiom of choice can possibly be normative.

However, arguably the most relevant set of results for infinite ethics comes from Basu and Mitra (2003, 2007).[9] They show that it's impossible to have a cardinal SWF comparing infinite locations obeying Completeness, Strong Pareto, and Finite Anonymity. Zame[10] (2007) extended the Basu-Mitra theorem by showing that, even if we drop the assumption of Completeness, the existence of a Strongly Paretian Finitely Anonymous social welfare function which ranks infinite utility streams is independent of Zermelo-Fraenkel set theory with choice. That means that the existence of even a partial ordering obeying desirable axioms *cannot* be explicitly constructed. That is a much stronger and more troubling fact than direct reliance on the axiom of choice.

These impossibility theorems impose quite severe constraints on the sorts of comparisons a utilitarian can make across infinite worlds. Zame showed that there are difficulties even with a partial ordering. Given that, the infinite utilitarian might switch her focus to looking for subsets of utility streams that can be compared while obeying desirable axioms, and making arguments for why those subsets are the only ones that really matter. A subtly different idea is to think about subrelations: $\simeq$ is a subrelation to $\geq$ if $x \simeq y$ implies that $x \geq y$. When it's provable that there is no SWF obeying certain desiderata, it's common practice in economics to search for a subrelation that does. In the case of infinite comparisons, this project is deeply incomplete.

[9] Basu and Mitra can be seen as having shown that the more well-known theorem of Diamond (1965) still holds if we drop the assumption of Continuity.
[10] Zame's paper assumes that utility levels are bounded between 0 and 1, and chapter 3 of (Askell, 2018) shows how to generalise the result to any bounded value.

Here is my proposal: by embracing a small amount of discounting, which may not matter for any practical purposes, utilitarians will be able to compare *many* more infinite utility streams. A major advance on this front was recently made by Jonsson and Voorneveld (2018). They consider a model where a utilitarian is discounting the future geometrically, and is ranking infinite utility streams by comparing the limit inferior of the discounted differences between them, as the discount rate tends to zero. In this context, limit inferior means that, if the differences between the streams oscillate forever, then the model picks the smallest of the accumulation values (values hit infinitely often), in order to create a more stable and conservative ranking. This results in a social welfare relation the authors call limit-discounted utilitarianism (LDU).

LDU obeys Strong Pareto, Finite Anonymity, Continuity and several other desirable axioms, while being able to compare a wider range of utility streams than any method previously considered. In particular, if you combine LDU with the results of a more recent paper by Jonsson (2023), it is provable that LDU allows comparison between any utility streams generated by a stationary Markov decision process. That extension of their theorem is not in the original papers, but it was confirmed by personal correspondence with the author. The details are not relevant, but arguably, for most realistic utility streams, there exists a stationary Markov decision process which could have outputted that stream (West, 2017). LDU is still only a partial ordering, but the utility streams it can't rank are particularly unrealistic.[11] The existence proof of LDU is constructive, and it does not directly invoke the axiom of choice.

[11] See §4 of their paper.

Discounting geometrically may sound arbitrary, but because LDU takes the limit as we consider smaller and smaller discount rates, the particular functional form will not change how any two given utility streams are ranked relative to each other. For utilitarians, LDU is perhaps the most appealing proposal yet devised for constructing an ordinal ranking across outcomes.

The case against discounting overwhelmingly hinges upon intuitions about how absurd it is to regard people as less morally important just because they live in the future (or far away). But economists have rapidly been making progress on formalising social welfare functions for which the discount rate is allowed to asymptote to zero. So, it may be that, for any practical purposes, utilitarians can choose as low a discount rate as they want. The infinite utility streams which remain incomparable by the latest methods are quite unrealistic.

This brings us to an ambiguity in my position as stated thus far. The social discount rate has traditionally been used in intergenerational economics to ensure that the total value of a future utility stream is finite. Am I claiming that a self-consistent aggregative consequentialist must discount in such a way that the total value they ascribe to the universe is finite? LDU suggests that the answer is: not necessarily. There is some appeal to this vision of utilitarianism which, for any given decision, discounts at a nonzero rate to avoid infinity paradoxes, but has a discount rate which asymptotically approaches zero, such that the total value of the universe is unbounded. It is reasonable to ask whether there is any difference between taking the limit as discounting shrinks to zero,

and not discounting at all. But in either case, discounting approaches would have shown themselves to be the solution to our stated puzzles.

I will make one final point about the frontier of research on discounting. Holden Karnofsky calls discounting in time and space the 'stupid version of' discounting, and speaks of how it's something that the most cutting-edge discount utilitarians have moved beyond (Wiblin et al., 2023). His favoured proposal is called UDASSA ('Universal Distribution' + 'Absolute Self-Sampling Assumption'). This advocates for giving moral weight according to how 'simple' it is to specify a world and a location within it, as measured by (something like) Kolmogorov complexity. Advocates claim that this resolves several related issues in anthropic reasoning, decision theory, and infinite ethics (Carlsmith, 2022, 60). I mention UDASSA not to consider its merits, but because it is important to note that work is being done on discounting approaches for which there is some independent argument for why the parameter being discounting with respect to is *not* morally irrelevant. Mueller (2017) has developed a version of UDASSA based on transition probabilities between observer moments, in which observers are not at all discounted for physical distance. Such an approach would heavily discount the welfare of simulations, Boltzmann brains, and other beings with high algorithmic entropy who (it is argued) are unlikely to have a strong interest in their continued existence into the future. The ultimate hope is that the consequentialist can rescue the sense of impartiality that she cares about, while discounting in such a way that may actually be a *virtue*.

6. Expansionism

I follow Askell (2018) in using the name Expansionism for a family of related approaches to infinite ethics. The foundational intuition for these methods is that we can compare two worlds by considering spheres expanding at the same uniform rate in each of them. If there is some time T, where, for all times after T, the total value inside the sphere in world 1 is strictly greater than the total value inside the corresponding sphere in world 2, then $w_1 \succ w_2$. Expansionism doesn't inherently assume that that value is numerically representable: we can still have dominance relations between worlds if value is only qualitative. The first rigorous development of this idea was by Peter Vallentyne and Shelley Kagan (1997), hereafter V&K, and their theory crescendos in the following:

> **Generalized Metaprinciple:** $w_1 \succ w_2$ if, for every point x in w_1, there is some radius R_x where, for every $r \geq R_x$, the total value in the ball of radius r around x in w_1 exceeds the total goodness in the corresponding ball around the matching point in w_2.

This is only applicable if we consider locations to have an 'essential natural order', which V&K suggest might be true of spatial and temporal regions, but not of people (1997, 9). They arrive at the Generalized Metaprinciple after considering the Basic Idea (equivalent to Weak Pareto), and several strengthenings of it. If we assume that locations do not come in an essential natural order, the strongest comparison V&K can make is the following:

> **Strengthened Basic Idea 1:** If w_1 and w_2 have exactly the same locations, and for any finite set of locations there is a finite expansion such that, for all further expansions, $w_1 \succ w_2$, then $w_1 \succ w_2$.

This theory results in only a partial ordering. Any worlds with a different relevant notion of 'distance' across locations are incomparable. And many world pairs, even where one is intuitively much better than another, will fail on the criterion that there exists some R_x where for *all* expansions r beyond it, the value inside r is greater.

But perhaps the bigger problem is that V&K's partial ordering is only *ordinal*. This theory, even if correct, tells us nothing about how to evaluate lotteries. In his reply to V&K, Bostrom (2011, 8) says that, in order to evaluate lotteries, such a method would first aggregate across locations of value, and then, after that, would account for uncertainty about which world will obtain. Arntzenius (2014) takes a different tack: he proposes that V&K theory be applied to *expected values* at locations, instead of the values themselves. The fact that the resulting ranking is only ordinal is not a problem, because in order to make decisions a utilitarian need only have an ordinal ranking over the expected values of different actions. The Arntzenius utilitarian can evaluate lotteries, like a 60% chance of world <2, 2, 2, … >, or world <1, 1, 1, … > for sure. This amounts to a choice between an 'expected world' of <1.2, 1.2, 1.2, … >, which dominates the 'expected world' of <1, 1, 1, … > by V&K theory. Because he is essentially applying the Generalized Metaprinciple, Arntzenius can compare worlds where the locations are not all the same, if there is an appropriate distance-preserving relation to tell us how the locations match between the worlds.

Arntzenius's proposal is clever, but there is a problem: Expansionism violates Strong Pareto. The proof of this is somewhat involved, but begins on page 83 of (Askell, 2018). Askell rejects Expansionism wholesale for this reason. But that is far from the only foundational objection. If spacetime regions define the essential natural order of locations, then Expansionism strongly favours worlds in which utility is more densely packed together. This is arguably a fundamentally finitary intuition misapplied to the infinite case (Askell, 2018, 210). For example, if Arntzenius uses a spatiotemporal ordering of locations, then his proposal endorses the addition of any finite number of dystopias to pull every planet in the universe 1 inch closer together (Carlsmith, 2022, 120). Merely moving value closer together does not seem to be of intrinsic moral significance, but by doing so at infinitely many locations, we will eventually get expected value dominance compared to the alternative world where value remains equally spaced apart, and any finite number of locations are made worse off.

Personally, I find the metaphysics of ethical comparisons being determined by dominance relations between hypothetical uniformly expanding spheres to be highly suspect. As typically formulated, the proposal relies on the spheres expanding at a uniform rate. There are many technical issues about how to define 'uniform' in this context; see (Arntzenius, 2014, 39). And Expansionism gives different views on cases in the style of Cain (1995)'s, depending on whether our imagined expanding sphere grows faster or slower than the Sphere of Happiness (Askell, 2018, 81). The Expansionists make strong claims, and their reward for doing so can still only compare quite a specific subset of worlds.

Kenny Easwaran (2021) has produced the first provably commutative version of Expansionism, in which we get the same result if we aggregate first across locations, and then across uncertainty (like Bostrom), or first across uncertainty and then across locations (like Arntzenius). The result can only compare worlds with the same locations, and, like the other methods, results only in a partial ordering (Easwaran, 2021, 299). While these ideas are certainly promising enough to be worthy of being developed further, Expansionism has not yet addressed nearly enough of the central issues to reasonably be considered the 'theory to beat' of infinite ethics.

7. Hyperreal numbers

Robinson (1966) showed how to construct a rigorous number system obeying the field axioms that contains:

- All the real numbers
- Infinitesimals (numbers that are positive, but smaller than any real number)
- Infinite numbers

This system later came to be known as the hyperreals. For infinite ethicists who assume that value is numerically representable, this offers an intriguing option. Because they form a field, all the standard arithmetical operations over these infinities and infinitesimals are well-behaved. And because that field is 'totally ordered', for any two elements, it is well-defined which one is bigger.

Bostrom (2011, 17)'s proposal was to associate infinite sequences of utility with specific infinite hyperreals. The hope is that, if every world is associated with a particular hyperreal, then we will both be able to rank outcomes, and also (since arithmetic is well-defined) be able to evaluate lotteries using the standard expected value framework.

The trouble is that defining a hyperreal field requires a completely arbitrary choice of 'non-principal ultrafilter'. This is actually an infinite set of arbitrary choices, in that for every subset of the 'index set' (usually $\mathbb{N}$), we need to choose whether to include that subset in the ultrafilter.

It doesn't matter for our purposes what an 'ultrafilter' actually is, but the upshot is this: ethical judgements about hyperreals require making arbitrary choices about an abstract mathematical object, which can determine, for example, whether a specific world is infinitely good or infinitely bad (Askell, 2018, 61). Various fixes have been proposed. The most straightforward amendment is to consider only the ethical rankings which remain constant under every possible choice of ultrafilter. Arntzenius (2014, 51) has shown that this collapses the hyperreal approach to become equivalent to V&K's Strengthened Basic Idea 1. And while Bostrom's proposal is ingenious, Arntzenius (2014, 52) argues that even in the best case, they wouldn't specify enough numerical relations to apply standard decision theory.[12]

[12] The first suggestion to apply hyperreals to infinite ethics came from V&K themselves (1997, 8). They rejected it because the study of hyperreals is silent in many cases where there are 'unboundedly' many numbers to be aggregated, as opposed to a specific infinite amount. Bostrom (2011, 17) seems to deny this distinction is relevant.

Discounting approaches are frequently criticised for advocating moral judgements which are based on an arbitrary choice of discount rate. But on the particular charge of arbitrarity, the hyperreals are surely much worse. The infinite consequentialist can at least discount in a way that has some basis in our moral intuitions, such as discounting welfare outside of his causal influence, or discounting welfare he has a low probability of influencing in a systematic direction, or simply discounting outside his communities. The choice of non-principal ultrafilter has no basis in our moral intuitions.

A different, and, I think, more promising application of hyperreals is to use them as part of a social welfare function. Pivato (2008) has proved that, unlike with $\mathbb{R}$, if the codomain is the field of hyperreals $^*\mathbb{R}$, then we *can* construct a Complete, Strongly Paretian, and Finitely Anonymous SWF. Pivato's approach has no time discounting, and can compare even utility streams that grow without bound. A hyperreal social welfare function is somewhat of a halfway house between a cardinal and an ordinal ranking: it can only compare across infinite, finite, and infinitesimal hyperreals in an ordinal fashion, but, within each category, it tells us how much an option is preferred. In principle, this gives us everything we need to evaluate lotteries.

As in Bostrom's method, whether one option is ranked higher by Pivato may depend on the particular choice of ultrafilter. However, he shows that this will only occur under quite specific circumstances (Pivato, 2008, 7). There is some promise to the approach of supplementing hyperreal social welfare functions with certain decision axioms that apply in the cases where the choice of ultrafilter affects the ranking. Although it is natural to be bothered by its

abstractness and ultrafilter-dependence, there is a narrow technical sense in which, using Pivato's method, the utilitarian is able to make comparisons between any infinite utility streams. I view this as the best alternative to discounting for utilitarians.

A view I have some sympathy with is that the order-dependency we've repeatedly met should be part of a richer conception of the nature of moral judgement. Our judgments might have some kind of 'canonical order' that locations of value come in. It seems to be an unavoidable feature of moral judgement that we value welfare at certain locations more than others, starting with our families, friends, and so on. That ordering might just be unavoidable in theory, as well as in practice. But it seems to me that the more mathematical approaches to infinite ethics come at the expense of allowing such an account to work. The choices about how to set up the hyperreals don't correspond to anything we ordinarily consider normative.[13]

8. Conclusion

Many moral theories, not only consequentialist, face an infinite aggregation problem. It's not possible to extend two minimal comparative principles, Strong Anonymity and Strong Pareto, into the infinite domain.

[13] A related proposal is to instead use Conway's (1976) 'surreal' numbers, which provably form the largest totally ordered field. The surreals avoid the problem of ultrafilter dependence – but other than that, they run into similar problems (Askell, 2018, 64).

Intergenerational welfare economics is relevant on two fronts. First, economists have been dealing with the problem of aggregating infinite value for almost a century, and have widely responded to it with discounting. Second, impossibility theorems place harsh limits on what kinds of infinite utility streams can possibly be compared, subject to desirable axioms. However, if we allow for discounting, the latest research suggests that we are able to compare almost all realistic utility streams. Economists have made much progress on solving the problem of infinite value aggregation, although it's not entirely clear to me whether philosophers have noticed.

Unlike the alternatives, the discount methods do not directly invoke the axiom of choice, and they can be explicitly constructed.

The first of the Expansionist proposals, from V&K, produced only a partial ordinal ranking. Further work is still to be done on Bostrom's suggestion to account for uncertainty across a V&K ranking. Arntzenius (2014)'s proposal to instead consider dominance relations between worlds where we consider expected value at locations is not even commutative. It matters whether we first aggregate across locations, and then uncertainty, or the other way around. The first provably commutative version of Expansionism, from Easwaran (2021), is, like the others, totally unable to compare many realistic worlds. If spatiotemporal regions define the essential natural ordering of locations, these methods place great importance on how densely we pack value together. I am not even necessarily opposed to that intuition, but that is a separate issue.

Prima facie, the hyperreals are totally inappropriate for this problem. The free choice of non-principal ultrafilter is at *least* as arbitrary as the choice of discount rate, and likely more so. Hyperreals are more promising in the context of social welfare functions, where they have finally allowed an infinite utilitarian to have a Complete ranking over outcomes. The result is so abstract that it's not clear how much comfort she should take.

It's a question for other work to what extent the adoption of a discount rate itself undermines the case for aggregative consequentialism. At least since the 1990s, scholars have wondered about whether, in infinite worlds, utilitarianism is the snake that bites its own tail – whether the use of a social discount rate erodes the theoretical elegance for which the theory was adopted in the first place.[14] I find it unlikely that discounting features in the 'true' moral theory, or as close as we can get to it, which is one reason why I am not personally a consequentialist.

I agree with the critics that the positive arguments for discounting are weak. But it compares favourably to the alternatives which have been proposed. While I was not exhaustive, the only major contender which has been developed that I did not discuss is the 'value-density' or 'averaging' approach, which is similar to Expansionism (Carlsmith, 2022, 115) (Bostrom, 2011, 16) (Askell, 2018, 53).

[14] This was the claim of Nelson (1991), the third time that the problem of infinite value aggregation was independently discovered, after Ramsey (1928) and Segerberg (1976). Nelson claimed that a utilitarian would either be totally unable to make decisions (like in 'infinitarian paralysis'), or would have to discount, which with an infinite horizon would inexorably lead to an 'eschatology' (like the 'ethics-free zone').

Derek Parfit spent much of his career in search of a ‘Theory X’ to resolve the paradoxes that arise when ethically comparing different populations. This effort culminated in the publication of ‘The Impossibility of a Satisfactory Theory of Population Ethics’, a proof that certain widely desired characteristics of such a theory are not jointly satisfiable (Arrhenius, 2011). I sense that infinite ethics may be nearing a similar synthesis. I will not hold my breath that UDASSA or the hyperreals will prove to be the Theory X for infinite ethics. But, in the course of exploring them, philosophers might demonstrate limitative results, about what properties such a theory can and cannot have.

Bibliography

Arntzenius, F. (2014). Utilitarianism, Decision Theory and Eternity. *Philosophical Perspectives, 28*(1), 31-58.

Arrhenius, G. (2011). The Impossibility of a Satisfactory Population Ethics. In *Descriptive and Normative Approaches to Human Behavior* (pp. 1–26). World Scientific.

Arrow, K. J. (2012 [1951]). *Social Choice and Individual Values: Third Edition.* Yale University Press.

Askell, A. (2018). Pareto Principles in Infinite Ethics. Retrieved from PhilArchive: https://philarchive.org/rec/ASKPPI

Basu, K., & Mitra, T. (2003). Aggregating Infinite Utility Streams with Intergenerational Equity: The Impossibility of Being Paretian. *Econometrica, 71*, 1557–1563.

Basu, K., & Mitra, T. (2007, March). Utilitarianism for infinite utility streams: A new welfare criterion and its axiomatic characterization. *Journal of Economic theory, 133*(1), 350–373.

Bostrom, N. (2009). Pascal's Mugging. *Analysis, 69*(3), 443-445.

Bostrom, N. (2011). Infinite Ethics. *Analysis and Metaphysics, 10*, 9-59.

Cain, J. (1995). Infinite Utility. *Australasian Journal of Philosophy, 73*(3), 401-404.

Carlsmith, J. (2022). A Stranger Priority? Topics at the Outer Reaches of Effective Altruism. Retrieved from https://joecarlsmith.com/2023/02/21/a-stranger-priority-topics-at-the-outer-reaches-of-effective-altruism

Carroll, S. (2020). *111 | Nick Bostrom on Anthropic Selection and Living in a Simulation.* Preposterous Universe. Retrieved from https://www.preposterousuniverse.com/podcast/2020/08/24/111-nick-bostrom-on-anthropic-selection-and-living-in-a-simulation/

Conway, J. H. (2001 [1976]). *On numbers and games.* Taylor & Francis.

Cowen, T. (1992). Consequentialism Implies a Zero Rate of Intergenerational Discounting. In *Philosophy, Politics, and Society* (pp. 162–168). Yale University Press.

Cowen, T. (2018). *Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals.* Stripe Press.

Diamond, P. (1965). The Evaluation of Infinite Utility Streams. *Econometrica, 33*(1), 170–177.

Dubey, R. S. (2011). Fleurbaey-Michel conjecture on equitable weak paretian social welfare order. *Journal of Mathematical Economics, 47*(4–5), 434–439.

Easwaran, K. (2014). Decision theory without representation theorems. *Philosophers' Imprint, 14*(27), 1–30.

Easwaran, K. (2021). A New Method of Value Aggregation. *Proceedings of the Aristotelian Society*, 299–326.

Hájek, A. (2024). *Pascal's Wager.* Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/sum2024/entries/pascal-wager/

Jonsson, A. (2023). An axiomatic approach to Markov decision processes. *Mathematical Methods of Operations Research, 97*, 117–133.

Jonsson, A., & Voorneveld, M. (2018). The limit of discounted utilitarianism. *Theoretical Economics, 13*(1), 19–37.

Koopmans, T. (1960). Stationary Ordinal Utility and Impatience. *Econometrica, 28*(2), 287–309.

Koopmans, T. (1965). On the Concept of Optimal Economic Growth. In *The Econometric Approach to Development Planning* (Vol. 28, pp. 225–87). Pontificiae Academiae Scientiarium Scripta Varia.

Koopmans, T. C. (1972). Representation of Preference Orderings Over time. *Decision and Organization, 57*(100).

Lauwers, L. (1995). Time-neutrality and linearity. *Journal of Mathematical Economics, 24*(4), 347–351.

McLaughlin, P. (2022a, October 7). *Getting on a different train: can Effective Altruism avoid collapsing into absurdity?* Retrieved from Effective Altruism Forum: https://forum.effectivealtruism.org/posts/8wWYmHsnqPvQEnapu/getting-on-a-different-train-can-effective-altruism-avoid

McLaughlin, P. (2022b). The Development of the Concept of Existential Risk. University of Cambridge Faculty of History.

Monton, B. (2019, June). How to Avoid Maximizing Expected Utility. *Philosophers' Imprint, 19*(18), 1–25.

Moore, A. W. (2018). *The Infinite.* Routledge.

Moore, G. E. (2004 [1903]). *Principia Ethica.* Dover Publications.

Mueller, M. P. (2017, December 5). *[1712.01826] Law without law: from observer states to physics via algorithmic information theory.* Retrieved April 12, 2025, from arXiv: https://arxiv.org/abs/1712.01826

Nelson, M. T. (1991). Utilitarian eschatology. *American Philosophical Quarterly, 28*(4), 339–347.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity.* Bloomsbury.

Parfit, D., & Cowen, T. (1992). Against the Social Discount Rate. In *Philosophy, Politics and Society : Second Series: A Collection* (pp. 144–61). Blackwell.

Pivato, M. (2008). Sustainable preferences via nondiscounted, hyperreal intergenerational welfare functions. *Munich Personal RePEc Archive*. Retrieved from https://mpra.ub.uni-muenchen.de/7461/

Ramsey, F. (1928). A Mathematical Theory of Saving. *The Economic Journal, 38*(152), 543-559.

Robinson, A. (1996 [1966]). *Non-standard analysis.* Princeton University Press.

Segerberg, K. (1976). A neglected family of aggregation problems in ethics. *Noûs, 10*(2), 221–244.

Singer, P. (1972). Famine, Affluence, and Morality. *Philosophy and Public Affairs, 1*(3), 229-43.

Tarsney, C. (2020, August). Exceeding expectations: stochastic dominance as a general decision theory. *Global Priorities Institute Working Papers Series, 3*. Retrieved from https://globalprioritiesinstitute.org/wp-content/uploads/Christian-Tarsney_Exceeding-expectations-stochastic-dominance-as-a-general-decision-theory.pdf

Vallentyne, P., & Kagan, S. (1997). Infinite Value and Finitely Additive Value Theory. *The Journal of Philosophy, 94*(1), 5-26.

Van Liedekerke, L., & Lauwers, L. (1997). Sacrificing the Patrol: Utilitarianism, Future Generations and Infinity. *Economics & Philosophy, 13*(2), 159 - 174.

von Neumann, J., & Morgenstern, O. (2007 [1944]). *Theory of Games and Economic Behavior: 60th Anniversary Commemorative Edition.* Princeton University Press.

West, B. (2017). *Big Advance in Infinite Ethics.* Retrieved from Philosophy for Programmers: https://web.archive.org/web/20200811134155mp_/http://philosophyforprogrammers.blogspot.com/2017/09/big-advance-in-infinite-ethics.html

Wiblin, R., Harris, K., & Karnofsky, H. (2023, July 31). *Holden Karnofsky on how AIs might take over even if they're no smarter than humans, and his 4-part playbook for AI risk.* Retrieved April 6, 2025, from 80000 Hours: https://80000hours.org/podcast/episodes/holden-karnofsky-how-ai-could-take-over-the-world/

Zame, W. R. (2007). Can intergenerational equity be operationalized? *Theoretical Economics, 2*(2), 187–202.