Infinite ethics and the utilitarian dream

Joe Carlsmith

September 2022

I Introduction

Most of ethics ignores infinities. They're confusing. They cause problems. Hopefully, they're irrelevant. And anyway, finite ethics is hard enough.

Infinite ethics is just ethics without these blinders. And taking off the blinders—at least sometimes—is good.[1] Infinities are a live issue in practice. And the problems they create are deeply revealing in theory.

This essay surveys some of these problems and reflects on their implications. I begin by briefly noting two prominent problems in the literature that seem to me less pressing: namely, problems about indifference to merely finite amounts of influence on an infinite world, and problems about "infinity-fanaticism"—e.g., tiny probabilities of infinite influence swamping all finite concerns. There are indeed worries here, but I also think that (especially for a certain sort of bullet-biting utilitarian), there are lines of response available that don't require substantial deviation from the sorts of principles we might've wanted to uphold in finite contexts.

I then turn to problems that offer no such comforts. I begin with some of the impossibility results in infinite ethics, which show that even in the context of merely ordinal rankings over infinite worlds (that is, rankings that tell you which worlds are better/worse relative to which others, but which don't say *how much* better/worse), a number of very plausible principles are incompatible with each other. And these principles, on their own, are far too weak to provide a full ordinal ranking regardless.

Even if we had a full ordinal ranking over infinite outcomes, though, that wouldn't solve the even more difficult problem of choosing between *lotteries* over such outcomes. I discuss various candidate solutions to this problem (notably, appeals to totals, discounts, averages, hyperreals, spatiotemporal expansions, and to what I call the "four-types view"). I suggest that none of these proposals are plausible, and that some are truly horrifying.

What's more, all of these proposals focus on a limited domain: namely, countable infinities. But there are much larger infinities as well. If our ethics has to deal with *those*, they seem likely to break whatever principles

[1] Though putting them on, for the purposes of simplifying a given discussion, may often be useful as well.

we settle on for the countable case.

With these problems in view, I turn to their implications. In particular, I argue that infinite ethics punctures the dream of a simple, bullet-biting utilitarianism; and it puts pressure on some of the broader intuitions underlying common arguments for "strong longtermism"—that is, the view that positively influencing the longterm future is the key moral priority of our time.[2] I also briefly touch on whether these problems constitute an argument against moral realism (my answer: maybe).

I close with a discussion of the practical implications of taking infinite ethics seriously, especially in the absence of solutions to the problems I've discussed. My guess is that beyond simply doing more research, and rather than looking for specific infinity-oriented interventions now, people who take infinite ethics seriously should work to make sure that our civilization reaches a wise and technologically mature future—one of superior theoretical and empirical understanding, and superior ability to put that understanding into practice. But reflection on infinite ethics can also inform our sense of how strange such a future's ethical priorities might be.

II The importance of the infinite

Why are infinities important to ethics? I see two main reasons: (1) we have to deal with them in practice, and (2) they're deeply revealing in theory.

Why do we have to deal with infinities in practice? For one thing, it's possible that we live in an infinite world.[3] But more importantly, it's possible that our actions, now, can influence what happens to an infinite number of value-bearing locations—for example, people. This could happen in two ways: causal, or acausal.

The causal way requires stranger science. It's not that infinite *universes* are strange: to the contrary, the hypothesis that we share the universe with

[2]See Greaves and MacAskill (2021); and see MacAskill (2022) for a more popular introduction.

[3]See Sean Carroll's comments on his and Bostrom's (20200 podcast (timestamp 13:01): "Just so everyone knows, this is an open question in cosmology. ... The possibility's on the table, the universe is infinite, there's an infinite number of observers of all different kinds, and there's a possibility on the table that the universe is finite, and there's not that many observers, we just don't know right now." See also the citations in Bostrom (2011): "Recent cosmological evidence suggests that the world is probably infinite..." (p. 2); Askell (2018), section 1.1; and Wilkinson (2021a): "you might be disappointed to find that the world around you is infinite in the relevant sense. I am sorry to disappoint you, but contemporary physics suggests just that. The widely accepted flat-lambda model predicts that our universe will tend towards a stable state and will then remain in that state for infinite duration ... Take any small-scale phenomenon which is morally valuable e.g., perhaps a human brain experiencing the thrill of reading philosophy for a given duration. Each of the above physical views predicts that our universe, in its infinite volume, will contain infinitely many such thrills" (p. 1919).

an infinite number of observers is very live. But current science suggests that our *causal* influence is made finite by things like lightspeed and entropy.[4] So exerting infinite causal influence probably needs new science. Maybe we learn to make computers that perform infinite amounts of computation,[5] or baby universes with infinite space-times.[6] Maybe we're in a simulation housed in a universe more friendly to infinite causal influence. Or maybe something else that we've never considered makes infinite causal influence possible.

The acausal way is compatible with more mainstream science. But it requires stranger decision theory. Suppose you're deciding whether to make a $5000 donation that will save a life, or to spend the money on a vacation with your family. And suppose, per various respectable cosmologies, that the universe is filled with an infinite number of people very much like you, faced with choices very much like yours. If you donate, this is strong evidence that they all donate, too. So evidential decision theory treats your donation as saving an infinite number of lives, and as sacrificing an infinite number of family vacations.[7] The stakes are high.

One response here is to reject unconventional science *and* unconventional decision theory. But very plausibly, you should at least have non-zero credence on them both.[8] EDT, after all, is a reasonably mainstream view in decision theory; and are you really *certain* that e.g. all religions that allow your actions influence over infinite afterlives are false? And it seems very possible to update, later on, towards the view that infinite causal influence is in fact possible (God, for example, could appear before you and offer you the chance to create a new infinite universe), or that evidential decision theory is correct (you could hear, for example, that there is a new, knock-down argument for it that all the leading philosophers have been persuaded by)—a possibility that suggests you shouldn't rule them out with certainty, now. But non-zero credence is enough to get many of the

[4] I'm ignoring situations where e.g. if I eat a sandwich today, then this changes what happens later to an infinite number of brains that randomly fluctuate out of a thermal bath ("Boltzmann Brains"—see Carroll (2017)), but such changes occur in a manner I can't ever predict. That said, this sort of scenario does raise problems: see Wilkinson (2021b) for some discussion.

[5] See Ord (2002) for more on hypercomputers, and Dyson (1979, p. 455-456) for discussion of physical possibilities for infinite computation.

[6] See Guth (1996).

[7] Exactly what it treats your decision as doing *overall* to the distribution of utility, given all the correlations across the infinite universe, is a further question—but the point that you are able to exert infinite (evidential) influence still stands. Other non-causal decision theories, like the "Functional Decision Theory" of Soares and Levinstein (2020), will plausibly behave in a similar way.

[8] Another possible response is to try to reject the idea that infinity is a coherent idea at all—but I find this quite hard to square with its role in live scientific cosmological hypotheses like the ones discussed above. As Russell (2022) puts it: drawing this conclusion, at least in response to the sorts of considerations discussed in this paper, would be a "striking bit of armchair physics" (timestamp 44:45).

relevant problems going—at least if you want to incorporate this credence into your decision-making.

Even if you insist on ignoring infinities in practice, though, they still matter in theory. In particular: whatever our world's (or our influence's) actual finitude, it seems strange if ethics falls silent in the face of the infinite—especially if we want to honor the idea (often assumed in normative ethics, but not unquestionable) that ethical truths hold in all possible worlds. Infinite worlds, after all, seem eminently possible—indeed, as just discussed, their *actuality* is a live scientific hypothesis. And forms of infinite causal influence seem possible as well (imagine, for example, choosing which of two infinite universes to create). What's more, we typically want our ethical principles to extend from the actual not just to other worlds physically similar to our own (e.g., strange trolley problems), but also to worlds involving other physical laws, too (e.g., trolley problems in which gravity works in a different way). It's not clear why infinite worlds (or worlds that allow infinitely consequential action) would be an exception.[9]

What's more, we have intuitions about infinity-involving choices. Suppose you're God, choosing whether to create an infinite heaven, or an infinite hell. Should you flip a coin? Definitely not. So that's one intuitive datapoint—and we have many others, which I'll draw on below. So naively, we have the makings of the familiar game of normative ethics, in which we attempt to identify general principles that fit, explain, and/or revise our intuitions about particular cases.

Except: infinities make this game much harder. Indeed, they break a lot of theories developed with only the finite in mind. This can be painful, but it's also instructive. In science, one often *hopes* to get new data that ruins an established theory. It's a route to progress: breaking the breakable is often key to fixing it.

So, on both practical and theoretical grounds, we need to grapple with infinite ethics. Let's look at what happens when we do.

III Locations of value

A few quick notes of set-up.

The standard approach to infinite ethics involves putting finite utilities on an infinite set (specifically, a countably infinite set) of value-bearing "locations." But it can make an important difference what sort of locations you have in mind.

Here's a classic example, adapted from Cain (1995). Consider two worlds:

[9] And indeed, it seems relatively clear to me that if infinities *didn't* create serious problems for normative ethics, we wouldn't be interested in excluding them from its domain of applicability.

ZONE OF SUFFERING: An infinite line of immortal people, numbered starting at 1, who all start out happy (+1). On day 1, person 1 becomes sad (-1), and stays that way forever. On day 2, person 2 becomes sad, and stays that way forever. And so on.

Person	1	2	3	4	5	
day 1:	⟨ -1,	1,	1,	1,	1,	...⟩
day 2:	⟨ -1,	-1,	1,	1,	1,	...⟩
day 3:	⟨ -1,	-1,	-1,	1,	1,	...⟩
etc...						

ZONE OF HAPPINESS: The same world, but the happiness and sadness are reversed: everyone starts out sad, and on day 1, person 1 becomes happy; day 2, person 2, and so on.

Person	1	2	3	4	5	
day 1:	⟨ 1,	-1,	-1,	-1,	-1,	...⟩
day 2:	⟨ 1,	1,	-1,	-1,	-1,	...⟩
day 3:	⟨ 1,	1,	1,	-1,	-1,	...⟩
etc...						

In zone of suffering, at any given *time*, the world has finite sadness, and infinite happiness. But any given *person* is finitely happy, and infinitely sad. In zone of happiness, it's reversed. Which is better?

My view is that the zone of happiness is better. It's where I'd rather live, and choosing it over zone of suffering fits with principles like "if you can save everyone from infinite suffering and give them infinite happiness instead, do it," which seem pretty solid to me. Of course, analogous principles for times also have appeal, but from a moral perspective, agents seem to me more fundamental.

My broader point, though, is that the choice of location matters. Unless otherwise stated, I'll focus on agents.

Also, to simplify the discussion, I'm generally going to be assuming that non-axiological considerations aren't playing an important role in the choices I discuss, and that we can move freely between talk about which worlds and options are "better," and which are more choiceworthy. And I'll often assume some relatively simple theory of welfare—e.g., hedonism—as well.

Finally: when I talk about "infinite worlds" or "infinite actions," I'll be speaking by default about worlds with an infinite number of value-bearing locations (with non-zero value), and about actions that affect the value at an infinite number of such locations. That is, I'm ignoring worlds with e.g. an infinite space-time, but only a finite number of people (assuming that people are the relevant locations of value), and actions that e.g. affect an infinite amount of space-time, but only a finite number of people.

IV Can finite influence matter in infinite worlds?

With this set-up in mind, let's start with some problems from the literature that I view as comparatively (if not absolutely) easy: namely, problems to do with infinities swamping merely finite concerns. I'll focus on two types: worries about finite actions not mattering in infinite worlds (this section), and worries about finite actions not mattering relative to arbitrarily tiny probabilities of infinite actions (next section).

A prominent proponent of the first type is Bostrom (2011), who focuses on the concern that actions with merely finite amounts of influence can't change the overall value of worlds containing infinite amounts of value and disvalue. So if your ethics is about changing the value of the world, living in such an infinite world would leave you indifferent to any finitely-influential action, however worthy or horrible.

Thus, for example, in a world of infinite people at 1, bumping any finite number up or down any amount leaves the total welfare—∞—unperturbed.[10] So naïve total utilitarians in such a world start shrugging at genocides. The same holds for prioritarians who first weight people's well-being according to how badly off they are, and *then* take the sum; and plausibly, average utilitarians are going to have similar troubles.[11] Bostrom calls this "infinitarian paralysis," and he treats it as a devastating problem.[12]

But I don't see it that way. In particular, in light of issues like this, it seems to me quite natural for the total utilitarian, at least, to refocus her ethical attention on the welfare she *adds* or *subtracts* from a world, rather than on the value of the world overall. Thus, in a world of infinite 1s, bumping 10 people up to 2 adds 10 units of welfare. So it's worth doing, even if the total welfare is unaffected. Prioritarians have similar options (e.g., weighting welfare additions/subtractions that affect the worse-off more strongly); and perhaps average utilitarians do as well.[13] Indeed, this sort of move seems to me a simple extension of a certain type of "size of drop, not size of bucket" reasoning often used in finite contexts.[14]

Bostrom considers a response of this type, but he argues that it faces two

[10] If you say that the total welfare is undefined, you get similar issues.

[11] Though we'd need to say more about how we're doing the averaging; see the section on average views below for more.

[12] Bostrom (2011): "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 in light of the theory's compensating virtues. The problem of infinitarian paralysis must be solved, or else aggregative consequentialism must be rejected." (p. 45).

[13] I haven't tried to work through the details here, partly because I see average utilitarianism as independently implausible.

[14] See MacAskill (2015): "It's not the size of the bucket that matters, but the size of the drop" (p. 25).

problems.[15] The first is that it requires giving up on the aspiration to change the overall value of the world—an aspiration Bostrom sees as core to the spirit of consequentialist ethics.[16] But this doesn't worry me much: I'm fine with merely aspiring to help people (rather than to change the value of the world), and I see this as a reasonably core aspiration in its own right (whether "consequentialist" or no).

Bostrom's second worry is that we might be able to take infinitely influential actions, and that this approach doesn't tell us how to choose between such actions (or even: small probabilities of them). Here I agree with the concern. But I see "how do we choose between different infinitely influential actions?"—particularly in the context of risk—as a separate and more fundamental problem than "how can finite changes matter in infinite worlds?" It's true that this solution to the latter doesn't solve the former. But the latter was (extremely) unsolved anyway (more below), and this solution to the former doesn't make it worse.[17] So Bostrom's focus on the former problem, in his paper, seems to me misplaced. The latter is the hard part.

V Infinite fanaticism

Before turning to this hard part, though, I want to touch on a different classic worry about the infinite swamping the finite: namely, fanaticism about infinite outcomes.

Fanaticism, in ethics, means paying extreme costs with certainty, but the sake of tiny probabilities of sufficiently high-stakes outcomes.[18] Thus, to take an infinite case: suppose that you live in a finite world, and everyone is miserable. You are given a one-time opportunity to choose between two

[15] See section 3.2 on "the causal approach."

[16] Bostrom (2011): "One consequence of the causal approach is that there are cases in which you ought to do something, and ought to not do something else, even though you are certain that neither action would have any effect at all on the total value of the world… The implication that you ought to 'do good' even when doing so does not make the world better must, from the standpoint of the aggregative consequentialist, be regarded as a liability of the causal approach" (p. 26).

[17] For example, it's not as if totalists have a great theory for comparing infinitely influential actions (see section on totalism below for more on these issues), but they run into problems when they start thinking about finite genocides in infinite worlds, such that switching focus to the welfare you add/subtract solves the finite genocide issue, but creates some new problem that totalism didn't already have.

[18] Wilkinson (2021) defines fanaticism, more generally, as: "*Fanaticism*: For any tiny (finite) probability $e > 0$, and for any finite value v, there is some large enough finite V such that L_{risky} is better than L_{safe} (no matter which scale those cardinal values are represented on).

L_{risky}: value V with probability e; value 0 otherwise

L_{safe}: value v with probability 1" (p. 5).

buttons. The blue button is guaranteed to transform your world into a giant (but still finite) utopia that will last for trillions of years. The red button has a one-in-a-graham's-number chance of creating a utopia that will last *infinitely* long. Which should you press?

Here the fanatic says: red. And naively, if an infinite utopia is infinitely valuable, then expected utility theory agrees: the EV of red is infinite (and positive), and the EV of blue, merely finite. But one might wonder. In particular: red seems like a loser's game. You can press red over and over for a trillion$^{\text{trillion}}$ years, and you're still basically guaranteed to not win the infinite prize you seek. Is that really what rationality looks like?

This isn't a purely infinity problem. Verdicts like "red" are hard to avoid, even for merely finite outcomes, without saying other very unattractive things.[19] Plausibly, though, the infinite version is worse. The finite fanatic, at least, cares about how tiny the probability is, and about the finite costs of rolling the dice. But the infinite fanatic has no need for such details: she pays *any* finite cost for *any* probability of an infinite payoff. Suppose that actually, I overestimated the probability of red paying out by a factor of a graham's number; and actually, red also kills a trillion children with certainty. The infinite fanatic doesn't blink. The moment you said "infinity," she tuned out finite considerations like those.

What's more, the finite fanatic can reach for excuses that the infinite fanatic cannot. In particular, the finite fanatic can argue that, in her actual situation, she faces no choices with the relevantly problematic combination of payoffs and probabilities. Whether this argument works is another question (I'm skeptical). But the infinite fanatic has trouble even voicing it.[20] After all, *any* non-zero credence on an infinite payoff is enough to bite her. And non-zero credences seem hard to avoid (to take a classic example: even if you're a confident atheist, can you really be *certain* that Catholicism is false, especially given how many people believe in it?).[21] Thus, no matter where she is, no matter what she has seen, the infinite fanatic never gives finite things any intrinsic attention.[22] When she kisses her chil-

[19]See Beckstead and Thomas (2021) and Wilkinson (2021) for discussion. That said, as Beckstead and Thomas (2021) discuss, fanaticism (even in purely finite cases) also leads to other problems (beyond basically counterintuitive verdicts like "red"): for example, violations of principles like prospect-outcome dominance. See also Russell (2021) for more.

[20]See Bostrom (2011), p. 32.

[21]Some views in epistemology may be able to accommodate certainty of this kind, but I won't delve into that possibility here. There are also views that discount or ignore sufficiently small probabilities in making decisions (see e.g. Smith (2014) and Monton (2019)). But as Beckstead and Thomas (2021), section 2.3, discuss, these views face very serious problems.

[22]Here I'm setting aside responses on which infinity-related considerations will always exactly balance out, such that you end up choosing on the basis of finite considerations despite your infinity-fanaticism. I find this very implausible: for example, to me it seems

dren, or prevents a genocide, she does it for some infinite prize, however improbable.[23]

So if we countenance infinitely valuable (or disvaluable) outcomes, then standard expected value theory leads to an especially unappealing form of fanaticism—an issue familiar from (though more general than) Pascal's Wager.[24] And indeed, weaker assumptions can lead to fanaticism of this kind as well. Thus, Beckstead and Thomas (2021, p. 26) show that a form of fanaticism about *finite* outcomes—what they call "recklessness"—leads to infinite fanaticism in conjunction with the assumption that an infinite payoff is *better* than any finite one, without ever appealing to expected value theory. And in many cases, such an assumption seems quite plausible. If saving more lives, for example, is better than saving fewer, then it's very natural to say that saving infinite lives is better than saving any finite number of lives x, since saving infinite lives saves x lives, *and also infinitely more.*

I grant that infinite fanaticism is a problem: it seems strange to ignore the finite entirely, and to obsess only about tiny probabilities of the infinite. But I don't think this is the biggest problem infinities create for ethics. For one thing, infinite fanaticism is similar to finite fanaticism in many ways, and it seems reasonable to expect a similar resolution—if a resolution is to be found. And indeed, some salient (though not, in my opinion, satisfying) ways of avoiding finite fanaticism (bounded utility functions, discounting sufficiently small probabilities) help with the infinite version, too.[25]

What's more, though, one salient response to finite fanaticism is to just bite the bullet, at least in sufficiently thought-experimental cases;[26] and this response can be applied to infinite fanaticism as well (indeed, as just noted, biting the finite bullet leads quickly to biting the infinite one, too). And biting has a familiar logic. After all, infinities of pleasure and pain really are *extremely* high stakes (higher stakes, very plausibly, than any finite amount of pleasure or pain). Indeed, there is a grand tradition of treating things like heaven and hell as lexically more important than the ephemera of this fallen world. Perhaps, then, we could live with obsession, if we had to. And in particular, I think, those who have reconciled themselves to biting the bullet on finite fanaticism can bite it here, too, without any

---

like there is some substantive and distinctive set of actions I could take in order to maximize my probability of going to one of the heavens (and avoiding the hells) posited by the various world religions, and I think I should have higher credence on those religions than on other made-up religions designed to specifically balance out the verdicts of the ones I see around me. See Bostrom (2011 p. 33) for more on this.

[23]See Beckstead and Thomas (2021), p. 31, for more on the revisionary-ness of this implication, even if the actual actions one performs for the sake of one's infinite fanaticism aren't so strange.

[24]See Hajek (2017) for an overview.

[25]See Beckstead and Thomas (2021) for discussion of these options.

[26]See Wilkinson (2021c) for advocacy.

substantial deviation from the commitments they held in finite contexts. It's a worse bullet, yes—but it's the same *type* of bullet, and one bites on broadly similar grounds.

The biggest problems for infinite ethics, in my opinion, are harder than this. In particular, I think, the biggest problems have to do with comparing infinities to other infinities (especially in the context of risk), rather than comparing the infinite to the finite. Let's turn to these now.

VI The impossibility of what we want

Whether you're obsessed with infinities or not, you might hope to be able to choose between them. After all, as noted above, ethics does not fall immediately silent in the face of the infinite. Heaven is better than Hell. An infinite Utopia is better than a single, immortal, barely-conscious, slightly-happy lizard, floating forever in space (or at least, I think so).

Can we identify plausible principles for making such comparisons? Let's start with an easy version of the task: namely, principles that could help generate a purely ordinal ranking of infinite worlds (that is, a ranking that tells us which worlds are better than which others, but which doesn't tell us *how much* better).

Consider the following very plausible principle:

> Infinite Agent-based Pareto: If two worlds (w1 and w2) contain the same people, and w1 is better for an infinite number of them, and at least as good for all of them, then w1 is better than w2.[27]

Infinite Agent-Based Pareto looks very good. But it immediately leads to problems. In particular, in infinite cases, it conflicts with:

> Agent-based Anonymity: If there is a welfare-preserving bijection from the agents in w1 to the agents in w2, then w1 and w2 are equally good.

By "welfare-preserving bijection," I mean a mapping that pairs each agent in w1 with a single agent in w2, and each agent in w2 with a single agent in w1, such that both members of each pair have the same welfare level. (The intuitive idea here is that we don't care more about some agents than others—at least not without good reason.[28] A world where Alice has 1, and Bob has 2, has the same value as a world where Alice has 2, and Bob has 1.)

[27] We can imagine many variants of this Infinite Agent-Based Pareto principle, including more comprehensive, finite variants that only require w1 to be better for *some* people (whether a finite number or an infinite number). I find the version in the infinite text especially plausible, though.

[28] What if Alice and Bob differ in some intuitively relevant respect, like the degree to which they *deserve* happiness vs. suffering? Following common practice, I'm ignoring such differences; but if you like, feel free to add further conditions like "provided that everyone is similar in XYZ respects."

To see the conflict between Infinite Agent-Based Pareto and Agent-Based Anonymity, consider the following example.[29] In w1, every fourth agent has a good life. In w2, every second agent has a good life. And the same agents exist in both worlds.

Agents	a1	a2	a3	a4	a5	a6	a7	
w1	1	0	0	0	1	0	0	...
w2	1	0	1	0	1	0	1	...

By Infinite Agent-Based Pareto, w2 is better than w1 (it's better for a_3, a_7, and so on, and just as good for everyone else). But there is also a welfare-preserving bijection from w1 to w2: you just map the 1s in w1 to the 1s in w2, in order, and the same for the 0s.[30] So by Agent-Based Anonymity, w1 and w2 are equally good. Contradiction.

Here's another example.[31] Consider an infinite world where each agent is paired with an integer, in a bijection, and where the integer in question determines the agent's welfare, such that each agent i is at i welfare. And now suppose you could give each agent in this world +1 welfare. Would this make the world better? By Infinite Agent-Based Pareto, yes. But by Agent-Based Anonymity: no. After all, there's a welfare preserving bijection from each agent i in the first world to agent $i - 1$ in the second:

Agents	...	a-3	a-2	a-1	a0	a1	a2	a3	...
w3	...	-3	-2	-1	0	1	2	3	...
w4	...	-2	-1	0	1	2	3	4	...

Indeed, Agent-Based Anonymity mandates indifference to the addition or subtraction of any uniform level of well-being in w3 (e.g., harming each agent by a million, or helping them by a trillion).

Clearly, then, we have to reject at least one of Infinite Agent-Based Pareto and Agent-Based Anonymity. Which should we choose?

I'm inclined to reject Agent-Based Anonymity. Infinite Agent-Based Pareto seems the more intuitively plausible principle to me, and Agent-Based Anonymity causes problems for other attractive principles as well. Consider, for example:

> Anti-infinite-sadism: Adding infinitely many suffering agents to a world makes it worse.

This principle seems extremely plausible to me. And it seems plausible even if we say that X's life of suffering is not *worse for X* than non-existence (such that adding suffering agents does not violate Infinite Agent-Based Pareto).

[29] This example is adapted from Van Liedekerke (1995).

[30] Thus: a_1 goes to a_1, a_2 goes to a_2, a_3 goes to a_4, a_4 goes to a_6, a_5 goes to a_3, and so on.

[31] This example is adapted from Hamkins and Montero (1999).

But now consider an infinite world where everyone is at -1. And suppose you can add another infinity of people at -1.

Agents	a1	a2	a3	a4	a5	a6	a7	...
w5	-1		-1		-1		-1	...
w6	-1	-1	-1	-1	-1	-1	-1	...

Agent-Based Anonymity is indifferent to this change, despite the fact that it creates an infinite number of suffering people, and changes nothing else. This seems to me a horrible conclusion.

That said, rejecting Agent-Based Anonymity is not easy: the principle has strong appeal. In particular: it can quickly start to seem like pairs of worlds like w3 and w4, and w5 and w6, really *are* ethically similar in the way that Agent-Based Anonymity assumes.

Here's a way of pumping the intuition. Consider a world just like w3/w4, except with an entirely different set of people (call them the "b-people").

Agents	...	b-3	b-2	b-1	b0	b1	b2	b3	...
w7	...	-3	-2	-1	0	1	2	3	...

Compared to w3, w7 looks equally good: switching from a-people to b-people doesn't change the value. But so, too, does w7 look equally good when compared to w4 (it doesn't matter which b-person we call b_0). But by Infinite Agent-Based Pareto, it can't be both. And we can pump the same sort of intuition with w5, w6, and another infinite b-people world consisting of all -1s (call this w8). This isn't to say there's no way to hold on to Infinite Agent-Based Pareto in the face of such cases (we could, for example, say that w3 and w4 are both incomparable to w7).[32] But letting go of Agent-Based Anonymity has strong intuitive costs.

We get the same conflict between Something-Based Pareto and Something-Based Anonymity if we focus on basic locations of value other than agents. Suppose, for example, that we replace each of the agents in the worlds above with spatio-temporal regions. "Infinite Space-Time-Based Pareto" (if you make some spatio-temporal regions better, and none worse, that's an improvement) will then conflict with "Space-Time-Based Anonymity" (if there's a value-preserving bijection between the spatio-temporal regions of two worlds, those worlds are equally good). And the same goes for person-moments, generations, and so forth.

This contradiction between Something-Based Pareto and Something-Based Anonymity is one relatively simple impossibility result in infinite ethics, but the literature contains a variety of others.[33] And note that we can get contradictions between Something-Based Pareto and

[32]See Askell (2018) for more discussion of solutions that posit large amounts of incomparability.

[33]See e.g. Zame (2007), Lauwers (2010), and Askell (2018).

Something-Else-Based Pareto as well: for example, Infinite Agent-Based Pareto and Infinite Space-Time-Based Pareto. The conflict between Zone of Suffering and Zone of Happiness above was one example of this (Zone of Suffering is better at an infinite number of times, and Zone of Happiness for an infinite number of agents).[34]

Pretty clearly, choosing between infinite worlds will require rejecting some principles that looked attractive in finite contexts.

VII Ordinal rankings are not enough

Suppose that, per these impossibility results, we choose one of Infinite Agent-Based Pareto or Agent-Based Anonymity to reject. We're still very far from generating an ordinal ranking over infinite worlds. Infinite Agent-Based Pareto, after all, is an extremely weak principle: it stops applying as soon a given world is better for one agent, and worse for another. And Agent-Based Anonymity stops applying without a welfare-preserving bijection.

Worse, though, *ordinal rankings aren't enough*. They tell you how to choose between *certainties* of one outcome vs. another. But real choices afford no such certainty. Rather, we need to choose between *probabilities* of creating one outcome vs. another.

Suppose, for example, that God offers you the following lotteries:

l1: 40% on a line of people at $\langle 1, 1, 1, 0, 1, 1, 1, 0, \dots \rangle$
60% on zone of suffering, plus an infinite lizard (always at 1) on the side.

l2: 80% on $\langle 1, -2, 3, -4, 5, \dots \rangle$
20% on zone of happiness, plus four infinite lizards (always at -62) on the side.

Which should you choose? It's not at all clear where to even begin.

Here I'll look at a few candidate principles for choosing amongst lotteries like this. This isn't an exhaustive survey, but my hope is that it can give a flavor for the challenge.

[34] Here's another example, from Arntzenius (2014). Consider a single room where Alice will live, then Bob, then Cindy, and so forth, onwards for eternity. In w9, each of them lives for 100 happy years. In w10, each lives for 1000 slightly less happy years, such that each life is better overall. w10 is better for every agent. But w9 is better at every time. So which is better overall? Here, following my verdict about the Zone of Happiness, I'm inclined to say w10: agents seem to me the more fundamental unit of ethical concern. But one might've thought that making an infinite number of spatio-temporal locations worse would make the world worse, not better.

VIII Totals

In finite contexts, many utilitarians look to the total welfare for guidance about the value of the world, including in the context of lotteries. Above I discussed one worry about this: namely, that finite changes can't alter the total welfare in an infinite world. But I think it's useful to note some other ways totals get weird in the context of the infinite.

For one thing, *infinite* changes don't necessarily alter the total welfare, either. Suppose, for example, that faced with a world with infinite people at 1, you can bump everyone up to 2. Per Infinite Agent-Based Pareto, shouldn't you do it? But the total welfare is the same: ∞. So finite influence isn't the totalist's main problem.

But the weirdness gets worse. Consider, for example, a world with infinite people at +2 welfare, and an infinite number at -1. What's the total welfare? It depends on the order you add. If you go: +2, -1, -1, +2, -1, -1 ... then the total oscillates forever between 0 and 2 (if you prefer to hang out near a different number, just add or subtract the relevant amount at the beginning, then start oscillating). If you go: +2, -1, +2, -1, you get ∞. If you go: +2, -1, -1, -1, +2, -1, -1, -1, you get $-\infty$. So which is it? If you're God, and you can create this world, should you?[35]

Or consider a world where the welfare levels are: 1, -1/2, 1/3, -1/4, 1/5, and so on.[36] Depending on the order you use, these can sum to *any welfare level you want*.[37] Pretty clearly, this isn't the type of situation the totalist is used to.

And naïve uses of totals break expected value theory, too. Thus: consider a one-in-a-graham's-number chance of heaven (and nothing otherwise) vs. a 100% chance of heaven. Which is better? Intuitively, and on a standard dominance analysis, the 100% chance is clearly better. But on a naïve expected value calculation, the EV is the same: ∞. And if we add a one-in-a-graham's-number chance of *hell* to either lottery, its EV becomes undefined.

Of course, we can look for more complex remedies for such problems.

[35]Note that the solution to Bostrom (2011)'s worry I mentioned above—namely, focusing on the total welfare you add or subtract from the world, rather than on the total welfare in the world—doesn't help here, on its own. Thus, suppose that you're faced with a world with infinite people at 0, and you're choosing whether to act in a way that will leave infinite people at 2, and infinite people at -1. How much welfare did you add/subtract? And here we have the same order-dependence issues that we have in finding the total within the latter world. Later in the piece, I discuss options for appealing to a definite order.

[36]Of course, one might worry about invoking arbitrarily precise welfare levels; but I'll skip over such issues for now. Those worried about them can discard this example; the problem for the previous one still stands.

[37]This follows from the Reimann Rearrangement Theorem. This example is the axiological analog of the "Pasadena Game" introduced by Nover and Hájek (2004).

Indeed, below I discuss a view that attempts to re-capture the spirit of total utilitarianism as fully as possible in light of these issues (but which, in my opinion, leads to truly horrible places). But naïve uses of totals, at least, look unpromising.

IX Discounts

Would it help if we weighted the locations of value unequally—for example, by applying some sort of exponential discount relative to some ordering of locations? Thus, for example, for a world w with ordered locations of utility $\langle l_1, l_2, l_3, \ldots, l_i, \ldots \rangle$, and for a discount rate α between 0 and 1 (non-inclusive), could we say that the value of w is $\sum_{i=1}^{\infty} \alpha^{i-1} l_i$?[38] This would allow us to say that a world of $\langle 1, 1, 1, 1 \ldots \rangle$ is better than a world of $\langle 2, 2, 2, 2 \ldots \rangle$, while assigning a finite cardinal value to them both. E.g., for a discount rate α of .5, the value of the 1s world is 2 (i.e., $1 \cdot 0.5^0 + 1 \cdot 0.5^1 + 1 \cdot 0.5^2 \cdots = 1 + 0.5 + 0.25 + \ldots$), and the value of the 2s world is 4 (i.e., $2 \cdot 0.5^0 + 2 \cdot 0.5^1 + 2 \cdot 0.5^2 + \cdots = 2 + 1 + 0.5 + \ldots$).

Approaches like this, applied to locations increasingly distant in time from the decision-maker, are common in economics.[39] As Bostrom (2011) notes, though, in order to handle spatially infinite worlds, we would need to treat spatial locations unequally as well—for example, discounting by spatial distance from the decision-maker, too. And we can also imagine other, more exotic ways of discounting, that don't focus directly on either time or space. Thus, for example, some theorists have been interested in the idea that locations of value that are in some sense "simpler" to describe, for some definition of "simpler," should be given ethical priority.[40]

All approaches that weight locations unequally, though, will face the charge that they are privileging the interests of some people over others without reflectively plausible grounds for doing so—and that in this sense, they are engaging in a kind of arbitrary discrimination. After all, people who are distant from us in space and in time, or whose locations are more "complicated," are no less real. And it seems very unappealing, on reflection, to think that we would improve the world by pulling them closer towards us (without changing their utility levels), or by moving them to "simpler" locations instead (a conclusion implied by these discounting views even in finite worlds)—and especially unappealing to say that we should pay

[38] This is a slightly simplified version of the discounted utilitarian rule discussed by Lauwers (2014), p. 6.

[39] See Cowen and Parfit (1992) for discussion.

[40] See e.g. Christiano (2011) and Garrabrant (2014) for comments in this vein. My understanding is that this position is partly inspired by an aspiration to apply some sort of simplicity weight in the context of both anthropic reasoning *and* ethical reasoning (Christiano's article is mostly focused on anthropics, but he also discusses ethical weights in his section on "splitting simulations"). See Carlsmith (2021c) for more on the relevant views in anthropics, here.

extreme costs to do this.[41]

And the costs at stake can be extreme indeed. Thus, for example, if we continue with our discount rate of .5 from above, then faced with a world with one happy person (utility 1) at the second location—that is, $\langle 0, 1, 0, 0, \dots \rangle$, total discounted value .5 ($0 + .5 + 0 + 0 \dots$)—we should be willing to create an infinite number of suffering people (utility -1) in locations 4 and up in order to move the person in location 2 to location 1, thereby yielding $\langle 1, 0, 0, -1, -1, -1, \dots \rangle$ and a total discounted value of .75 ($1 + 0 + 0 - .125 - .0625 \dots$).[42] But causing infinite suffering for the sake of such a re-arrangement seems (at least to me) ethically out of the question. (And note that views with less extreme discounts will lead to qualitatively similar conclusions.)

For reasons like this, many philosophers have found discounting views quite unappealing, and I am inclined to agree.[43] And even if we set such reasons aside,[44] exponential discounts don't, on their own, solve the problems for totalism above. After all, utilities can grow as fast or faster than the discounts shrink.[45] Thus, if our discount rate is .5, but the utility at each location i is 2^{i-1}, then the discounted total is infinite ($1+1+1+1+\dots$); and so, too, is it infinite in worlds where the utility at each location is a million times larger ($1M + 1M + 1M + \dots$). So we've lost Infinite Pareto over the locations in question, and we're back to having to incorporate infinite values into our evaluations of prospects.

X Averages

Could we appeal to averages? After all, if we want to say that $\langle 2, 2, 2, 2, \dots \rangle$ is better than $\langle 1, 1, 1, 1, \dots \rangle$, one option for capturing this would be to assess the value of an infinite world via the limit

$$\lim_{n \to \infty} \frac{\text{total welfare of the locations counted so far}}{\text{number of locations counted so far}},$$

[41] Indeed, weighting based on "simplicity" seems an even poorer fit for our ethical intuitions than weighting based on space or time, since it's not clear that the concept of a location's "simplicity" plays any role in our everyday picture of the world.

[42] See Chichilnisky (1996, p. 240) and Lauwers (2014, p. 8) for more on examples in this vein.

[43] See e.g. Sidgwick (1907), Ramsey (1928), and Parfit (1984), who writes: "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" (p. 486). And note that we can also ask questions about what could make it the case that the discount is a particular value; and about whether discounts that refer to the location of the decision-maker make sense in the context of attempts to place objective axiological values on a world.

[44] Russell (2022) suggests that we should at least consider doing this.

[45] And more extreme discounts lead to even more extreme indifference to what happens beyond a certain zone.

relative to some counting order n. Thus (and no matter how you count), the 2s have a limiting average of 2; and the 1s, a limiting average of 1.

But this approach suffers from a myriad of problems. Here's a sample:

- It's always indifferent to helping finitely many locations, and to adding finitely many suffering locations to a world, since this won't change the limit of the average.
- It's order-dependent in a manner analogous to totalism. For example: if I have infinite locations at 2, and infinite locations at -1, I'll get a different average depending on whether I alternate 2s and -1s (limiting average: 1/2), vs. adding a 2 after every three -1s (limiting average: -1/4). Indeed, I can make the average swing wildly, both above and below zero, depending on the order.[46]
- Even if we fix an ordering, it's indifferent to many ways of helping infinitely many locations, like moving from $\langle 1, 2, 3, 4, \dots \rangle$ to $\langle 2, 3, 4, 5, \dots \rangle$ (limiting average: ∞ in both cases).
- It becomes undefined whenever you end up adding locations in an order like $\langle 1, -3, 5, -7, \dots \rangle$, where the average utility keeps flipping back and forth between -1 and 1.
- It becomes undefined on cases that mix together infinitely good and infinitely bad locations (e.g. $\langle \infty, -\infty, -\infty, -\infty, -\infty, -\infty, \dots \rangle$ vs. $\langle -\infty, \infty, \infty, \infty, \infty, \infty, \dots \rangle$).
- Naively, it implies average utilitarianism about finite worlds. But average utilitarianism is widely thought to be an unattractive view (for example, it endorses creating suffering people, instead of a larger number of happy people who will together drag the average down more).[47]

So appeals to averages of this kind face significant challenges as well.

XI Hyperreals

Could we look for new ways of representing infinite quantities?

One option in this vein comes Bostrom (2011), who suggests mapping infinite worlds to *hyperreal numbers*.[48] I won't examine this proposal in detail here, but here's a brief description.[49] Hyperreal numbers are extensions

[46] One solution to order-dependence is to appeal to the limit of the utility per unit space-time volume, as you expand outward from some/all points. I discuss principles of this type in the section on expansionism.

[47] See Parfit (1984) for canonical discussion.

[48] There are other options as well. For example, see Askell (2018), p. 61, for discussion of whether Conway's (2000) "surreal numbers" might be useful in this context—but she's not optimistic.

[49] See Arntzenius (2014), section 5, for an especially clear introduction.

of the real numbers. They can be both larger and smaller than any real number, while remaining distinct, ordered, and amenable to operations like addition and multiplication. We can represent hyperreals using sequences of real numbers; the hyperreal representation of a real number is just that real number repeated infinitely (e.g., the hyperreal for 3 is $\langle 3, 3, 3, \dots \rangle$); and we say that one hyperreal is bigger than another if it's bigger at a "large" number of locations—where "large" is defined such that no finite set of locations can be large, but where any infinite set of locations whose complement is also infinite can be large, depending on a choice of something called an "ultra-filter."

Equipped with an ordering of the value-bearing locations in your infinite worlds, then, one could imagine making value comparisons between such worlds in the same way one makes size comparisons between hyperreals. Thus, $\langle 1, 1, 1, \dots \rangle$ would be worse than $\langle 2, 2, 2, \dots \rangle$, because it's worse at a large number of locations (for any ultra-filter you pick). However, this approach makes any finite differences between worlds axiologically irrelevant (since such differences will only apply to a small number of locations), so Bostrom proposes a more complicated alternative: namely, mapping a world to the hyperreal corresponding to the sum of the utility as you proceed along the ordering. Thus, the world $\langle 1, 1, 1, \dots \rangle$ would correspond to the hyperreal $\langle 1, 2, 3 , \dots \rangle$; and if you bumped up the first person to 2 (such that the world is now $\langle 2, 1, 1, \dots \rangle$), that would change its corresponding hyperreal to $\langle 2, 3, 4, ... \rangle$—which is bigger.

But Bostrom's proposal suffers from a number of serious problems. First: like appeals to totals and averages, its verdicts are dependent on the ordering you use. A world with infinite 2s and infinite -1s, for example, could yield a hyperreal worse than, or better than, any finite number (since, as we discussed earlier, its total can hang out above or below any finite number indefinitely). On top of this, though, Bostrom's proposal's verdicts are also dependent on how finely you carve the locations in question.[50] Thus, suppose that the locations are times, and that the seconds in the world have utilities $\langle 1, 1, 1 , \dots \rangle$, then your world gets better depending on whether you use one-second time intervals (hyperreal: $\langle 1, 2, 3, \dots \rangle$), two-second time intervals (hyperreal: $\langle 2, 4, 6 , \dots \rangle$), three-second time intervals (hyperreal: $\langle 3, 6, 9, \dots \rangle$), and so on.[51]

And even after you've fixed your ordering and your carving of locations, your verdicts are *additionally* sensitive your choice of ultrafilter, which determines, for any infinite set of locations S whose complement is also infinite, whether S or its complement will be treated as "large." Thus:

[50]See Arntzenius (2014), p. 49.

[51]This example is inspired by one in Askell (2018), p. 59, footnote 61. Perhaps you could say that if we appeal to agents as our locations, we will have a privileged carving of locations; but agents are also the least well-suited locations for a natural ordering, and natural orderings seem necessary for avoiding totally-arbitrary order-dependence.

- The world $\langle 1, -2, 1, 1, -2, 1, 1, \ldots \rangle$ can be made better, worse, or equal to an empty world (since its corresponding hyperreal, $\langle 1, -1, 0, 1, -1, 0, \ldots \rangle$, can be made to equal $\langle 1, 1, 1, \ldots \rangle$, $\langle -1, -1, -1, \ldots \rangle$, or $\langle 0, 0, 0, \ldots \rangle$ at a large number of locations).
- A world whose corresponding hyperreal reaches every finite value infinitely many times (for example, worlds with utilities chosen by a random walk) can be made equally valuable to a world of any finite value: just make the set of locations in the hyperreal with that finite value large.
- The world $\langle 2, 2, -2, 2, -2, \ldots \rangle$ is either twice or four times as good as a single person at 1 (its corresponding hyperreal is $\langle 2, 4, 2, 4, 2, \ldots \rangle$, and is thus equivalent to either 2 or 4, depending on whether the set of even or the set of odd-numbered locations is large).[52]

This last sensitivity—to the choice of ultrafilter—seems to me especially dire: as far as I can tell, it corresponds to nothing of plausible ethical relevance.[53]

XII Expansionism

Let's turn to "expansionism"—an approach focused on the utility contained inside expanding bubbles of space-time.

Vallentyne and Kagan (1997) suggest that if we have two worlds with the same locations, and these locations have an "essential natural order," we can compare the value of the two worlds by comparing the amounts of utility contained in a "bounded uniform expansion" from any given location. In particular: if there is some positive number k such that, for any bounded uniform expansion, the utility inside the expansion eventually stays larger by more than k in world_i vs. world_j, then world_i is better.

Thus, for example, in a comparison of $\langle 1, 1, 1, 1, \ldots \rangle$ vs. $\langle 2, 2, 2, 2, \ldots \rangle$, the utility inside any expansion is bigger in the 2 world. And similarly, in $\langle 1, 2, 3, 4, \ldots \rangle$ vs. $\langle 2, 3, 4, 5, \ldots \rangle$, expansions in the latter will always be greater by 1.

Vallentyne and Kagan don't define "essential natural order" fully, but importantly, on their view, things like agents and person-moments don't have it (agents can be listed by their height, by their passion for Voltaire, etc), but space-time does (there is a well-defined notion of a "bounded-region of space-time," and we can make sense of the idea that in order to get from a to b, you have to go through c).[54] Pinning down "uniform expansion" also

[52] See Bostrom (2011), p. 23, and Arntzenius (2014), p. 50-1, for more on these objections.

[53] In particular, to me it seems worse than sensitivity to spatio-temporal structure, which at least has some grounding in our intuitions about which worlds are "dense" with value. That said, perhaps the choice of ultra-filter can draw on similar intuitions.

[54] Vallentyne and Kagan (1997): "The notion of locational order that we have in mind is

requires some subtlety (see Arntzenius (2014) for discussion), but broadly: the relevant bubble of space-time should be growing at the same rate in all directions.[55]

A major problem for Vallentyne and Kagan is that their principle only provides an ordinal ranking. But Arntzenius (2014) suggests a modification that generalizes to choices between lotteries: instead of looking at the *actual* value at each location, look at the *expected* value. Thus, suppose you're choosing between the following lotteries, all for with the same locations of value:

l3: 50% on $\langle 1, 1, 1, 1, \dots \rangle$
50% on $\langle 1, 2, 3, 4, \dots \rangle$

l4: 50% on $\langle -1, 0, -1, 0, \dots \rangle$
50% on $\langle 1, 4, 9, 16, \dots \rangle$

Arntzenius uses the expected values at the locations to make lotteries involving multiple worlds into single "equivalent worlds" comparable using the Vallentyne-Kagan methodology. That is: l3 is equivalent to $\langle 1, 1.5, 2, 2.5, \dots \rangle$, and l4 is equivalent to $\langle 0, 2, 4, 8, \dots \rangle$. The latter is better according to Vallentyne-Kagan, so Arntzenius says to choose it.[56]

I'll note two major problems with this approach:

1. It leads to results that are unattractively sensitive to the spatio-temporal distribution of value.

2. It fails to deliver verdicts in lots of choices.

To get a flavor of problem 1: consider an infinite line of planets, each of which houses a Utopia, and none of which will ever interact with any of the others. On expansionism, it is *extremely good* to pull all these planets an inch closer together: so good, indeed, as to justify any finite addition of

that of a topological manifold. We shall not define it precisely, but the rough idea is that locations are connected to each other so that the notion of a (continuous, or unbroken) path is well defined and all locations are path connected... Naturalness is the most difficult notion-indeed, we are embarrassed to say that we cannot give a crisp definition of what we mean by it!" (p. 12-14).

[55] Arnzenius (2014): "Vallentyne and Kagan want their theory to apply to cases in which there is more than one relevant dimension, e.g. in an infinite 3-dimensional space, or an infinite 4-dimensional space-time. In that case Vallentyne and Kagan say that a 'uniform' expansion of an initial spatial region S1 means that at each step in the sequence one 'adds a band of constant width to the previous region'. This, of course, presupposes that the space in question comes equipped with a metric, but that doesn't seem to be a too severe a restriction. Vallentyne and Kagan do not make precise what they mean by a 'band of constant width'. But I suggest that we can take it to mean the following. A band of width w around a region R consist of all the points that are within distance w of some point in R. (This will not really look like a band of constant width if the region in question has 'deep dents', but this does not matter.)" (p. 39).

[56] See Wilkinson (2021) for similar themes.

Dystopias to the world.[57] After all, pulling on the planets so that there's an extra Utopia every x inches will be enough for the eventual betterness of the uniform expansions to compensate for any finite number of hellscapes. But this seems wrong. In particular: *no one benefits* (indeed, no one notices) when you pull the planets closer together—it's the same population, with the same welfare levels, either way. But a *lot* of extra people suffer when you add arbitrary finite numbers of dystopias.

For closely related reasons, expansionism violates *both* Infinite Agent-Based Pareto *and* Agent-Based Anonymity. Consider the following example from Askell (2018), p. 83, in which three infinite sets of people (x-people, y-people, and z-people) live on an infinite sequence of islands, which are either "Balmy" (such that three out of four agents are happy) or "Blustery" (such that three out of four agents are sad). Happy agents are represented in black, and sad agents in white.

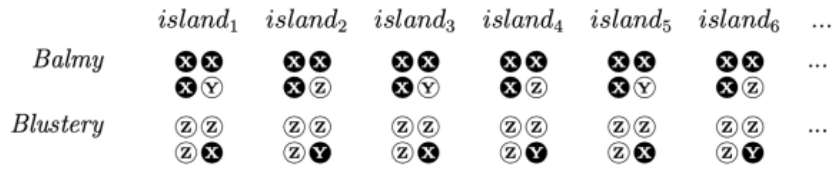

Figure 31: Balmy and Blustery

Figure 1: From Askell (2018), p. 83; reprinted with permission

Here, expansionism prefers Balmy to Blustery—and intuitively, we might agree. But Blustery is better for the y-people, and worse for no one: so we've violated Infinite Agent-Based Pareto. And there is a welfare-preserving bijection from Balmy to Blustery as well: so we've violated Agent-Based Anonymity as well.

The basic issue, here, is that expansionism's moral focus is on *space-time positions*, rather than people or person-moments. In some cases (e.g. Balmy vs. Blustery), this actually does fit with our intuitions: universes that seem dense with value also seem better. But stated abstractly, such a moral focus is quite alien; and I find that when I reflect on how much suffering I want to cause in order to pull happy planets closer together, the appeal from intuition starts to wane.

Let's turn to problem 2: expansionism fails to provide guidance in lots of cases—and in particular, cases where the worlds in question don't all have the same locations.[58]

[57]Thanks to Amanda Askell, Hayden Wilkinson, and Ketan Ramakrishnan for discussion.

[58]Expansionism also fails to give verdicts in various cases with the same locations. For example: it becomes undefined on "zone of suffering/happiness" type cases, because

Consider, for example, the choice between creating a spatially-finite world with an immortal person trudging from hell to heaven, the welfare of whose days corresponds to the sequence $\langle \ldots, -2, -1, 0, 1, 2, \ldots \rangle$, and a spatially-infinite universe that only lasts a day, with an infinite line of people the welfare of whose one-day lives corresponds to $\langle \ldots, -2, -1, 0, 1, 2, \ldots \rangle$. How shall we match up the locations in these worlds? Depending on how we do it, we'll get different expansionist verdicts (i.e., we can make every location in the first world better than its counterpart in the second, or vice versa). And we'll hit even worse arbitrariness if we try to e.g. match up locations for worlds with different numbers of dimensions (e.g., pairing locations in a 2d world with locations in a 4d one), let alone worlds whose differences reflect the full range of logically-possible space-times.

One option is to accept incomparability in such cases. But note that this incomparability infects our lotteries as well. Thus, for example, suppose that infinite space-times A and B can't be matched up with each other in any non-arbitrary way. And now suppose that I'm choosing between lotteries like:

l5: 99% on a A-world of -1s
1% on a B-world of 2s.

l6: 99% on a A-world of 2s
1% on a B-world of -1s.

The problem is that because these worlds can't be matched up, we can't turn these lotteries into single worlds we can compare via the Vallentyne-Kagan approach. So even though it looks plausible that l6 is preferable, Arntzenius's approach is silent.[59]

Will this problem arise in practice? Arntzenius (2014) and Wilkinson (2021)

different expansions will give different verdicts, depending on whether they grow faster or slower than the "zone of x" does (see Askell (2018) p. 81).

[59]We might look for ways to get an overall ranking out of comparing the A-world in l5 with the A-world in l6, and same for the B-worlds, and then to get to an overall verdict that way. This isn't Arntzenius's approach, though: his approach specifically tries to make an individual *lottery* into an "equivalent world," with expected utilities at the locations rather than utilities proper. And we would still face additional problems in e.g. cases where we have different probabilities on the world-types in each case (i.e., 98% on the A-world in l5, vs. 99% in l6), cases where one of the lotteries includes a world-type that can't be compared to any of the others at stake, and so on. That said, I'm not, here, trying to delve into all the possible moves and countermoves available in the context of expansionism; rather, I'm trying to give a flavor for the sorts of challenges that arise.

seem to think not.[60] But I disagree.[61] We should already have non-zero credence on our living in various space-times that can't be matched up, and (absent small-probability discounting), it doesn't matter how small the probability on the B-world is in the case above. What's more, we should have non-zero credence that in the future, we'll be able to create all sorts of different infinite baby-universes—including ones whose their causal relationship to our universe doesn't support a privileged mapping between their locations.

This seems like a general problem for any approach infinite ethics that requires identity or counterpart relations between spatio-temporal locations across worlds.[62]

XIII What's the most bullet-biting utilitarian response we can think of?

As a final sample from the space of possible views, let's consider the view that seems to me most continuous with the spirit of a simple, bullet-biting hedonistic utilitarianism.[63] This view doesn't care about people, or expanding bubbles of space-time, or Infinite Agent-Based Pareto. All it cares about is *the amount of pleasure vs. pain in the world*. Pursuant to this single-minded focus, it groups worlds into four types:

1. Positive infinities. Worlds with infinite pleasure, and finite pain.

[60] Arntzenius (2014), p. 40: "More precisely: while there will be a large amount of pairs of worlds whose relative utility will be indeterminate due to the absence of, or vagueness of, the relevant counterpart relations, this will typically not be the case when we are considering worlds that according to an agent's credences are likely consequences of different actions between which the agent is deciding. That is to say, in the context of my, yet to be detailed, solution in terms of expected probabilities, the third problem will rarely, if ever, lead to indeterminacy." Wilkinson (2021), p. 1922: "Alternatively, suppose that our locations are spacetime positions. Then we may not have such a theory of identity. But that is no trouble—there is an obvious identity/counterpart relation, at least in any pairs of worlds we'll ever need to compare. Since our actions necessarily cannot change the past, any such worlds will share the same past events (at the same positions). In worlds like these with common histories up to the present, let us map all past positions to those occupied by the same events (e.g., we can map the position of Runnymede in 1215 in one world to the same position in every other world, as that's where the signing of the Magna Carta occurs in all of them). Then we can map future positions too: each such x is uniquely specified by its spatial and temporal distance from (any four) past points, so we can specify its transworld identities/counterparts as the positions which are also those same distances from those same points."

[61] And even if the problem didn't arise in practice, I would still see it as an issue in theory.

[62] See e.g. Wilkinson (2022) and Easwaran (2021).

[63] This view is closely akin to Bostrom's (2011) "Extended Decision Rule" (p. 29), though Bostrom's view simply ignores *Mixed infinity* worlds, whereas the view in the text treats them as 0. This difference matters in cases where e.g. you can shift the probability of living in a mixed infinity world, but it's not important for the present discussion.

Value: ∞.

2. Negative infinities. Worlds with infinite pain, and finite pleasure. *Value:* $-\infty$.

3. Mixed infinities. Worlds with infinite pleasure *and* infinite pain. *Value: o* (the good and bad infinities cancel).[64]

4. Finite worlds. Worlds with finite pleasure and finite pain. *Value: as given by total utilitarianism.*

This view's decision procedure is: first, maximize the probability of positive infinity minus the probability of negative infinity (call this quantity "the diff"). Then, if more than one available action maximizes the diff, use the EV from mixed infinites and finite worlds to break the tie (though whether ties will come up very often in practice is a further question—I'm skeptical).[65]

Call this the "four types" view. To see what it implies, consider the following worlds:

- Heaven: Infinite people living the best possible (painless) lives you can imagine, forever.
- Infinite Lizard: A single barely-conscious, slightly-happy lizard floating in space for eternity.
- Heaven+Speck: Infinite people living in bliss for eternity, but each gets a speck in their eye one time.
- Hell+Lollypop: Infinite people being tortured for eternity, but each gets to lick a lollypop one time.
- Infinite Speck: Infinite barely-conscious mice who pop into existence, feel a mildly-irritating dust-speck in their eye, then wink painlessly out of existence.

[64] We can also imagine versions that make mixed infinities incomparable to finite worlds, and to each other (though worse than positive infinities, and better than negative infinities). But my paradigm bullet-biting utilitarian doesn't like incomparability; and regardless, positing it leads to problems similar to the ones I discuss below.

[65] See e.g. Bostrom (2011): "The epistemic probabilities that enter into the calculation can be sensitive to a host of imprecise and fluctuating factors: the estimated simplicity of the hypotheses under consideration, analogies (more or less fanciful) derived from other domains of our changing experience, the pronouncements of miscellaneous authorities, and all manner of diffuse hunches, inklings, and gut feelings. It would seem almost miraculous if these motley factors, which could be subjectively correlated with infinite outcomes, always managed to conspire to cancel each other out without remainder. Yet if there is a remainder—if the balance of epistemic probability happens to tip ever so slightly in one direction—then the problem of fanaticism remains with undiminished force" (p. 33).

- HELL: Infinite people being tortured for eternity (with no pleasure whatsoever).[66]

On the four types view:

- HEAVEN and INFINITE LIZARD are equally good; INFINITE SPECK and HELL are equally bad; and HEAVEN+SPECK and HELL+LOLLYPOP are both equivalent in value to an empty world, and to each other.
- Faced with a choice between HEAVEN + SPECK, or a lottery with a one-in-a-graham's-number chance of INFINITE LIZARD, and HELL+LOLLYPOP otherwise, the four types view chooses the lottery.
- Faced with a choice between HEAVEN + SPECK, or a finite world where one person eats a sandwich and then dies painlessly, the four types view goes for the sandwich.
- The four types view is indifferent to adding an infinity of eternally happy people to any world that already has infinite pleasure (for example, the first four worlds), or to preventing the addition of an infinity of suffering people to any world that already has an infinity of pain (e.g., any of the last four worlds). In both cases, it would rather focus on eating another bite of sandwich in a finite world.

We can see the four types view as continuous with a certain kind of "pleasure/ pain-anonymity" principle. That is, if we assume that pleasure/pain come in units that can always be aggregated and weighed against each other (such that e.g., there is some amount of lizard time that outweighs a moment in heaven; some number of dust specks that outweigh a moment in hell, etc—a classic utilitarian thought), then you can build the evaluative equivalent of every positive infinity world by re-arranging INFINITE LIZARD, of every negative infinity world by re-arranging INFINITE SPECK, and of every mixed infinity world by re-arranging both in combination. It's the same (quality-weighted) *amount* of pleasure and pain regardless, says this view, and *amounts* of pleasure and pain (as opposed to densities, or placements in different people's lives, or whatever) were what utilitarianism was supposed to be all about.[67]

[66]Note that the relationship between HEAVEN and INFINITE LIZARD is distinct from the relationship between a finite Utopia and a lizard with a sufficiently long life that its world contains more total welfare (i.e., the relationship at stake in the standard repugnant conclusion). In particular, it's not the case that the total welfare in INFINITE LIZARD is higher than the total welfare in HEAVEN, and various principles would choose HEAVEN over INFINITE LIZARD that would not apply in the finite case (for example, if we make the lizard a citizen of heaven and grant that happy existence is better for someone than non-existence, then Infinite Agent-Based Pareto chooses HEAVEN over INFINITE LIZARD; but a comparable argument does not apply to the standard repugnant conclusion). And the something similar holds for INFINITE SPECK and HELL (e.g., we can make the mice the citizens of HELL, such that HELL is worse for everybody).

[67]Thanks to Amanda Askell for discussion of the sense in which utilitarians really care

There is a certain logic to it. But also: it's horrifying. Trading a world where an infinite number of people have infinitely good lives, for an effective guarantee of a world where infinitely many people are eternally tortured, to get a one-in-a-graham's-number chance of creating a single, immortal, barely-conscious lizard? To me this seems much worse than e.g. paying to pull planets together, or not knowing what to say about worlds with non-matching space-times.

But also: such a choice doesn't really make sense on its own terms. INFINITE LIZARD is getting treated as lexically better than HEAVEN + SPECK, because it's possible to map all of INFINITE LIZARD's barely-conscious happiness onto something equivalent to all the happiness in HEAVEN+SPECK, with the negative infinity of the dust specks left over. But so, equally, is it possible to map all of INFINITE LIZARD's barely-conscious happiness onto everyone's first nano-seconds in heaven, to map those nano-seconds onto each of their dust specks in a way that would more than outweigh the dust-specks in finite contexts, and to leave everyone with an infinity of fully-conscious happiness left over. That is, the "Infinite Lizard Has All of Heaven's Happiness" and "No Amount Of Time In Heaven Can Outweigh The Dust Specks" mappings aren't, actually, privileged here: one can just as easily interpret HEAVEN + SPECK as ridiculously better than INFINITE LIZARD (indeed, this is my default stance). But the four types view fixates on those particular mappings anyway.

XIV Bigger infinities and other exotica

I've now discussed six possible approaches to infinite ethics that go beyond ordinal rankings: totals, discounts, averages, hyperreals, expansionism, and the four types view. All of them seem to me unattractive, and some seem downright horrifying. And while this hasn't been an exhaustive survey,[68] we know that no theory on offer will avoid the impossibility results already discussed.

I also want to note, though, that all the discussion thus far has been mostly focused on a specific range of cases: namely, countable infinities. But there is an un-ending hierarchy of larger infinities, too, which we haven't yet attempted to grapple with.[69]

about the *amount* of utility.

[68]For positive views I'm not discussing, see e.g. Jonsson and Voorneveld (2018), Easwaran (2021), and Wilkinson (2022)—though the proposals from Wilkinson and Easwaran, at least, both require that the worlds being compared have exactly the same locations.

[69]In particular: according to Cantor's theorem, the powerset of any set A (including any infinite set) has strictly greater cardinality than A—so you can reach endlessly bigger infinities simply by continually taking powersets. And we might also wonder about large cardinals, inaccessible via power-setting.

Do we need to? I'm not sure. Certainly, it's quite difficult to imagine worlds with e.g. one person for every real number (let alone larger infinites than that), and salient scientific hypotheses involving infinite worlds don't ask us to. So considering larger infinities requires a further step in the direction of the exotic—and perhaps, the incoherent/impossible. And countable infinities are certainly hard enough on their own.

On the other hand, to the extent that you were impressed, ethically, by the stakes of countably infinite payoffs, relative to finite ones, it seems plausible that you should be similarly impressed by the stakes of uncountably infinite payoffs, relative to countable ones. And while it may be hard to imagine payoffs of such cardinality, it also seems hard to rule out ever getting evidence for their availability (God, for example, could appear before you, announcing the chance to create such a large-cardinality heaven—are you certain the offer is fake?). So plausibly, the same logic that asks us to grapple with small probabilities of merely countable infinities would ask us to grapple with (obsess about?) larger infinities as well.

And if we do have to grapple with these larger infinities, it seems likely to me that they will break whatever principles we worked so hard to develop for the countable case. After all, countable infinities have very different properties from even the smallest uncountable infinity; some of the approaches above rely specifically on counting all the locations in a given order (something you can't do with uncountable infinities); and an infinite hierarchy of ever-larger infinities seems, to say the least, a daunting challenge to handle comprehensively. In this sense, ignoring uncountable infinities might be a recipe for the same kind of rude awakening that countable infinities give to finite ethics. Yes, the ethical problems in some limited domain (e.g., finite worlds, countably infinite worlds) are in some sense "hard enough"—but if your solutions predictably break as soon as you leave that domain, and you need to leave that domain eventually, then over-focus on it risks wasted effort.

And we can imagine other exotica breaking our theories as well. Thus, for example, expansionism relies on all the worlds we're considering having something like a space-time (or at least, a "natural ordering" of locations). But do worlds with space-times, or worlds with any natural orderings of locations, exhaust the worlds of moral concern? I'm not sure. Admittedly, I have a tough time imagining standardly valuable things existing without something akin to space-time; but I haven't spent much time on the project, and I have non-zero credence that if I spent more, I'd come up with something.

That said, exactly which exotica it makes sense to try to incorporate into one's theorizing and decision-making seems to me a tricky question. For we might wonder: how easy is it to end up with non-zero credences on any old crazy and questionably-coherent proposition? After all, God (or

some more mundane epistemic superior) could always appear before you announcing that roughly *any p* is true. Should this always count as at least some evidence for *p*? What's your credence, for example, that contradictions can be true? Or that probabilities need not add up to 1? Or that 1+1=3? Or that phenomenal consciousness is constituted by sufficiently cheesy sourdough bread? Need our ethics accommodate such possibilities? And if not, should we say the same about uncountably infinite payoffs, worlds without space-times, and so on?

I don't have a worked-out view about how to draw the relevant lines, here—and it seems possible to me that uncountably infinite payoffs will fall on the "OK to ignore it" side. I'm very skeptical, though, that countable infinities will. Unlike the exotica above, countably infinite cases seem readily imaginable, and we have strong ethical intuitions about many of them (e.g., Heaven + Speck vs. Hell + Lollypop). What's more, we have very credible scientific theories that say that our actual universe contains a countably infinite number of people; credible decision theories that say that we can have infinite influence on that universe; widely-accepted religions that posit infinite rewards and punishments; and a possibly technologically-extravagant future ahead of us where baby-universes/wormholes etc appear much *more* credible, at least, than "consciousness = cheesy-bread." Indeed, as Bostrom (2011, p. 38) notes, *conditioning* on absence of infinities (or *ignoring* infinity-involving possibilities) leads to weird behavior in other contexts—e.g., refusing to fund scientific projects premised in infinity-involving hypotheses, insisting that the universe is actually finite even as more evidence comes in, etc.

So even if we ignore the exotica above, I don't think we can ignore the challenges of infinite ethics more generally.

XV The death of the utilitarian dream

In the discussion thus far, I've been aiming to convey a sense of how difficult infinite ethics can be. Even beyond "how can finite genocides matter in an infinite world?" and "should I pay any finite cost for any probability of an infinite payoff?", we've got bad impossibility results even just for ordinal rankings; we've got a smattering of theories that are variously incomplete, order-dependent, Pareto-violating, and otherwise unattractive/horrifying; and we've got an infinite hierarchy of further infinities, waiting in the wings to break whatever theory we settle on.

Of course, there's much more to say about all of these issues, my survey of the available views has not been exhaustive, and I expect further work on the topic to lead to further clarity about the best overall response. But even without this response in hand, I think we're in a position to draw out some interesting implications from the issues discussed thus far. The rest of this essay focuses on a few of these implications.

The first is that I think infinite ethics punctures a certain type of utilitarian dream. It's a dream I associate with a utilitarian friend of mine, who once warned me, when I was in the midst of offering a possible counterexample to his view: "I bite all the bullets." In my caricatured picture, it's the dream of hitching yourself to some simple ideas—e.g., expected utility theory, totalism in population ethics, hedonism about well-being—and riding them wherever they lead, no matter the costs. Yes, you push fat men and harvest organs; yes, you destroy Utopias for tiny chances of creating zillions of evil, slightly-happy rats (plus some torture farms on the side). But you always "know what you're getting"—e.g., more expected net pleasure. And because you know what you're getting, you can say things like "I bite all the bullets," confident that you'll always get at least this one thing, whatever else must go.

Plus, other people have problems you don't. They end up talking about vague and metaphysically suspicious things like people, whereas you only talk about valanced experiences—which, you assume, are on much more solid ground. They end up writing papers entirely devoted to addressing a single category of counter-example—even while you can sense the presence of many others, just offscreen. And more generally, their theories are often complicated, ad hoc, intransitive, or incomplete.

Indeed, even people who reject this dream can feel its allure. If you're a deontologist, scrambling to add yet another epicycle to your already-complex and non-exhaustive principles, to handle yet another counterexample, you might hear, sometimes, a still, small voice saying: "You know, the utilitarians don't have this kind of problem. They've got a nice, simple, coherent theory, that takes care of this case and a zillion others in one fell swoop, including all possible lotteries (something my deontologist friends barely ever talk about). And they always get more expected net pleasure in return. They sure have it easy…" In this sense, "maximize expected net pleasure" can hover in the background as a kind of default. Maybe you don't go for it. But it's there, beckoning, and making a certain kind of sense.

But I think infinite ethics changes this picture. In the land of the infinite, the bullet-biting utilitarian train runs out of track. You have to get out and wander blindly. The issue isn't that you've become fanatical about infinities: that's a bullet, like the others, that you're willing to bite. The issue is that once you've resolved to be 100% obsessed with infinities, *you don't know how to do it*. Your old approach (e.g., "just sum up the pleasure vs. pain") doesn't make sense in infinite contexts, so your old trick—just biting whatever bullets your old approach says to bite—doesn't work (or it leads to *horrific* bullets, like trading Heaven + Speck for Hell + Lollypop, plus a tiny chance of the lizard). And when you start trying to craft a new version of your old approach, you run headlong into Pareto-violations, incompleteness, order-dependence, spatio-temporal sensitivities, appeals to

persons as fundamental units of concern, and the rest. In this sense, you start having problems you thought you transcended—problems like the problems the other people had. You start having to rebuild yourself on new and more complicated foundations. You start writing whole papers about a few counterexamples, using principles that you know don't cover all the choices you might need to make. Your world starts looking stranger, less elegant, more ad hoc. You start to feel, for the first time, genuinely lost.[70]

This isn't to say that the problems of infinite ethics are hopeless. My point, rather, is that we can already tell that the best response to these problems won't look like the simple, complete, impartial, totalist, hedonistic, EV-maximizing utilitarianism that some hoped would answer every ethical question—and which it is possible to treat as a certain kind of fallback. Maybe the best view will look a lot *like* such a utilitarianism in finite contexts—or maybe it won't. But regardless, a certain type of dream will have died. And if we know it will die eventually, it should die now, too.

XVI Everyone's problem

That said, infinite ethics is a problem for everyone, not just utilitarians. Everyone (even a virtue ethicist) needs to know how to choose between Heaven + Speck vs. Hell + Lollypop, given the opportunity. Everyone needs decision procedures that can handle some probability of infinitely-consequential actions. Faced with impossibility results, everyone has to give something up. And sometimes the intuitions and principles you give up matter in finite contexts, too.

A salient example to me, here, is the ethical significance of space-time. Utilitarian or no, many philosophers want to deny that a person's location in space and time has intrinsic ethical (or at least, axiological) significance. Indeed, claims in this vicinity play an important role in standard arguments against discounting the welfare of future people, and in support of a thesis called "strong longtermism"—the view that positively influencing the long-term future is the key moral priority of our time—that has recently received an increasing amount of academic and popular attention.[71] But notably, various prominent views in infinite ethics (notably, expansionist

[70]Obviously, I'm not trying to cover all the logically possible positions with respect to these issues. Rather, I'm trying to gesture at a cluster of related tendencies. I think this is worthwhile partly because I think the allure of the utilitarian dream in question is, at least partly, an aesthetic one—and that encounter with infinite ethics renders this aesthetic unsustainable. Of course, even those initially enamored of such an aesthetic can end up ambivalent about it for other reasons, too—and professional philosophers, in particular, might be amply acquainted with the reasons on offer.

[71]See e.g. MacAskill and Greaves (2021); and see MacAskill (2022) for a popular introduction (though one focused on a somewhat weaker thesis—namely, that positively influencing the long-term future is *a* key moral priority of our time).

views—but also all views that appeal to space-time as a source of natural ordering) reject this sort of indifference to moving people around in space-time, while leaving their welfare unaffected. On these views, locations in space and time matter *a lot*—enough, indeed, to make e.g. pulling infinite happy planets an inch closer together worth any finite amount of additional suffering. On its own, this isn't enough to get conclusions like "people matter more if they're nearer to *me* in space and time" (the claim that strong longtermism, for example, most needs to reject)—but it's an interesting departure from ethical indifference to spatio-temporal location, and one that, if accepted, might make us question other similarly-flavored intuitions.

And the logic that leads to non-indifference about space-time is understandable. In particular: infinite worlds look and behave very differently depending on how you order their value-bearing locations, so if your view focuses on a type of location that lacks a natural order (e.g., agents), it often ends up indeterminate, incomplete, and/or in violation of Infinite Pareto over the locations in question. Space-time, by contrast, comes with a natural order, so focusing on it cuts down on arbitrariness, and gives us more structure to work with.

Something somewhat analogous happens, I think, with persons vs. experiences as units of concern. Some philosophers are tempted, in finite contexts, to treat experiences (or "person-moments") as more fundamental.[72] But in infinite contexts, refusing to talk about persons makes it much harder to distinguish between worlds like Heaven + Speck and Hell + Lollypop—worlds that prompt intuitions plausibly driven by the fact that in Heaven + Speck, everyone's lives are infinitely good, but in Hell + Lollypop, everyone's lives are infinitely bad. So to retain such intuitions, it becomes tempting to bring persons back into the picture.[73]

We can see the outlines of a broader pattern. A certain kind of reductionist impulse in finite ethics often tries to ignore structure.[74] It calls more and more things (e.g., the location of people in space-time, the locations of experiences in lives) irrelevant, so that it can hone in on the fundamental unit of ethical concern. But infinite ethics *needs* structure, or else too much dissolves into re-arrangeable equivalence. So it often starts adding back in what finite ethics threw out.[75]

[72]See e.g. Campbell (2021) for discussion and some motivation. Parfit (1984) seems to me to be channeling this impulse when writes that on a reductionist view about personal identity: "it is ... more plausible to focus, not on persons, but on experiences, and to claim that what matters morally is the nature of these experiences" (p. 446).

[73]See Askell (2018), p. 198, for more on this.

[74]See e.g. Chappell (2011) on "value atomism."

[75]Indeed, we might wonder if there is even more structure to be not-ignored. Perhaps, indeed, a methodology that attempts to derive the value of the whole from the value of some privileged type of part is worse than one might've thought (see Chappel (2011) for some considerations; and thanks to Carl Shulman for discussion).

These are a few examples of finite-ethical impulses that infinites put pressure on. I expect there to be many others. Indeed, I think it's good (though dispiriting) practice, in finite ethics, to make a habit of checking whether a given proposal breaks immediately upon encounter with the infinite. If so, that doesn't make the proposal useless; but it at least suggests a need for further refinement.

XVII Is this an argument for meta-ethical despair?

So far I've been focusing on the implications of infinities for normative ethics. But I also want to briefly touch on their implications for meta-ethics as well.

In particular: in my experience, some people exposed to the challenges of infinite ethics see them as evidence against moral realism.[76] This is often more of an inchoate intuition than a structured argument, but it's sufficiently common that I think it's worth examining directly. Consider the following reconstruction:

1. If morality does not have property X, then moral realism is less likely to be true.

2. Infinite ethics suggests that morality does not have property X.

3. Thus, moral realism is less likely to be true.

Here, candidates for property X might include: simplicity, completeness, simultaneous compatibility with initially attractive principles like Infinite Agent-Based Pareto *and* Agent-Based Anonymity, or "making some kind of intuitively resonant *sense*."

Is this a good argument? I think it depends. I'm happy to grant (2) for various of the candidate Xs just listed. But (1) seems weaker. After all, stated abstractly, moral realism—i.e., the thesis that there are mind-independent moral facts, and that moral claims are truth-apt—does not strictly entail that morality should have any of the properties above.[77] So we need a story about why morality's having property X is nevertheless *more likely*, conditional on moral realism being truth, than it is on moral realism being false.

Are stories of this kind available? We can at least speculate. For example, we might think that moral realism makes morality more analogous to

[76] See e.g. Rob Wiblin's remarks in his and Askell's (2018) podcast: "So this all sounds like a council of despair to some extent. How much of an update is this against moral realism or naturalism or at least against consequentialism?" Note that the meta-ethical despair at stake here is distinct from the normative-ethical paralysis involved in e.g. thinking that your actions don't make a difference to the value of infinite worlds.

[77] I won't, here, attempt to tackle questions about the best way to define moral realism. My paradigm moral realist, though, is Enoch (2011).

physics, and thus something we should expect to be simple in the way we generally expect physical theories to be simple; whereas moral anti-realism makes morality more closely related to human psychology, and thus more likely to be complicated, messy, and vague.[78] Or we might think that insofar as our moral intuitions are tracking some mind-independent moral reality, we should expect those intuitions to cohere with each other, rather than to lead to the type of contradictions and impossibility results we get from infinite ethics. Or we might see moral realism as importantly bound up with the idea that reflection on morality should cause it to make *more* sense, rather than less—but infinite ethics seems like a permanent move towards "less." And perhaps there are other possible stories as well. Indeed, the fact that infinite ethics can so readily *seem* an argument against moral realism, to some people, suggests that some story of this kind was operative in their own background meta-ethical thinking—whether justifiably or no.

Drawing out and justifying any one of these stories is beyond the scope of the present paper. But I think they're worth attention. Indeed, I see them as interestingly continuous with other sorts of despair one occasionally finds prompted by normative-ethical difficulties. Thus, for example, Sidgwick famously took the contradiction between intuitions in favor of impartiality and egoism as a counsel for some sort of despair about rationalizing morality;[79] some population ethicists have seen its difficulties, even in finite cases, as arguments for meta-ethical anti-realism;[80] and some deontologists see the difficulty of generating plausible systematizations of our deontological intuitions in a similar light.[81] So despair induced by infinite ethics looks like an instance of some broader connection between "it looks like normative ethics isn't going to give us what we wanted"

[78]Ramakrishnan (unpublished) suggests this: "At that point we may need to accept that moral reality is—as consequentialists have long urged—unified, coherent, and profoundly counterintuitive. Or else we may need to abandon the assumption that our moral convictions reflect an external reality at all. If moral truths are simply like parochial facts about social custom, it is not surprising that our moral convictions should prove recalcitrant to capture by consistent principle. If morality is just a welter of attitude and feeling—if there is no external benchmark, supplied by the reality of the world, to which these attitudes and feelings are beholden—there is little or no pressure to clean its messiness up."

[79]See Sidgwick (1907): "the Cosmos of Duty is thus really reduced to a Chaos: and the prolonged effort of the human intellect to frame a perfect ideal of rational conduct is seen to have been foredoomed to inevitable failure." That said, the nature and rationale for Sidgwick's despair is a matter of controversy in the literature.

[80]See e.g. McMahan (2013): "Problems in the morality of causing people to exist seem to me the most difficult and intractable of all the problems of which I am aware in normative and practical ethics. They suggest that it is a real possibility that any moral theory that is both complete and coherent will have implications that are intuitively intolerable. It is these problems, therefore, rather than arguments in metaethics about the queerness of objective values, the connections between normativity and motivation, and so on, that seem to me to pose the greatest challenge to realism in ethics" (p. 34).

[81]See Ramakrishnan (unpublished).

and some doubt about whether the project of rationalist ethics is in good order—a connection that seems worth understanding.[82]

XVIII Infinities in practice

Overall, then, infinite ethics seems to me an important and largely unsolved dimension of normative ethics. I haven't, here, tried to solve it myself. Rather, my aim has to distinguish the easier issues (e.g., comparing infinites to finite quantities) and from the harder ones (e.g., comparing infinities to each other, especially in the context of risk); to illustrate the hardness of the harder issues clearly and vividly (e.g., in the context of impossibility results, and via the difficulties for the six theory-types I considered); and to point at some of the implications we can already discern (centrally, with respect to the viability of simple utilitarianism; but also, possibly, with respect to Space-time-based and Person-moment-based Anonymity, and perhaps with respect to moral realism as well) even in the absence of a settled best answer.

I'll close with a few thoughts on practical implications. First: whether we are fanatical about infinity payoffs or not, and regardless of whether we have good ways of comparing them to each other, I think we should acknowledge that they are, at least, an extremely big deal—and in particular, at least as big of a deal as any finite payoff of equivalent moral "currency" (e.g., an infinity of happy lives compared to any finite number). Indeed, when I imagine a future civilization looking back on our current attitudes towards infinitely consequential actions, to me it seems plausible that they will be horrified at how *little* attention we paid to such actions relative to more local concerns; and I think it reasonable for those who aspire to prioritize in scope-sensitive ways to take the possibility of having infinite impact very seriously.

What does being serious about this look like in practice? Various philosophers in the literature have focused on the possibility that the infinity-oriented (or obsessed) should prioritize ensuring that our civilization reaches a wise and empowered future.[83] After all, if we reach such a future, we'll be able to understand the ethical issues here much more deeply. We'll also know much more about what sort of infinitely consequential actions we're able to perform, and we'll be much better able to execute on infinite projects we deem worthwhile (building hypercomputers, creating baby-

[82] We might also look for more psychological diagnoses. E.g., perhaps infinite ethics reminds us too hard of our cognitive limitations; of the ways in which our everyday morality, for all its pretension to objectivity, emerges from the needs and social dynamics of fleshy creatures on a finite planet; of how few possibilities we are in the habit of actually considering; of how big and strange the world can be. And perhaps this leaves us, if not with a rigorous argument nihilism, then with some vague sense of confusion and despair.

[83] See e.g. Bostrom (2011) and Thomas and Beckstead (2021).

universes, etc). Or, to the extent we were always performing infinitely consequential actions (for example, acausally), we'll be wiser, more skillful, and more empowered on that front, too.

Now, absent an actual theory of how to choose between infinity-involving lotteries (the type of theory we're hoping a wiser and more empowered future will supply), it's hard to get a fully rigorous argument going for focusing on reaching such a future, vs. other candidate infinity-oriented projects—e.g. converting as many people as possible to whichever religion posits the largest-cardinality heaven/hell.[84] Heuristically, though, and without surveying all the alternatives, the former looks like a fairly reasonable path forward to me—and in particular, one that seems comparatively robust to our current level of ignorance about both the empirical and philosophical issues that infinities raise. And as Bostrom (2011) and Beckstead and Thomas (2021) both note, such a path plausibly looks good (suspiciously good?) on finite-ethical grounds, too.[85]

I also want to highlight, though, a few ways in which this orientation towards the future differs from the standard sort of longtermism I mentioned earlier, which focuses specifically on the implications of our actions for the welfare of the astronomic but finite numbers of people that conventional physics suggests that the future might contain. As Bostrom (2011) and Beckstead and Thomas (2021) both note, the ultimate moral focal points at stake here (e.g., finite benefits to future generations, vs. better prospects for infinity-oriented action) are distinct, and they can come apart. Beckstead and Thomas (2021, p. 31) discuss future trade-offs between infinity-oriented projects and finite benefits in this respect (when do you stop researching the possibility of creating baby-universes and focus on building a merely finite Utopia instead?), but we can also imagine practical implications with nearer-term relevance. In particular, to me it seems plausible that infinity-oriented perspectives on the future should value marginal future resources differently than, for example, the sort of finitely-oriented total utilitarian perspective that the simplest arguments for longtermism rely on.[86] This sort of perspective values additional resources linearly, because they can create linearly more future people (and also, one assumes, linearly more optimally-efficient pleasure)—but it is much less clear that the success and value of infinity-oriented projects (e.g., creating baby universes, breaking out of simulations, exerting positive acausal influence across an infinite cosmology) scales with resources in this way (though resources do seem useful regardless).[87] So relative to a longtermism based on simple to-

[84]And this is not to mention the balance between more altruistic infinity-focused projects, and more prudentially-focused ones—e.g., avoiding hell yourself, maximizing the probability that you can live a large-cardinality life, and so on.

[85]See e.g. Ord's (2020, Chapter 8) discussion of the "long reflection."

[86]See e.g. Bostrom (2003b) for a classic example.

[87]Thanks to Nick Beckstead for suggesting this point.

talism, perhaps an infinity-focused perspective would be more focused on getting to a wise, technologically-mature, and reasonably-resourced future *at all* (rather than on a *big* version of such a future), and would comparatively unwilling to trade e.g. a guarantee of one galaxy for a .00001% chance of a billion galaxies (though obviously, it's hard to say).[88]

More generally, though, imagining a future focused on infinity-oriented projects—creating baby-universe, acausally bargaining with the aliens, etc—is just a different vision from a future focused on e.g. using the resources within our lightcone to create some large but finite amount of optimally-pleasurable experience. Both are strange; but the infinity-oriented one seems stranger—and in this sense, it's a broader reminder of just how strange a wise future's ethical priorities might get.

All in all, I think of infinite ethics as a lesson in humility: humility about how far standard ethical theory extends; humility about what priorities a wise future might bring; humility about just how big the world (both the abstract world, and the concrete world) can be, and how little we might have seen or understood. We need not be pious about such humility. Nor need we preserve or sanctify the ignorance it reflects: to the contrary, we should strive to see further, and more clearly. Still, the puzzles and problems of the infinite can be evidence about brittleness, dogmatism, overconfidence, myopia. If infinities break our ethics, we should pause, and notice our confusion, rather than pushing it under the rug. Confusion, as ever, is a clue.[89]

[88] We can speculate about the value of resources growing superlinearly, but note that this applies in the finite case, too.

[89] Thanks to Leopold Aschenbrenner, Amanda Askell, Paul Christiano, Katja Grace, Cate Hall, Evan Hubinger, Ketan Ramakrishnan, Carl Shulman, and Hayden Wilkinson for discussion of the issues in this essay. And thanks to Hilary Greaves for written comments.