A Stranger Priority?

Topics at the Outer Reaches of Effective Altruism

A thesis submitted for the degree of

Doctor of Philosophy

University of Oxford

Michaelmas Term 2022

Joseph Carlsmith

St. John's College

A Stranger Priority?
Topics at the Outer Reaches of Effective Altruism

Joseph Carlsmith
St. John's College, University of Oxford

Submitted for the degree of DPhil in Philosophy, Michaelmas Term 2022

Word count: ~67,000 without citations. ~69,000 with citations. Chapter 1: ~24,000. Chapter 2: ~17,000. Chapter 3: ~17,000.

Abstract

This thesis examines three philosophical topics relevant to the project of identifying and acting on the most effective ways of doing good: anthropic reasoning, simulation arguments, and infinite ethics. These topics are unified in their potential to disrupt the empirical and normative assumptions underlying the most straightforward case for "strong longtermism" (that is, the view that positively influencing the long-term future is the key moral priority of our time). The first chapter examines the debate between the Self-Indication Assumption ("SIA") and the Self-Sampling Assumption ("SSA") in the context of anthropic reasoning – a debate with important implications for the size of humanity's future (and, plausibly, of the universe itself). I argue that SIA is the superior view, and that the most prominent objection to SIA – the so-called "Presumptuous Philosopher" – is a bullet that we should consider biting. The second chapter formulates what I see as the strongest version of a "simulation argument" – that is, an argument that we should be highly confident that at least one of the following is true: either the ratio of simulated beings ("sims") to non-simulated beings ("non-sims") of certain types is not high, or we are sims. I distinguish between ways of making this argument that rely on empirical assumptions (for example, about the computational power available to advanced civilizations) and those that do not; I suggest that the latter are superior and independently forceful; and I explore some of the complications and uncertainties that the latter lead to. The third chapter surveys a variety of problems that infinities create for ethics, and it reflects on the implications of those problems. In particular, I argue that these problems puncture the dream of a simple, bullet-biting utilitarianism, and that they put pressure on some of the broader intuitions underlying common arguments for strong longtermism as well.

Acknowledgments

This thesis owes much to many. Acknowledgements to people who helped me on the individual essays are included in a footnote at the end of each. Here I especially want to express my gratitude to:

- My supervisors, Hilary Greaves and Jeff McMahan, for their guidance and support – and especially to Hilary for working with me so closely in the last few months before submission.
- Katja Grace, for being a true friend, as well as a patient and insightful philosophical interlocutor.
- Ketan Ramakrishnan, for many years of amazing friendship and philosophical dialogue; here's, hopefully, to many more.
- Cate Hall, for patience, love, and inspiration.
- Luke Muelhauser and Holden Karnofsky for flexibility and support with respect to my finishing the thesis.
- Iris Geens, for helping me to navigate the administrative aspects of pursuing a somewhat unorthodox schedule in completing my degree.
- William MacAskill, Luke Ding, and Roger Crisp for helping to make it possible for me to transfer to Oxford from NYU, and to pursue the schedule just mentioned.
- Samuel Scheffler, Sharon Street, and David Chalmers, for guidance and support during my time at NYU; and to Arden Koehler (and Ketan, again), for helping to make that time magical.
- John Gardner, for supervision and support during my time on the BPhil. "It is a mistake to think of duties as burdens that get in the way of living our lives well."
- Nick Bostrom, whose work has exerted a huge amount of influence on my thinking (and on my life as a whole), and who I think exemplifies a number of rare and extremely valuable philosophical virtues.
- Carl Shulman and Paul Christiano, for talking with me at length about where the crazy train goes; and for thinking through so many things.
- Toby Ord, William MacAskill, Nick Beckstead, Howie Lempel, and Jacob Trefethen, for friendship, guidance, and inspiration.
- Duncan Carlsmith, Lynn Keller, and Caroline Carlsmith, for love and support of so many kinds over the years.
- The Future of Humanity Institute, for the DPhil scholarship that funded my work.

Earlier and more informal versions of these essays were published online at handsandcities.com (though Chapter 2 has since been re-written entirely).

Contents

Introduction

This thesis examines three philosophical topics relevant to the project of identifying and acting on the most effective ways of doing good (roughly, "Effective Altruism").[1] The first is the debate between the "Self-Sampling Assumption" (SSA) and the "Self-Indication Assumption" (SIA) in the context of anthropic reasoning (Chapter 1);[2] the second is the question of whether we are living in a computer simulation (Chapter 2); and the third is the ethics of living in infinite worlds and performing infinitely impactful actions (Chapter 3).

These topics are united in their ability to disrupt empirical and normative assumptions that underly common patterns of reasoning among effective altruists – and in particular, the patterns of reasoning that lead most directly to the conclusion that positively influencing the long-term future should be our top ethical priority (roughly, "Longtermism").[3] This integrative chapter outlines the patterns of reasoning I have in mind, it summarizes the three main chapters of the thesis, and it describes some of the disruptions at stake. The

[1] See MacAskill (2019), quoted below, for a more detailed definition; and see MacAskill (2015) for a popular introduction.

[2] Bostrom (2002a) defines anthropic reasoning as the study of "observation selection effects" – i.e. cases in which "our data is filtered not only by limitations in our instrumentation but also by the precondition that somebody be there to 'have' the data yielded by the instruments" (p. 2). To me the topic seems broader, though; I tend to think of it as the attempt to grapple systematically with questions about how to simultaneously assign credences to both *de dicto* hypotheses (i.e., those about the nature of the objective world) and *de se* hypotheses (i.e., those about which observer in an objective world you are, and about what time it is), especially in cases where there are multiple observers with your evidence within a single world (see Lewis (1979) for classic discussion), and where the *de dicto* hypotheses stake involve different numbers of people. That said, the precise definition of "anthropic reasoning" does not matter for present purposes; the specific questions at stake are what count.

[3] This is a rough version of Greaves and MacAskill (2021) call "deontic strong longtermism," which they define as the claim that "In the most important decision situations facing agents today, (i) one ought to choose an option that is near-best for the far future, and (ii) one ought to choose an option that delivers much larger benefits in the far future than in the near future" (p. 26). See MacAskill (2022) for a more general introduction to longtermism.

final, concluding chapter goes into more depth about how we should proceed in light of the discussion overall.

I. Effective Altruism and Longtermism

MacAskill (2019) defines Effective Altruism (EA) as:

> (i) "the use of evidence and careful reasoning to work out how to maximize the good with a given unit of resources, tentatively understanding 'the good' in impartial welfarist terms, and
>
> (ii) the use of the findings from (i) to try to improve the world."

So defined, EA is both an intellectual and a practical project. And as a practical project, it has grown into a full-fledged social movement – one with thousands of active members, billions of dollars in funding; and a growing presence in politics and popular culture.[4]

Relative to most contemporary social movements, though, effective altruism has unusually deep roots in analytic philosophy. Key early figures, like William MacAskill and Toby Ord, were analytic philosophers inspired by Peter Singer, Derek Parfit, John Broome, and the utilitarian tradition more broadly; the movement's priorities have been notably influenced by topics and intellectual tools (population ethics, Bayesian epistemology, expected value reasoning) associated with analytic philosophy; the movement funds and is influenced by organizations doing philosophical work (notably: the Future of Humanity Institute and the Global Priorities Institute); and key decision-makers have detailed philosophical views that they treat as directly relevant to their decision-making.[5]

---

[4] See Todd (2021) for funding and membership estimates as of July 2021 (though these have since gone down quite a bit); and see Matthews (2022) for updated estimates and a more qualitative description of EA's activities in politics and beyond.

[5] See, for example, Karnofsky (2018), which focuses on a variety of thorny philosophical issues that feed directly into questions about budget allocations between causes at Open Philanthropy, one of the two largest funders of EA work.

One of the most dramatic manifestations of this philosophical influence is the EA movement's focus on *longtermism*, which MacAskill (2022) defines as "the idea that positively influencing the longterm future is a key moral priority of our time" (p. 4). Indeed, many EAs endorse "strong longtermism," on which positively influencing the long-term future is *the* key moral priority of our time – perhaps by an overwhelming margin.[6] A large portion of the movement's financial resources are on track to go to longtermist causes;[7] 80,000 hours (an EA career advising organization) focuses heavily on directing people towards longtermist careers;[8] and figures like MacAskill and Ord have pivoted to a focus on the longterm future as well.[9]

An early and influential argument for longtermism comes from Bostrom's (2003) "Astronomical Waste," which argues that an astronomical number of happy lives could be sustained in the accessible universe if humanity were to reach technological maturity and settle space, and that consequently, aggregative utilitarians should prioritize maximizing the probability that humanity creates a future of this type.[10] Beckstead (2013) offered a version of this argument that uses a broader set of assumptions, and which focuses on the quality of the "general trajectory along which our descendants develop"; and the argument

[6] See Greaves and MacAskill (2021) for more precise formulations of some different types of longtermism, and Beckstead (2013) for a framing that focuses on the "overwhelming" importance of longtermist considerations. In what follows, I'm not going to be especially strict about the distinction between longtermism and strong longtermism, because the arguments for longtermism I'm interested in are typically arguments for strong longtermism as well, and because I see the longtermist EA community as centrally motivated by these arguments in particular, even if, in practice, they only advocate for and act on weaker forms of longtermism.

[7] In the context of Open Philanthropy's budget allocation, Karnofsky (2018) writes: "We will probably recommend that a cluster of 'long-termist' buckets collectively receive the largest allocation: at least 50% of all available capital." Open Philanthropy currently accounts for the large majority of all EA funding.

[8] See their key ideas pages, https://80000hours.org/key-ideas/, for more.

[9] Ord (2020, introduction) and MacAskill (2022, introduction) both describe this shift in priorities.

[10] See also Bostrom (2013).

has received further refinement and sensitivity analysis since then – notably, in Greaves and MacAskill (2021).[11]

In all these cases, the argument for longtermism rests on (a) an empirical claim to the effect that we're in a position to positively influence the lives of a very large (though still finite) number of future people, in expectation,[12] and (b) a normative claim that, in virtue of (a), positively influencing the long-term future warrants the focus of our moral attention. The empirical claim plausibly falls out of a relatively standard scientific worldview, together with some further arguments about the tractability of longtermist interventions.[13] The normative claim falls most directly out of total utilitarianism, together with a decision theory that directs you to maximize expected value – but it can be justified, in various forms, by a variety of population axiologies (and perhaps, a variety of decision theories as well).[14]

The topics I examine in this thesis – SIA vs. SSA, simulation arguments, infinite ethics – all make trouble for these claims. In particular, SIA, SSA, and simulation arguments make trouble for the empirical scenario that longtermism directs us to focus on – namely, the scenario on which we live very early in the history of a civilization that will one day fill the accessible universe (or at least, earth's future) with life – and for a standard scientific worldview more generally. And infinite ethics both makes trouble for the views and

[11] See also e.g. Cowen (2018), Tarsney and Thomas (2020), and Mogensen (2020) for more discussion. The working paper series from the Global Priorities Institute (available at https://globalprioritiesinstitute.org/papers/) is an especially rich source for analyses of (and objections to) longtermism.

[12] Note that this doesn't require space settlement, and that I am here combining claims about the expected size of the future with claims about the tractability of influencing it.

[13] At least when coupled with the assumption that within a relatively short period of time, we will be able to drive the annual risk of extinction very low. See Thorstad (2022) for discussion of this aspect. I largely agree with Thorstad that longtermism requires endorsing what he calls a "Time of Perils" hypothesis. But I find such hypotheses plausible, especially in light of possible developments in advanced artificial intelligence (a line of reasoning Thorstad discusses).

[14] See Greaves and MacAskill (2021) for discussion.

intuitions in population ethics that most naturally undergird longtermism, and it directs our attention away from the sorts of finite (if astronomical) impacts that the argument for longtermism focuses on, and towards infinite impacts instead.

But the topics I'll discuss are also united in the *character* of the disruption to longtermism that they suggest. Let's turn to that now.

II. Riding the crazy train

Some objections to longtermism accuse it of being "too weird" – that is, of suggesting too great a deviation from our everyday thinking, and of following strange philosophical reasoning off of a cliff. Thus, for example, Matthews (2015) accuses longtermists focused on existential risk of falling victim to a "Pascal's mugging," in which a tiny probability of an extremely large pay-off dominates an expected value calculation. Alexander (2022) suggests that he would sooner abandon philosophy altogether than accept MacAskill's (2022) claims about population ethics.[15] Holt (2021) accuses Ord's (2020) logic of making us "moral slaves" to the future, and of privileging the needs of the future over the needs of the present to an alarming extent. And Torres (2021) emphasizes the strangeness of focusing on futures filled with "massive numbers of technologically enhanced digital posthumans inside huge computer simulations spread throughout our future light cone" – especially relative to more immediate global problems.

These objections are especially salient in the context of Effective Altruism's early focus on global health and on especially rigorous forms of evidence (in particular, randomized

[15] Alexander (2022) writes: "But I'm not sure I want to play the philosophy game. Maybe MacAskill can come up with some clever proof that the commitments I list above imply I have to have my eyes pecked out by angry seagulls or something. If that's true, I will just not do that, and switch to some other set of axioms. If I can't find any system of axioms that doesn't do something terrible when extended to infinity, I will just refuse to extend things to infinity" (section IV).

controlled trials) for the effectiveness of the interventions it recommended, like buying anti-malarial bednets for children in poor countries. Relative to focusing your moral attention on tiny changes in the probability of one day filling the lightcone with digital post-humans, buying bednets seems like a comparatively accessible and intuitively robust way of doing good. And the same is plausibly true, though to a lesser extent, of some of the more unusual interventions recommended by Effective Altruists, like funding corporate campaigns for cage-free eggs.[16] It takes a bit more of a deviation from many people's common-sense morality to prioritize animal welfare (at sufficient scale) over human welfare. But longtermism can easily seem a substantially further step – and one that goes quite a bit too far.

In thinking about hesitations in this broad vicinity, Ajeya Cotra offers the metaphor of a "train to crazy town":

> when the philosopher takes you to a very weird unintuitive place — and, furthermore, wants you to give up all of the other goals that on other ways of thinking about the world that aren't philosophical seem like they're worth pursuing — they're just like, stop… I sometimes think of it as a train going to crazy town, and the near-termist side is like, I'm going to get off the train before we get to the point where all we're focusing on is existential risk because of the astronomical waste argument. And then the longtermist side stays on the train, and there may be further stops.[17]

Loosely speaking, we can think of the objections canvassed above as accusing longtermist EAs of having ridden the train to crazy town too far – of having taken philosophy, as it were, too seriously. Of course, this isn't the only available diagnosis: a salient alternative is simply that longtermists are doing philosophy *wrong*, and that more careful philosophizing will reveal their mistake. I'm especially interested here, though, in objections that focus

[16] See Bollard (2019) for more.
[17] See Wiblin and Cotra (2019); Samuel (2022) also discusses Cotra's metaphor.

more directly on a sense that "this is just too weird," even if they don't try to explain where the reasoning goes wrong.

But while this sort of objection can have significant intuitive force, it also fits uneasily with effective altruism's intellectual culture and its historical relationship with philosophy. After all, even in the less controversial contexts mentioned above, Effective Altruists generally endorse a variety of claims that would seem counterintuitive and even offensive to many: for example, the idea that it can be better to work as an investment banker and donate your money than to work as a doctor;[18] or that you should take a 1% chance of saving a thousand lives over a guarantee of saving one life; or that failing to provide aid to someone in the developing world can be a moral failure comparable in significance to letting a child in front of you drown. EAs endorse these counterintuitive claims partly in virtue of the philosophical case in their favor – and indeed, Effective Altruism's *willingness* to follow philosophical logic where it leads has plausibly been one of its key strengths. Why, then, would one stop at longtermism in particular?

Indeed, in my experience (I've been involved in the Effective Altruist community since ~2013, and especially active during the past five years), a certain type of Effective Altruist prides themselves on their willingness to bite bullets of the type that objectors like Matthews, Alexander, Holt, and Torres cannot stomach. They don't want to get off of the philosophy train just because it's leading them to strange and/or uncomfortable places. Indeed, to discover something strange *but true* can be a significant contribution, if the topic is important enough; and conversely, a refusal to countenance strange truths can render your attempts at altruism ineffective and maybe even actively harmful.

[18] See MacAskill (2014) for more on this.

The topics in this thesis all offer objections to longtermism continuous with this sort of spirit. And they are continuous, as well, with other, more specific strains of longtermist thought – notably, a willingness to take *scale* seriously, including in contexts where its interaction with Bayesian epistemology and expected value reasoning leads to extreme changes in belief and action. Indeed, in my experience, the objections to longtermism I discuss are among the ones that philosophically-minded effective altruists take most seriously, even if these objections do not appear prominently in public treatments of longtermism like Ord (2020) or MacAskill (2022). Cotra, for example, explicitly mentions all three, in the interview quoted above, as playing a role in prompting her to question the view that longtermism should be Effective Altruism's sole priority.

In effect, these topics suggest, per Cotra's comments above, that the train to crazy town may not stop at longtermism – and especially not the sort of longtermism often assumed by expected altruists in practice (i.e. one focused on the finite impacts on future people in our light-cone).[19] Rather, further philosophical reflection leads to some combination of (1) even stranger places, and (2) more general perplexity. That is: rather than being "too weird," longtermism is revealed as "not weird enough" – at least, by the lights of the type of philosophizing that gave rise to it in the first place.

If true, I think this is important for a few reasons. On the one hand, to the extent we choose to stay on the train, our sense of our destination may change in practically relevant ways. Indeed, in some cases, I'll suggest, we simply run out of track – that is, our existing theories start to simply break down, rather than to give merely counterintuitive results –

[19] Thinner definitions of longtermism can in principle accommodate a wider range of points of moral focus, and how we define things like "the future" starts to matter (do impacts in the level of reality simulating you count? What about impacts on infinite afterlives?). But I'm especially interested, here, in pointing at disruptions to longtermism as commonly conceived of and practiced, rather than to its commitments as given by some particular abstract statement.

and further philosophical work is required to proceed further (though we also need to figure out how to act in the meantime – a topic worthy of philosophical analysis in itself).

On the other hand, to the extent we choose to *get off* the train, longtermism ceases to seem an obviously privileged stop, and the possibility of getting off even earlier becomes more salient. Cotra describes experiencing this sort of shift in perspective:

> I was more pro going all in on the astronomical waste argument before thinking about some of the further weird things that come up as the train keeps moving to crazy town… I sort of was like, "Okay, actually, this line of thinking takes me to a place weirder than I am comfortable with". And I sort of therefore have sympathy for people for whom the immediately previous stop was weirder than they were comfortable with, and I was more able to listen to the parts of myself that found that uncomfortable.

I won't, here, try to resolve the question of exactly how we should relate to the crazy train (though see the conclusion for a bit more discussion). At the very least, though, I think we should be clear about where it leads.

In what remains of this introduction, I'll summarize the three main chapters of the thesis, with a special emphasis on their interconnections, and their implications for longtermism in particular.[20]

III. First essay: SIA vs. SSA

The first essay in the thesis – "SIA vs. SSA" – argues that one prominent theory of anthropic reasoning (SIA) is better than another (SSA). Roughly speaking, SIA favors worlds where there are a lot of people with your evidence; whereas SSA favors worlds

[20] I've tried to keep the chapters themselves relatively self-contained (though this sometimes involves repeating a few definitions and notes that those familiar with the previous chapters wouldn't need). And the chapters themselves are not generally focused on the disruptions to longtermism I discuss in the introduction and conclusion. Rather, they are focused more directly on the specific issues (anthropic reasoning, simulation arguments, infinite ethics) in question – issues that I think of as interesting partly in virtue of their implications for longtermism, but which are also worth understanding in their own right.

where the people with your evidence are a larger fraction of the "reference class" – a (in my opinion, objectionably mysterious) set of observers that in some (in my opinion, objectionably unclear) sense you "could have been."

The most prominent objection to SIA is the so-called "Presumptuous Philosopher": SIA sometimes becomes extremely confident in a hypothesis, even despite other more empirical evidence against it, simply in virtue of that hypothesis positing a sufficient number of observers. This is indeed a problem; and SIA has other problems, too (notably: with infinities). But I argue that SSA's problems are worse, and that in the absence of an alternative superior to both SSA and SIA, the presumptuous philosopher is a bullet we should consider biting. What's more, I argue, attractive alternatives to SIA are unlikely to avoid this sort of bullet.

The debate between SIA and SSA matters for longtermism for a number of reasons. First, SSA leads to the so-called "Doomsday argument," on which we should update against worlds with highly populated futures, in virtue of their making it unlikely that we would find ourselves so early in human history.[21] If sound, this argument makes it extremely unlikely that we ever reach the sort of highly-populated future that gives longtermism its most straightforward moral heft. SIA, though, avoids this sort of doomsday argument. In this sense, my defense of SIA gives longtermism resources to defend itself against the charge that its empirical narrative places us implausibly early in humanity's history.

Unfortunately, though, SIA suggests doomsday arguments of its own – for example, that we should favor hypotheses on which the "great filter" that explains why we don't see aliens (that is, the step or steps along the trajectory from dead matter to large-scale space

[21] See Leslie (1996) for a classic statement. What's more, the strength of this update is proportional to the size of the future in question, making it harder to get high expected populations simply by increasing the number of people posited by a given hypothesis.

colonization that make completing the entire trajectory very unlikely) occurs *after* civilizations in the cosmos reach our current stage, rather than before.[22] And it plausibly favors the hypothesis that we live in a computer simulation as well -- and in particular, a simulation run by a highly-resourced civilization especially interested in simulating people like us (in this sense, the topic of the first chapter interacts with the topic of the second). Beyond this, though, an SIA-agent (at least when SIA is naively interpreted) plausibly becomes *certain* that the world is infinite, and that it contains an infinite number of observers with exactly our evidence – a conclusion that generally makes trouble for our ability to do anthropic reasoning at all (including via SIA), and which the third chapter suggests is problematic for our ethics as well.[23]

In this sense, while SIA allows us to avoid the main objection to longtermism posed by SSA, it does not, thereby, return us to the standard empirical narrative that longtermism wants to draw on. Rather – as the accusation of "presumptuousness" suggests – it leaps, from the armchair, in much more extreme directions, including ones that our ethics and epistemology are poorly equipped to handle.

IV. Second essay: Simulation arguments

The second essay in the thesis formulates and defends what I see as the strongest version of a "simulation argument" – that is, an argument that we should have high credence that one of the following is true: either the ratio of simulated beings ("sims") to non-simulated beings ("non-sims") of certain types is not high, or we are sims. In particular, I distinguish between what I call "Type 1" simulation arguments, which treat various empirical claims

[22] See Grace (2010b). Olson and Ord (2021) discuss some similar considerations.
[23] By "SIA-agent," I mean an agent who reasons in the manner suggested by SIA (and the same for SSA and an "SSA-agent"). For the sake of brevity, I will also sometimes speak of SIA as an SIA-agent – e.g., "SIA updates as follows" or "SIA expects X" (and again, same for SSA).

(for example, about the computational power available to advanced civilizations) as highly likely, and "Type 2" versions, which do not – and which focus, instead, on arguing that conditional only on most observers of a certain sort being sims (and on our being observers of this sort), the rest of our evidence is similarly likely conditional on our being sims vs. non-sims. Type 2 arguments seem to me forceful, and they avoid various objections that have been offered in the literature – for example, about problematically unstable or inconsistent relationships to our everyday beliefs.

The essay also explores some of the uncertainties and complexities we face if we begin to take Type 2 arguments seriously – and in particular, how we should adjust our credences overall in light of these arguments. In this context, I respond to a recent argument from Thomas (2021), to the effect that we should be basically *certain* that we're sims, because conditional on being non-sims, the *expected* ratio of sims to non-sims is very high. I suggest that once we incorporate the lesson of Type 2 simulation arguments into our reasoning, we need not accept this claim about the expected ratio of sims to non-sims conditional on being non-sims (though doing so is one option). Nevertheless, Thomas's argument points at a further substantive constraint that our overall credences must obey, if we accept Type 2 arguments. I describe a few other pressures that our credences must navigate, and I also explore the question of what range of cases these arguments apply to – a range that I think may be wider and stranger range than the existing literature has focused on.

Simulation arguments matter for longtermism for a few reasons. First: if we live in a simulation, it becomes much less likely that our descendants go on to create the sort of highly-populated future longtermists are focused on (since simulating this sort of future would require large amounts of computational resources) – so it becomes very pressing to resolve whether we should adopt views like the one that Thomas (2021) suggests; and

pressing, more generally, to understand what sort of overall credence we should assign to the ratio of sims to non-sims being high.[24] What's more, if we're sims, questions about how our actions might impact the world outside the simulation, including the future choices of our simulators, take on greater importance.[25]

Even if we maintain credence on being non-sims, though, simulation arguments tell us that in those scenarios, the ratio of sims to non-sims (of certain kinds) is almost guaranteed to be very low, thereby imposing an additional constraint on the sorts of futures we might hope to causally impact – one with implications for whether we should expect to make it to a much more advanced stage of technological development at all (since failing to do so is one salient way to avoid creating lots of sims).

Whether these considerations suffice to actually *block* longtermism (as opposed to merely lower the expected value of longtermist interventions) isn't clear.[26] At the very least, though, taking simulation arguments seriously involves a disorienting revision to our everyday picture of our empirical situation – one that we might expect to have implications for our efforts at effective altruism as well.

Simulation arguments also interact in interesting ways with the other topics I discuss. For one thing, I mentioned above, SIA plausibly bolsters the case for worlds with large

[24] Also, even if we are in a simulation with a big future, we would then need to apply simulation arguments *again* to the ratio of "double sims" (e.g., sims being simulated by people in a sim) vs. "single sims" (sims being simulated by non-sims) within our simulation. And conditional on that ratio being high, then simulation arguments will direct us to place high credence on being double sims (and then triple sims, and so on). So the overall upshot that it's very unlikely that we're in a position to causally affect a big future filled with sims will persist. But I don't discuss this issue in depth.

[25] Green (2020) suggests, for example, that investigation of simulation hypotheses may itself pose an existential risk, since it may increase the probability that our simulators shut us down.

[26] It depends on a variety of factors, notably: our overall probability that we're sims, the types of futures we can affect if we're non-sims, the probability that we have a big future ahead of us even if we are sims, what sort of decision theory we accept, what sort of opportunities for nearer-term altruism are available, and our tolerance for focusing on especially small probabilities of especially large impacts. I don't try to tease apart all of these factors, but see Tomasik (2016) for discussion of some of the complexities involved.

numbers of simulations, since those worlds have more observers with evidence broadly similar to ours. What's more, though, simulation arguments rest on principles that make most sense in finite worlds – but SIA leads to certainty that the world is infinite, thereby complicating the question of whether and how simulation arguments apply. What's more, though, if we are in a sim (and especially, in a sim housed in an infinite unsimulated reality), this makes it more plausible that our actions can have infinite causal consequences, since the physical laws suggesting that this is impossible may not hold in the reality simulating us.

V. Third essay: Infinite ethics and the utilitarian dream

The third and final main essay of the thesis surveys a variety of problems that infinities create for ethics, and it reflects on the implications of those problems. I argue that some classic issues in the literature – notably, those to do with ethical indifference to merely finite amount of influence on the world – are minor in comparison with more pressing problems: notably, those to do with impossibility results in generating ordinal rankings; extreme difficulties generating plausible principles for choosing between lotteries over infinite outcomes; and additional looming complexities if we have to bring uncountable infinities into the picture as well.

I argue, in particular, that these problems puncture the dream of a simple, bullet-biting utilitarianism; and that they put some pressure on broader intuitions underlying common arguments for longtermism as well – notably, intuitions on which someone's location in time does not intrinsically matter, morally. I also briefly touch on whether these problems constitute an argument against moral realism (my

answer: maybe), and on what taking infinite ethics seriously in practice might look like – especially in the absence of solutions of the problems I've discussed.

Infinite ethics matters for longtermism for a number of reasons. In particular: in my experience, those drawn to the utilitarian dream that I think infinites puncture tend to be unusually drawn towards longtermism – and in particular, strong longtermism -- as well, and to view its extreme and counterintuitive conclusions as precisely the sort of bullet that they (unlike so many others) are willing to bite. Indeed, arguments like Bostrom's "Astronomical Waste" proceed most naturally from a combination of (a) a simple total utilitarian ethic, and (b) a willingness to let tiny probabilities of extreme impacts dominate one's overall decision-making.[27] But infinite ethics makes serious trouble for (a) – and indeed, for population ethics more generally – while also suggesting that, if we're going in for (b), the sorts of extreme impacts that longtermism focuses on (e.g., protecting the potential for some astronomical but finite number of flourishing lives within the affectable universe) aren't sufficiently extreme to warrant our priority, relative to the possibility of *infinite* impacts.

[27] Exactly how much willingness of this sort is required is a topic of some dispute amongst longtermists. Thus, Greaves and MacAskill (2021) write: "We regard [problems with focusing on small probabilities of enormous payoffs] as one of the most plausible ways in which the argument for strong longtermism might fail. Our view is that at present, the question cannot be confidently settled, since research into the possibility of a non-fanatical decision theory is currently embryonic. However, initial results suggest that avoiding fanaticism might come at too high a price" (p. 25). That is, a key aspect of their response to worries about fanaticism is simply that plausibly we should accept fanaticism after all. That said, they also mention the possibility that the probabilities in question are not problematically small (though what counts as "problematically small" is not defined), especially if we focus on societal decision-making rather than individual decision-making. I find this response unsatisfying, though, partly because effective altruism often focuses on individual decision-making in other contexts, and partly because it seems like a defensive move, made in the absence of some principled way of identifying which sorts of small probabilities are problematically small vs. acceptably small (one possible response here is to say "we mean for the notion of 'problematically small' to stand in for whatever line a viable non-fanatical decision theory will allow us to draw between e.g. the rationality of wearing a seat-belt and the rationality of fanaticism, if there is such a theory" (thanks to Hilary Greaves for discussion) – but the question of whether such a viable decision theory is available remains open, and one reason to think not is that the relevant line seems likely to be unprincipled).

Of course, the future may be the best place for infinite impacts, too (though it's not the only possibility: you could, for example, focus on converting present-day people to your highest-probability religion with an infinite afterlife).[28] But prioritizing it for that reason is not the mainstream longtermist narrative, and I suggest in the chapter that an infinity-focused longtermism may lead to substantively different places than the more traditional version.

VI. Conclusion

Taken together, then, these three essays all point towards serious (and interconnected) complications to the empirical and normative picture underlying mainstream presentations of the longtermist worldview. What's more, these complications are continuous in philosophical aesthetic with the sorts of arguments and impulses that most naturally undergird that worldview. In effect, they all suggest that big worlds can lead to strange philosophical places fast, especially in the context of epistemic and normative principles that aspire to treat scale seriously and quantitatively; that longtermism is by no means obvious or privileged terminus of this strangeness; and that often, these tools start to actively break down as we push them to their limits.

In the concluding chapter of the thesis, I step back and reflect more on what this means for how we should proceed, now. In particular, I suggest that while we should proceed humbly in light of the uncertainties the thesis's discussion suggests, we should not ignore the clues – however strange -- that this sort of reflection offers as to the empirical or normative considerations that may ultimately govern the upshot of an effective altruist ethic. That is, I don't think we should treat the crazy train as "too crazy"; but we should

[28] Here I am assuming that the effects in the afterlife are not in the "long-term future" as traditionally understood. That said, it's true that this sort of intervention wouldn't be focused on the "near-term," either.

ride it with taste, caution about robustness of the conclusions we are drawing, and attention to whether those conclusions have left us truly convinced. Indeed, I suggest that this sort of approach may ultimately recover and re-emphasize one of the key intermediate priorities of standard longtermism: namely, to make it to a much *wiser* and more empowered civilization, with superior capacity to understand and act on issues of this scope and strangeness.[29] In a few cases, though (for example, simulation hypotheses), there may also be implications that cannot be so easily punted to some much wiser set of future people, because the relevant people may not exist.

[29] See Ord (2020, Chapter 8) on the "long reflection."

Chapter 1

SIA vs. SSA

I. Introduction

This essay argues that one prominent approach to anthropic reasoning (the "Self-Indication Assumption" or "SIA") is better than another (the "Self-Sampling Assumption" or "SSA").[30] Consider:

> *God's extreme coin toss*: You wake up alone in a white room. There's a message written on the wall: "I, God, tossed a fair coin. If it came up heads, I created one person in a room like this. If it came up tails, I created a million people, also in rooms like this." Conditional on the message being true, what should your credence be that the coin landed heads?

SIA says: ~one in a million. SSA says: one in two. (I explain why, in each case, below.)

I open with this case because it's one of the worst for SIA, the approach I favor. In particular: we can construct more scientific analogs, in which SIA becomes extremely confident in a given hypothesis, simply in virtue of that hypothesis positing many more

[30] Bostrom (2002a) defines anthropic reasoning as the study of "observation selection effects" – i.e. cases in which "our data is filtered not only by limitations in our instrumentation but also by the precondition that somebody be there to 'have' the data yielded by the instruments" (p. 2). To me the topic seems broader, though; I tend to think of it as the attempt to grapple systematically with questions about how to simultaneously assign credences to both *de dicto* hypotheses (i.e., those about the nature of the objective world) and *de se* hypotheses (i.e., those about which observer in an objective world you are, and about what time it is), especially in cases where there are multiple observers with your evidence within a single world (see Lewis (1979) for classic discussion). That said, the precise definition does not matter for present purposes.

observers. Various philosophers treat this implication (known as the "Presumptuous Philosopher") as a basically decisive objection to SIA.[31]

But I think that the objections to SSA are stronger, and that in the absence of an alternative approach superior to both SSA *and* SIA ("Anthropic Theory X"), the Presumptuous Philosopher is a bullet we should consider biting.[32]

I begin by explaining how SSA and SIA work, motivating them both in contrast to an approach that attempts to simply stick with your prior, and contrasting various "just-so" stories used to illustrate/justify their logic. I then discuss a variety of objections to SSA. In particular, SSA implies:

- scientific presumptuousness comparable to SIA's;
- extreme confidence that fair coins, yet to be flipped, will land heads;
- expecting a rolling boulder to leap out of the way of a puppy, depending solely on whether you form the intention to create lots of observers if it doesn't;
- an unexplained and indeterminate ontology of "reference classes" (I also argue that Bostrom's (2002a) attempt to use this ontology to avoid SSA's unattractive implications fails);
- sensitivity to differences that seem epistemically irrelevant (like whether an observer you know you're not is killed vs. never created);

[31] See e.g. Leslie (1996); Ćirković (2001); Bostrom (2002a); and Arntzenius and Dorr (2016).
[32] Here the name "Anthropic Theory X" is a reference to Parfit's (1984) use of "Theory X" to denote the elusive theory of population ethics that would give us all of what we want. Thanks to Nick Beckstead for suggesting this.

- strong updates towards solipsism.

After an aside about the complexities of arguments about betting in anthropics, I then turn to a discussion of SIA's problems. In particular:

- I suggest that we should be least *open* to biting the bullet about the Presumptuous Philosopher.

- Infinities are a problem for SIA, but they're a problem for SSA, too (though perhaps not quite so bad of one), and for anthropic reasoning more generally.

- SIA "learns something" when it wakes up in Sleeping Beauty; but if you think of yourself as a person-moment, I think this becomes more intuitive.

- Given some values and decision theories, SIA implies inconsistencies between the policy you'd want to commit to *ex ante* and your behavior *ex post*. But again, so does SSA. What's more, these inconsistencies are common in other contexts, and if you're worried about them, I suggest addressing them at the level of decision theory rather than epistemology (while not conflating the two).

I close by discussing whether we should expect to find an alternative superior to both SSA *and* SIA — the "Anthropic Theory X" above. My current answer is: maybe, but Anthropic Theory X should probably keep SIA's good implications (like "thirding" in Sleeping Beauty). And the good implications seem closely tied to (some of) the bad. I also briefly touch on SIA's real-world implications, which, in light of SIA's problems with infinities, seem notably unclear.

II. A bit of set-up

Cases like *God's extreme coin toss* involve assigning credences to both *de dicto* hypotheses (i.e. hypotheses about what sort of objective world exists) and *de se* hypotheses (i.e. hypotheses about which observer you are in an objective world, and at what time – I'll often call this your "location").[33] Thus, in the case above, the hypothesis that God's coin landed tails is *de dicto*; and the hypothesis that God's coin landed tails AND you are the person in room 1 at time *t* is *de se*.

Let's use "objective world" to refer to a fully-specific *de dicto* hypothesis, and "centered world" to refer to the pairing of an objective world with a subject and a time within that world.[34] Further, let's call all of a subject's evidence at a given time her "epistemic situation."[35] Importantly, two different times within the same person's life can have the same epistemic situation (for example, if I put your brain in a vat, give you a tea-drinking experience at t_1, then wipe your memory and give you the exact same experience at t_2 as well).[36] I want the anthropic principles I discuss to treat uncertainty about whether you are

---

[33] See Lewis (1979) for a classic discussion.

[34] Elga (2004) and Manley (unpublished) also call this a "predicament." Formally, we can think of a centered world as a triple $<w, s, t>$ where w is an objective world, s is a subject in that world, and t is a time.

[35] Here I'm mostly following the set-up in Manley (unpublished). Also following Manley, I'll generally assume that a subject's epistemic situation includes her apparent-memories and qualitative experiences, construed in an internalist way such that molecule-for-molecule copies of a given subject have the same epistemic situation. I think that attempting to make greater room for externalism of various kinds could well make a difference to the analysis (see e.g. Weatherson (2005), p. 617, for some discussion), but I won't attempt that here. Similarly, and again following Manley (see p. 5, fn. 6), I am going to try to avoid wading into the issue of exactly what sort of epistemic access you have to what your evidence is. Manley suggests that we need not get too hung up on a subject's epistemic access to her conformity with doxastic norms, since she can implement them (or fail to implement them) without being a position to know that she is doing so. I am hopeful that something like this is true, and that even if subjects in the cases I discuss are not in a position to know that they are adjusting their credences in the right way, we can proceed with our analysis of how they should be adjusting their credences regardless. That said, insofar as both SIA and SSA involve counting the number of people in your epistemic situation (in a given objective world), it will indeed be difficult to implement these rules, in practice, without representing to yourself what your epistemic situation is. I will generally assume that this step does not introduce extra problems or uncertainties (at least at the level of normative analysis – even if it makes it harder to know whether you're in conformity with the norms in question), but I acknowledge that there is room for complexity here, and that a more complete analysis may need to grapple more directly with questions about your epistemic access to what your evidence is. However, and importantly, such questions are not, as far as I can tell, a source of disagreement *between* SIA and SSA – and it is the disagreement between the two views that is my central interest here.

[36] While both objective worlds and centered worlds are maximally specific (I define them this way partly to reflect the convention used in e.g. Elga (2004) and Manley (unpublished), and partly to avoid confusions that sometimes arise when the question is left open), for convenience in what follows I will often speak of them

Bob-at-t_1 or Bob-at-t_2 the same way they treat uncertainty about whether you are Bob-at-t_1 or Sam-at-t_1, so I'll formulate those principles in a manner that focuses not on persons or observers per se, as epistemic subjects, but on what Bostrom (2002a) calls "*observer-moments*" – i.e., observers at a given time (though for simplicity, I'll often drop the reference to moments and speak more loosely about observers/people).[37]

Let's call two centered worlds "similar" if they share an objective world and if they are centered on subjects with the same epistemic situation. Both SIA and SSA (at least as I'll understand them) take for granted the following indifference principle (adapted from Elga (2004)):

> *Indifference*: Similar centered worlds deserve equal credence.[38]

---

in a more coarse-grained way. Thus, I will sometimes refer to the "tails world," where in fact I mean to refer to a large set of objective worlds where the coin came up tails (e.g., worlds with mountains vs. forests surrounding the white rooms, where the coin bounced in *x* vs. *y* way, and so on). Similarly, I will say that in e.g. the tails world in God's extreme coin toss, there are a million people in your epistemic situation. In fact, though, what really matters is that conditional on tails, your maximally specific epistemic situation (e.g., the precise shade of white the room is painted, the specific clothes you're wearing, the specific pattern of sensation on your feet) is equally likely conditional on being in any of the rooms. That is, conditional on tails, God need not make every person identical. But your specific characteristics don't provide any information about what room you're in. This simplification will not affect the debate between SIA and SSA (I'll explain why, in a footnote, once I've introduced SIA and SSA below). However, if you'd like, you can reformulate cases like God's extreme coin toss to make it unnecessary – e.g., you can imagine a version of where there are only two possible maximally-specific objective worlds in play, and where the tails world involves exact copies of all the people. And you're free to imagine that I've offered such a version in each case.

[37] The difference between observers and observer-moments is reflected, in Bostrom's work, via the difference between the self-sampling assumption (SSA) and the *strong* self-sampling assumption (SSSA). My version of SSA is as equivalent to Bostrom's SSSA. Note that in focusing on "observer-moments" here, I don't mean to take a stand on broader questions about "time-slice rationality" in the sense of e.g. Hedden (2015), except insofar as doing so is necessary to reproduce the specific credal dynamics I discuss. That said, I do think that puzzles of the sort created by the cases I discuss provide one motivation for more time-slice focused conceptions of rationality as a whole (see e.g. section 2.2 of Hedden (2015)).

[38] The ideal way of setting up this principle isn't completely clear, especially in the context of cases involving infinite subjects in the same evidential situation, but the issues for *Indifference* aren't part of the disagreement between SIA and SSA, so I'm not going to try to resolve them here -- see Weatherson (2005) for various objections and complications, and Manley (unpublished) for a more formal treatment designed to handle some of them. I'll note, though, that Elga's original formulation focuses on centered worlds that are "subjectively indistinguishable," which is a subtly different notion from having the same epistemic situation, and which leads to possible intransitivity in cases where A is indistinguishable from B, and B from C, but not A from C (see Weatherson (2005) and Manley (unpublished, fn. 5) for more). And note, as well, that the viability of *Indifference* does not depend on whether you are always in a position to know what your epistemic situation is, or which observers share it (see Manley (unpublished), fn. 6).

Suppose, for example, that God has created ten exact copies of you in ten white rooms, labeled 1-10, but where you can't see the labels. What should your credence be that you're in room 1? Plausibly: 10%. What about room 2? Also 10%. It would seem strange, in a case like this, to prefer some rooms, epistemically, over others: to be at e.g. 13% on room 1, but 52% on room 2. And we can make various other arguments for *Indifference* as well.[39]

Despite its common-sense credentials, though, *Indifference* is controversial and sometimes problematic – as are related variants. However, I find something in the ballpark quite plausible; and because the principle is not at issue in the debate between SIA and SSA, I won't focus on it here.

The cases I'll consider will generally involve a prior over objective worlds, corresponding to the fair coin in *God's extreme coin toss*.[40] Here I am following others in the literature who treat SIA and SSA as distinct approaches to updating a prior probability distribution that the two theories otherwise agree on.[41] In cases involving coin tosses that determine which of two objective worlds get created, this probability distribution (at least on the standard set-up) is set by the coin, at 50% on each world. In other, messier cases, it's set by some more general build-up of empirical evidence -- for example, about the respective plausibility of different cosmological theories, the likelihood that humanity goes extinct within the next few centuries, or the likelihood of some strange physical event like a

[39] See Elga (2004) for a classic discussion, and Weatherson (2005) for some objections.
[40] I use the term "prior" here to reflect the fact that there is (plausibly) some further updating yet to do – namely, the updating suggested by SIA or SSA (in the next section, I also discuss the possibility of not updating the prior at all – a possibility that I view as unpromising). I do not mean, though, to refer to a more fundamental type of prior, corresponding to credences that do not yet incorporate *any* of your evidence (what's sometimes called an "ur prior" – see e.g. Meacham (2016) for discussion). That said, the prior in question is still *hypothetical*, in the sense that it need not correspond to an epistemic state you have ever occupied – or even, could rationally occupy. Indeed, the prior need not incorporate the information that *you* exist – despite the fact that epistemic states that fail to reflect this information are plausibly irrational. See Manley (unpublished), p. 15, and Isaacs et al (2021), p. 7, for more on priors of this kind.
[41] See e.g. Manley (unpublished), p. 20-22 and Issacs et al (2021), p. 20, who formulate SIA and SSA in a manner very similar to the way I do.

wounded deer suddenly appearing before you.[42] I'll sometimes call the evidence that informs the prior probability distribution "non-anthropic evidence," to distinguish its epistemic role from the specific, additional updates suggested by SIA and SSA.

How should we understand this prior? It's not, itself, an "ur prior" – that is, a more fundamental type of prior, reflecting probabilities that do not yet incorporate *any* of your evidence.[43] Rather, my default is to view it as the product of *updating* an ur prior (over objective worlds) on all of the evidence that makes no appeal to *your* location in particular – what we might call your "*de dicto*" evidence.[44] That is, if you're in some evidential situation *x*, you would get to the prior I have in mind by updating your ur prior on the fact that *some observer-moment* is in evidential situation *x*, and then SIA and SSA tell you where to

[42] Thus, for example, Bostrom's (2002, p. 124) characterization of the Presumptuous Philosopher involves the empirical cosmological considerations – which Bostrom imagines come from considering the futuristic science of "super-duper symmetry" – being indifferent between two theories, T1 and T2. These considerations would set the prior in question. And similarly, discussion of the Doomsday Argument (discussed in Section VI below) often contrasts the dramatic verdict of the argument (i.e., that humanity is extremely unlikely to have a highly-populated future) with what a naïve assessment of the empirical evidence might've suggested – an assessment on which a highly populated future would've seemed reasonably plausible. See also Bostrom (2002a, p. 143) for the wounded deer example, in which Bostrom imagines that "the prior probability of a wounded deer limping by [the cave of Adam and Eve] is one in ten thousand, say" – a probability that the sort of "telekinesis" I discuss in section VII below might work to counteract. The general thought here seems to be that there is some probability distribution that a common-sensical scientist would have about some domain, absent exposure to the discourse about anthropics (see also Sean Carroll's comments at 27:24 of his (2020) conversation with Bostrom, in which he imagines a first step of "give theories prior probabilities by how elegant or reasonable they seem" – and then updating according to SSA or SIA from there). And much of the hand-wringing about anthropics I discuss below is prompted by hesitations about letting armchair anthropic reasoning move us far away from this (supposedly) more common-sensical epistemic position.

[43] After all, the sorts of naïve, common-sensical probabilities I above – on things like different cosmologies, human extinction, the wounded deer walking up, and so on – *do* incorporate various types of evidence. See Meacham (2016) for more on the concept of "ur priors," which can be understood in a variety of different ways ("common candidates include: the credences a subject should have if she had no evidence, a subject's initial credences, a subject's evidential standards, and any function that plays the right diachronic role" (p. 1-2)). Of course, to the extent we're relying on some notion of ur priors, there is the further (quite fundamental) question about where such priors come from and what standards (if any, beyond probabilistic coherence) are applicable to them. I don't have answers to these questions, and I don't think we need such answers to proceed with the sort of analysis this chapter engages in (indeed, if we needed solid stories about fundamental priors before engaging in a broadly Bayesian approach to some philosophical issue, much of Bayesian-inspired epistemology would be stymied). For what it's worth, though, my own leading candidate for an "ur prior" appeals to some notion of the *simplicity* of different hypotheses. See Carlsmith (2021b) for more on this, and Carlsmith (2021c) for more on a possible application to anthropics in particular.

[44] See Manley (unpublished) and Issacs et al (2021) for more on this sort of story. On such a story, the "non-anthropic evidence" I discussed above would be just: the de dicto evidence.

go from there – that is, how to incorporate the fact that "*I* am in evidential situation *x*" (in the next section, I discuss why this further step is necessary).

This understanding of the prior raises various questions, which I won't attempt to get to fully resolve here (though see footnote for some discussion).[45] Indeed, I don't view this particular story about priors as especially central to my main argument; and more generally, I'm open to the idea that the debate between SIA and SSA is best formulated and understood in a quite different way – for example, as a debate about what your ur prior should be.[46] What matters most, in my opinion, is the ultimate *verdicts* of the theories

[45] One issue concerns conditionalization. In particular, and following Manley (unpublished) and Isaacs et al (2021), I am not assuming that posterior probabilities can be calculated simply by conditionalizing your ur prior on *all* of your evidence (including your *de se* evidence). Rather, the set-up I'm using assumes that you start with an ur prior over objective worlds, conditionalize that ur prior on your *de dicto* evidence, and then proceed according to SIA/SSA from there – i.e., updating your credences on objective worlds *again* in proportion to either the number of people in your epistemic situation (on SIA), or the fraction of the people in your reference class that people in your epistemic situation make up (on SSA), and then distributing this credence amongst *de se* hypotheses according to *Indifference* above (in this sense, I am assuming that your final credences should be determined not just by your evidence, but also by your ur prior and your choice between SIA and SSA). This procedure is admittedly somewhat cumbersome, and I would be happy to discover a way of making it compatible with the simpler and more theoretically unified procedure of simply conditionalizing on the ur prior (indeed, I think that SIA, at least, can be made to satisfy this constraint). However, for reasons I explain in the next footnote, I think that attempting to formulate the *disagreement* between SIA and SSA entirely at the level of ur priors assumes answers to some questions that some advocates of SSA want to leave open (specifically, whether the reference class can vary depending on your epistemic situation), so I don't attempt it here.

I also want to acknowledge a few other outstanding issues. First: I'm also mostly passing over questions about the required relationship between your credences and the objective chances (see e.g. Thomas (2021a) for some discussion) – an issue that I think could well bear on the right way to formulate the most plausible positions in the vicinity of SIA and SSA. Second: in messy cases like the Doomsday Argument or the Presumptuous Philosopher, I'm leaving ambiguous the specific role of the ur prior vs. the *de dicto* evidence in producing the specific prior at work in the case. Third: I'm formulating this set-up specifically with SIA and SSA in mind – if we tried to incorporate a broader range of views into the discussion, the set-up might well require alteration.

[46] Thus, for example, you could imagine formulating the disagreement between SIA and SSA as centrally about whether the *de se* evidence that "I exist" is more likely, on the ur prior, conditional on objective worlds with a larger number of observer-moments – where SIA says yes, SSA says no, but they agree on what to do once this ur prior is in place (namely, split your credence between *de se* hypotheses per *Indifference*, and conditionalize on your evidence in the standard way). But various versions of SSA are difficult to formulate in this framework. In particular, this framework effectively assumes that we're using "all observer moments" as the reference class for SSA (see section IV for more on what I mean by "reference class"), and that this reference class does not vary depending on the type of observer-moment you end up being – assumptions that some advocates of SSA, like Bostrom, do not take for granted. Thus, for example, suppose that God flips a coin. If heads, he creates one human and nine chimps. If tails, he creates ten humans. And suppose you wake up as a human. SIA is at $1/11^{th}$ on heads, here (see next section for the calculation leading to that verdict). Some versions of SSA, though, are at 1/2 on heads, because your reference class need not include chimps (see section IV for more on this sort of case). But this difference is very hard to square with

in question in the types of cases I'll discuss (for example, *God's extreme coin toss*) – verdicts produced by the underlying way in which SIA and SSA favor different types of objective worlds.[47] I've found the formulation I use a productive way of isolating and analyzing the disagreement that leads to this difference in verdicts, and similar formulations are common in the literature, but if superior alternatives are available (alternatives that also reproduce the verdicts and forms of favoritism in question), all the better – I expect much of what I say in what follows to apply regardless.

III. SIA and SSA

Equipped with a prior of this kind, we can characterize the difference between SIA and SSA as follows: SIA updates the prior in proportion to the *number* of observer-moments in your epistemic situation in a given objective world.[48] SSA, by contrast, updates it in proportion to the *fraction* of the observer-moments-with-your-epistemic-situation that are in your *reference class*, in that world. (What's a reference class? Let's hold off on that for now – I'll say more about it soon, and it's easiest to understand by looking at how it functions in practice. In general, though, and unless I say otherwise, I'll assume that the reference class consists of all the observer-moments discussed in a given case, except for God.)

---

formulating the disagreement between SIA and SSA as a disagreement about whether "I exist" is more likely, on the ur prior, in worlds with more observer-moments. On such a formulation, one would expect SIA and SSA to both be at 1/2 on heads conditional only on "I exist" (since both heads and tails imply the same number of observer-moments), and then to end up in the same place, as well, once they incorporate "I am a human." (You can make SIA and SSA give the same verdict, here, if you force SIA to use a reference class in a manner similar to SSA – but for reasons I discuss in Section V below, I don't want to do this.) Of course, the fact that some versions of SSA are difficult to formulate in a manner consistent with straightforward conditionalization on the ur prior is a mark against their plausibility (see section X for more discussion). But I prefer not to assume one side of that debate at this stage in the chapter's set-up.

[47] Specifically, as I'll discuss in the next section, SIA favors worlds with more observer-moments in your epistemic situation, and SSA favors worlds in which the observer-moments in your epistemic situation are a larger fraction of some other set of observer-moments – the "reference class." Whether this favoritism occurs at the level of priors, or via differences in how one updates one's priors, seems to me of secondary importance.

[48] I don't have a specific method in mind for counting observer moments, but regardless of how you do so, ten minutes of observer-time should have 10x the number as one minute -- and it's the ratios that will matter in what follows.

More formally, suppose that you have non-zero credence on objective worlds O_1 through O_w; suppose that n is a function indicating the number of observer-moments in a given world in your epistemic situation; r is a function indicating the number of observer-moments in a given world in your reference class; p_r is your prior over objective worlds, and p is your posterior credence after your anthropic updating. Then, to calculate your posterior credence on a given objective world O_x, SIA says:

$$\text{SIA: } p(O_x) = \frac{p_r(O_x) \times n(O_x)}{\sum_{i=1}^{w} p_r(O_i) \times n(O_i)}$$

And SSA says that:

$$\text{SSA: } p(O_x) = \frac{p_r(O_x) \times \frac{n(O_x)}{r(O_x)}}{\sum_{i=1}^{w} p_r(O_i) \times \frac{n(O_i)}{r(O_i)}}$$

(In what follows, I'll mostly present calculations of this form in terms of odds-ratios, which I find easier to think about.[49])

Having made this update, then per *Indifference*, both theories apportion their new credence on each objective world equally amongst the centered worlds compatible with that objective world.[50]

---

[49] The odds ratio between a hypothesis H_1 and a hypothesis H_2 is just the ratio of their probabilities: i.e., $p(H_1)$: $p(H_2)$. So SIA says that given two objective worlds O_x and O_y, the posterior odds ratio $p(O_x)$: $p(O_y)$ is $p_r(O_x) \times n(O_x)$: $p(O_y) \times n(O_y)$, and SSA says that it's $p_r(O_x) \times \frac{n(O_x)}{r(O_x)}$: $p(O_y) \times \frac{n(O_y)}{r(O_y)}$.

[50] For similar definitions, see Isaac's et al (2021, p. 20), and Manley (unpublished, p. 21). The relationship between these definitions and Bostrom's (2002a) original usage is somewhat more complicated, partly because Bostrom's qualitative definitions of the two principles (SSA: "One should reason as if one were a random sample from the set of all observers in one's reference class" (p. 57); SIA: "Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist" (p. 66)) are notably unclear in their implications, and partly because Bostrom thinks of SIA as an assumption that you can *add* to SSA, rather than as an alternative – a way of thinking that I discuss below in the context of what I call "R-SIA," but which seems to me unhelpful. And still others in the literature may use the terms differently again. But the names are not important for present purposes.

To see how this works, consider the following case:

> *God's coin toss with equal numbers*: God tosses a fair coin, and he creates ten people in white rooms either way. If heads, he gives one person a red jacket, and the rest, blue jackets. If tails, he gives everyone red jackets. You wake up in a white room and see that you have a red jacket. What should your credence be on heads?

Here, both SSA and SIA give the same verdict, but for different reasons. SIA reasons: "Well, my prior is 1:1. But on tails, there are 10x the number of people in my epistemic situation — e.g., red-jacketed people. So, I update 10:1 in favor of tails. So, 1/11th on heads."

SSA, by contrast, reasons: "Well, my prior is 1:1. But on heads, the people in my epistemic situation are a smaller fraction of the reference class. In particular, on heads, the red-jacketed people are 1/10, but on tails, they're 10/10 (assuming that we don't include God). Thus, I update the prior 10:1 in favor of tails. So, 1/11th on heads."

Having made this update about the objective world, SIA and SSA then both think of themselves as 1/11th likely to be each of the red-jacketed people.[51]

---

[51] We're now in a better position to explain why the simplification I noted in footnote 36 above -- that is, the simplification that allows us to speak about objective worlds and epistemic situations in coarse-grained ways (e.g., talking about large sets objective worlds/epistemic situations as though they are single objective worlds/epistemic situations), despite the fact that objective worlds/epistemic situations are supposed to be maximally specific – makes no difference to the debate between SIA and SSA. Suppose, for example, that there are exactly two maximally specific epistemic situations, A and B. And consider two cases:

1. God tosses a coin. If heads, he creates one person with A. If tails, he creates two people with A.
2. God tosses a coin. If heads, he creates one person; if tails, he creates two people. Then, for each person he creates, he tosses another coin to decide whether to put them in A or B.

Further, let's assume that except for the epistemic situations, heads and tails worlds are maximally specific and similar. And let's suppose that you wake up in epistemic situation A.

Case 1 is maximally fine-grained and specific. That is, there are only two objective worlds – Heads-A and Tails-AA – with a prior odds ratio of 1:1. Case 2, by contrast, features four possible objective worlds -- Heads-A, Tails-AA, Tails-AB and Tails-BA – with a prior odds ratio of 2:1:1:1. (If this is unclear, first

This case is useful to keep in mind, because it's a kind of "square one" for anthropics. In particular, it helps answer the question: "Why are we updating the prior at all? Why not just stick with ½?" A key answer is: if you don't update the prior, and instead skip straight to apportioning your prior credence amongst the red-jacketed people in each world (per *Indifference*), you get this case wrong. Thus, you reason: "Well, 50% on heads. So 50% that I'm the one red-jacketed heads-world person. And 50% on tails, so 5%, for each of the tails-world people, that I'm them." Here, very plausibly, you've failed to learn the right thing from your red jacket. In particular, you've failed to learn that the coin probably landed tails.

---

consider the six possible worlds in play prior to learning that you have A: namely, Heads-A, Heads-B, Tails-AA, Tails-AB, Tails-BA, and Tails-BB. The prior odds ratio here is 2:2:1:1:1:1 – so when you cross off Heads-B and Heads-BB upon learning that you have A, you're left with 2:~~2~~:1:1:1:~~1~~.)

But SIA and SSA both treat these cases identically. In Case 1, SIA updates towards the tails world (because it has 2x the A-people) to 1:2, whereas SSA sticks with the prior of 1:1 (because A-people are the same fraction of the reference class either way). In Case 2, by contrast, SIA updates towards the Tails-AA world, for a final odds ratio of 2:2:1:1 (and so, 1/3rd on heads); whereas SSA updates *against* the Tails-AB and Tails-BA worlds (since A-havers are only ½ of the reference class in those worlds), for a final odds ratio of 2:1:$\frac{1}{2}$: $\frac{1}{2}$ (and so, ½ on heads). Thus, you get the classic "thirding" and "halfing" behaviors regardless.

This dynamic generalizes to a version of the case where we simply specify that God creates one person if heads, and two people if tails. Even if A is a highly unlikely epistemic position for any person to end up in (say, one of out a million possibilities), as long as it's *equally likely* that any of the people are in A, then we get the same results from SIA and SSA if we just coarse-grain our description and say that the tails worlds involves two people in your epistemic position, vs. splitting it out into the many different fine-grained possibilities and running the more detailed calculation.

That said, if we were considering a wider array of views in anthropics, things would get more complicated (and to the extent we understand the "prior" in these cases as the product of updating an ur prior over objective worlds on our de dicto evidence, we might also need to be more careful about whether we're talking about fine-grained or coarse-grained objective worlds). Views like Neal (2006), for example – views which try to avoid anthropic updating at all, but which also use a very fine-grained notion of epistemic situation -- treat cases 1 and 2 very differently. These views just stick with the prior in the cases above. Thus, they end up as 1/2-ers in Case 1, and 2/5-ers in Case 2 – but their similarity to thirders increases quickly the more possible epistemic situations are in the mix. These views, though, give the wrong verdicts in versions of *God's coin toss with equal numbers* (discussed in the main text) where we specify that all of the red and blue-jacketed people have exactly the same epistemic situation except for their jacket colors. And relatedly, in universes that are sufficiently big that there is at least one of observer in every physically possible epistemic situation (for example, universes that feature sufficiently large numbers of "freak observers" generated by random fluctuations of physical conditions), these views can't update towards their epistemic situation being common vs. rare – a failure that Bostrom (2002b, section I) argues threatens their ability to believe that their scientific observations reflect the universe's actual conditions (since, e.g., even if the universe's real conditions are very different, *some* observer will make the observations you are making).

To illustrate why you need to learn this, suppose you haven't yet seen your jacket. Then, surely, you should be 50-50, and split your credence equally amongst all the people in each world. Then suppose you see that your jacket is red. This observation was much more likely conditional on tails rather than heads. Thus, it seems like basic Bayesianism to update.[52]

IV. Storytelling

SIA and SSA both get this "square one" right; but they differ in their verdicts in other cases (like *God's extreme coin toss* above). Before getting to those cases, though, can we say anything about what SIA and SSA are doing on a qualitative level? What underlying story about our epistemic and metaphysical situation motivates these approaches, and their differences?

In my opinion, it's unclear in both cases. Indeed, both approaches can be given multiple qualitative rationales, and none of the rationales on offer seem to me especially satisfying. I don't think of the theories as *defined* by such rationales, though. Rather, what matters is the quantitative mechanics of how they update. That said, such rationales can be helpful, at least, for pumping some intuition about SIA and SSA-like reasoning. Here I'll sketch two that I find especially useful in this respect.[53]

Let's start with SIA. On one story, SIA treats you as a *specific possible person-in-your-epistemic-situation,* who might or might not have existed, even conditional on there being *someone* in

[52] Now, some views actually endorse not making this update: these views that say you should be a 50-50 before you see that your jacket is red, and 50-50 afterwards, too. See Halpern (2006), Meacham (2008), and Neal (2006) (assuming a version of the case where all the red epistemic positions are exact copies). In my opinion, though, this isn't to their credit. See Manley (unpublished, section 3) and Bostrom (2002b) for more on the problems here. Regardless, though, my aim here is to debate the merits of SIA vs. SSA in particular; and they agree on this case.

[53] See also Bostrom's (2002a, p. 122) "heavenly messenger" analogy, and Manley's (unpublished, section 5) analogy with marbles.

that situation. And it thinks of worlds as pulling some number of people-in-your-epistemic-situation from the "hat" of the platonic realm. That is, and put fancifully: before you were created with a red jacket in a white room, God said to himself "I need to create X number of people with red jackets in white rooms." He then reached into the platonic realm and groped around randomly in the area labeled "people with red jackets in white rooms." You were there, in your red jacket, huddled together with some untold number of other red-jacketed souls (a number large enough that God can draw as many people as he wants out, without meaningfully altering the probability that he draws you). But yet, by a stroke of very serious luck, God's great fingers wrapped around your ghostly non-body. You got pulled, as the other red-jacketed souls looked on. Thus, you found yourself alive. It was, indeed, quite a lottery-win. But importantly, it was more likely in worlds where God reached in more times.

This story has problems. Saliently, for example, it makes most sense if we imagine that the population of red-jacketed, white-roomed souls in the platonic realm is finite (but very large). If that population is infinite instead, then it becomes much less clear how to think about the probability that you get drawn. And it feels laden with suspicious metaphysical baggage more generally.

That said, it's not the only story available. Thus, for example, we can also think of SIA as treating you as a random sample from the people-in-your-epistemic-situation who *might* exist, weighted by the probability that they *do* exist.[54] However, I think this story may run into instabilities (see footnote), so I tend to stick with the story above.[55]

[54] See e.g. Olson and Ord (2021) for an example of this sort of framing.

[55] On the "random sample from possible people" story, SIA is centrally about a kind of principle of indifference about who you are, applied to all the people you might be (including people in different possible worlds), but weighted by probability that those people exist (thanks to Katja Grace for suggesting formulations in this vein). That is, an SIA-agent notices that they exist in their epistemic situation, then says:

Let's turn to SSA's story -- or at least, a certain version of it. Unlike SIA, SSA assumes that you exist in any world where *someone* is in your epistemic situation. That is, it imagines that once God decides to create a world where *someone* will have your memories, experiences, etc, he goes looking for you in the hat of possible people, and then "inserts you" into that world – regardless of the how many people-in-your-epistemic-situation it contains. Importantly, though, when God creates you and inserts you into the world, he does so in a particular way: namely, he makes you a random member of some "reference class" *other than* the people in your epistemic situation. (What sort of reference class? It's not clear. I'll return to this problem later.) That is, in any given world containing someone in your epistemic situation, SSA imagines that God hones in on some set of people you "could have been" — even though for some of them, you know you *aren't* — and then makes one of them, at random, you.

"Ok, who am I?" They then looks at all the people in that epistemic situation who *might* exist, and tries to not-be-opinionated-with-no-reason about people like that who are equally likely to *actually* exist. Thus, in a case like *Sleeping Beauty* (described in section VIII below – here I'll assume familiarity), the SIA-agent reasons: "Ok, I've woken up. So, which person-moment am I? Well, I might be Heads-Monday, I might be Tails-Monday, and I might be Tails-Tuesday. Heads-Monday is 50% likely to exist, and Tails-Monday and Tails-Tuesday are both 50%, too. So, they're all equally likely to exist. Thus, with no special reason to favor any of them, I split my credence evenly: 1/3rd on each. Thus, I'm 1/3rd likely to be in a Heads world."

And if the original coin had been weighted, say, 25% on Heads, and 75% on tails, the SIA agent would adjust accordingly, to make sure that they stay equally likely to be equally-likely-to-exist people-in-its-epistemic-situation: "Ok, Heads-Monday is only 25% likely to exist. Whereas Tails-Monday and Tails-Tuesday are both 75% likely. If they were all equally likely to exist, I'd be 1/3rd on each; but actually, the tails people are 3x more likely to exist than the heads person. So, upweighting each of those people by 3x, I end up at 1/7th on Heads-Monday, and 3/7ths of each of Tails-Monday and Tails-Tuesday. Hence, 1/7th on Heads."

But this sort of framing raises the question of why you don't update *again,* once you've decided that tails is more likely than heads. That is, granted that tails is 2/3rds, it's now 2/3rds that Tails-Monday and Tails-Tuesday exist, and only 1/3rd that Heads-Monday does. So why doesn't the SIA agent reason as follows? "Ah, actually, the tails people are each twice as likely to exist as the heads people. So, instead of 1/3rd on each, I'll be 2/5ths on each of the tails people, and 1/5th on the heads person. But now, actually, it looks like tails is 4/5ths and heads is 1/5th. So actually, instead of 1/5th on heads person, I'll be 1/9th…" and so on, until they become certain of tails. So while I think this framing has advantages over the "possible people in the platonic hat" framing, it also risks a kind of instability, and/or a convergence towards false certainty.

Of course, we can just specify that SIA only makes an update of this kind once (see e.g. Manley (unpublished)), but at the level of qualitative rationale, it's not clear to me what justifies this. Partly for that reason, I currently don't lean heavily on a story of this kind.

(Bostrom (2002a), an advocate for SSA, is at pains to emphasize that SSA doesn't involve positing any actual physical mechanism — akin to a time-traveling stork — for randomly distributing souls across members of the reference class. Rather, SSA is just a way of assigning credences. That said, we might wonder what would *make* such a way of assigning credences track the truth, absent such a mechanism — and Bostrom does not offer an account.)

To see where the reference class aspect of SSA starts to make an important difference, consider this variation on *God's coin toss with equal numbers*:

> *God's coin toss with chimpanzees*: God tosses a fair coin. If heads, he creates one person in a white room, and nine chimpanzees in the jungle. If tails, he creates ten people in white rooms. You wake up as a human in a white room. What should your credence be on heads?

Here, SIA reasons as it did in the original case, when people in blue jackets were in the role of the chimps. Thus, and using the language of the story above: "On tails, there are 10x the number of people in my epistemic situation, and so 10x the number of 'draws' from the hat of the platonic realm, and so 10x the chance of drawing me. Thus, I update 10:1 in favor of tails: 1/11th on heads."

SSA, though, gives different answers depending on whether you count chimpanzees in the jungle as in your reference class or not. Thus, and using the language of the story above, it reasons: "Well, I know that both heads and tails create at least one person in my epistemic situation, so I'll assume that I would've existed in either of those worlds no matter what. What's more, if heads, then I was randomly inserted into a reference class of nine chimps in the jungle, and one human in a white room. Thus, on heads, it would have been only

10% likely that I find myself in my epistemic situation; I would have expected to be a chimp instead. By contrast, on tails, I was randomly inserted into a reference class consisting entirely of humans in white rooms, so it would have been 100% that I find myself in my epistemic situation. So I update 1:10 in favor of tails: 1/11th on heads."

By contrast, if SSA *doesn't* count chimps in the jungle as in your reference class, then it reasons as before: "It's 100%, on either heads or tails, that I'd find myself a human in a white room, so I don't update at all: 50%." Thus, whether you "could have been a chimp," in the sense relevant to the reference class, ends up a crucial question. And the same will be true, in other cases, of whether you could have been a bacterium, an ant, a genetically engineered post-human, a brain emulation, a nano-bot, an alien, and so on. Indeed, as I'll discuss below in the context of the Doomsday Argument, on SSA, our views about the very future of humanity plausibly hinge on such questions.

(Note that the "could have" here need not be the "could" of metaphysical possibility. But somehow, on SSA, the reference class needs to be such as to license surprise, conditional on heads and chimps-in-the-reference-class, that you find yourself a human — and if there's *no* sense in which you could've been a chimpanzee, it's unclear why you'd be surprised that you're not one. Regardless, I'll continue to use "could have been a chimpanzee" in whatever sense is required to justify such surprise — I'm happy for the sense to be minimal.)

Perhaps you're wondering: can SSA just use the simple and attractive reference class of "people in my epistemic situation" (call this the "minimal" reference class)? No, it can't, because then it loses the ability to update the prior at all with respect to worlds that feature at least one person in your epistemic situation, since the percentage of observers in your

reference class who are in your epistemic situation will always be 100%.[56] Thus, with a red jacket in *God's coin toss with equal numbers* above, it ends up at 50% on heads, and 50% on tails — even though on heads, only one person out of ten had a red jacket, but on tails, everyone did. Thus, it falls afoul of basic Bayesianism in the way discussed above.

In my opinion, SIA gets around this problem elegantly. It honors the "minimal reference class" intuition that what matters here is *people in your epistemic situation*, and that focusing attention elsewhere is arbitrary. But those people don't need to be a fraction of some larger (and hence more arbitrary) set, in order for their numbers given tails vs. heads to provide information. Rather, the bare fact that there are *more people in your epistemic situation* given tails vs. heads is enough.

V. SIA without reference classes

I want to pause here to explicitly distinguish between the version of SIA I just presented, and a version sometimes presented in the literature — a version I consider less attractive, even though extensionally equivalent.

I'll call the version I have in mind "Reference-class-SIA" (or R-SIA). Like SSA, R-SIA thinks of you as a member of some reference class. But it also thinks that *you are more likely to exist if more members of your reference class exist*. That is, it imagines that God populates the *reference class* with souls, by pulling them out of the possible-people-in-that-reference-class hat, then throwing them randomly into the bodies of reference class people. And since you are in that hat, more people in the reference class means more chances for you

[56] Indeed, a central problem motivating Bostrom is that he thinks that if you can't make updates like favoring tails in cases like *God's coin toss with equal numbers*, then you can't do science given the possibility of "big worlds" — that is, worlds where, for any given observation, there is some observer who makes it, even if it is false. In comparing big world hypotheses, Bostrom thinks, we need to be able to favor the worlds in which a larger *fraction* of observers in the relevant reference class makes the observation in question — but the minimal reference class makes this impossible. See his (2002b) for more.

to get pulled. Thus, unlike SIA as presented above, which scales the prior in proportion to the number of people in your epistemic situation (call this *n*, as above), R-SIA scales the prior in proportion to the number of people in your reference class (call this *r*, as above).

If you *combine* R-SIA with SSA, you get SIA as I presented it above. That is, if you first scale in proportion to *r,* and then in proportion of *n*/*r*, the *r* cancels out, and *n* is the only thing that matters.[57] Thus, tacking R-SIA onto SSA eliminates the problematic dependence on the reference class that SSA otherwise implies: whatever reference class you choose, you get the same answer. And it exactly cancels SSA's other counterintuitive implications, like the Doomsday Argument (discussed below). The image, here, is of what I'll call an "inflate-and-claw-back" dynamic: that is, first you *inflate* your credence on worlds with many people in your reference class, via R-SIA, and then you *claw it back* in proportion to the fraction of those people who are in your epistemic situation, via SSA. You're left with the version of SIA above.

But I think this framing undersells SIA's appeal. The appeal of SIA with respect to reference classes isn't that you can pick whatever reference class you want. Rather, it's that you don't have to think in terms of the dubious concept of reference classes at all; you can just think entirely in terms of "people in your epistemic situation" — that is, in terms of *n*. In this sense, R-SIA + SSA feels to me like its ceding too much ground to SSA's reference-class focused ontology.

Similarly, the appeal of SIA with respect to SSA's counterintuitive implications isn't that it adds just the right additional extreme update to counteract SSA's other extreme update.

---

[57] That is, on R-SIA + SSA, the posterior odds ratio $p(O_x)$: $p(O_y)$ is $p_r(O_x) \times r(O_x) \times \frac{n(o_x)}{r(o_x)}$: $p_r(O_y) \times r(O_y) \times \frac{n(o_y)}{r(o_y)}$, which reduces to $p_r(O_x) \times n(O_x)$: $p(O_y) \times n(O_y)$.

It's not that SIA lunges a million miles left, to balance out SSA's lunging a million miles right. Rather, the appeal is that (at least in doomsday-like cases) SIA doesn't lunge at all. In this sense, SIA as I presented it above feels to me simpler than R-SIA + SSA, and in that sense, more attractive.

VI. The inevitability of presumptuousness

Let's turn, now, to a more in-depth evaluation of which of SIA or SSA is better. I emphasize "better," here, both because I don't think of either of these views as especially attractive in an absolute sense; and they aren't the only two anthropic approaches available. I focus on them because they are two quite prominent approaches; because I find myself opinionated about their comparative merits; and because they illustrate basic tensions that any plausible approach to anthropics will have to navigate.[58] At the end of the chapter, I'll return to the question of what other options might be available.

To get an initial flavor of some trade-offs between SIA and SSA, let's look at the basic dialectic surrounding two versions of *God's extreme coin toss*:

> *God's extreme coin toss with jackets:* God flips a fair coin. If heads, he creates one person with a red jacket. If tails, he creates one person with a red jacket, and a million people with blue jackets.
>
> - *Darkness*: God keeps the lights in all the rooms off. You wake up in darkness and can't see your jacket. What should your credence be on heads?
>
> - *Light+Red*: God keeps the lights in all the rooms on. You wake up and see that you have a red jacket. What should your credence be on heads?

[58] I discuss some other candidate views in the footnotes of section XVI below.

(I'll assume, for simplicity, that the SSA reference class here is "people," and excludes God. I talk about fancier reference-class footwork below.)

In *Darkness*, SIA is extremely confident that the coin landed tails, because waking up at all is a million-to-one update towards tails. SSA, by contrast, is 50-50: you're the same fraction of the reference class either way. In *Light+Red*, by contrast, SIA is 50-50: there's only one person in your epistemic situation in each world. SSA, by contrast, is extremely confident that the coin landed heads. On heads, after all, you're 100% of the reference class; but on tails, you're a tiny sliver.

Thus, both views imply an extreme level of confidence in some version of the case. Indeed, various prominent problem cases for each view basically amount to a restatement of this fact.[59] I'll suggest, though, that while such confidence can be made counterintuitive in both cases, SSA's version is worse.

Let's start with the *Presumptuous Philosopher*:

> *The Presumptuous Philosopher*: There are two cosmological theories, T1 and T2, both of which posit a finite world. According to T1, there are a trillion observers. According to T2, there are a trillion *trillion* observers. The (non-anthropic) empirical evidence is indifferent between these theories, and the scientists are preparing to run a cheap experiment that will settle the question. However, a philosopher who accepts SIA argues that this experiment is not necessary, since T2 is a trillion times more likely to be correct.

[59] For example, in Bostrom's work, the Presumptuous Philosopher is basically just a restatement of SIA's verdict in *Darkness*. The Doomsday Argument, Adam and Eve, UN++, and Quantum Joe are all basically just restatements of SSA's verdict in *Light+Red*.

It seems strange, in this case, for the philosopher to be so confident about the true cosmology, simply in virtue of the number of observers at stake. After all, isn't cosmology centrally an *empirical* science? Don't we need to look at the world, to know how many observers there are? Extreme confidence about a question like that, reached from the armchair, seems unjustified.[60]

Indeed, we can make the presumptuous philosopher look even more foolish. We can imagine, for example, that the empirical evidence favors T1 a thousand to one. Still, the philosopher bets hard against its prediction about the next experiment, and in favor of T2.[61] Unsurprisingly to the scientists, she loses. Now the evidence favors T1 a million to one. Broke, she mortgages her house to bet again, on the next experiment. Again, she loses. At this point, the scientists are feeling sorry for her. "The presumptuous philosopher," Bostrom and Ćirković (2003) write, "is making a fool of [her]self" (p. 9).

Many people basically get off the boat with SIA at this point: presumptuousness of this kind is just too much to accept. And I agree that this is a very bad result. For now, though, after nit-picking a little bit about the example as presented, I want to argue that SSA's implications are (a) at least as bad (and presumptuous, unscientific, etc), and (b) worse.

Let's start with the nit-picks. First, it's important to the example that we don't know enough about our location in the universe to rule out being the extra observers in question. Suppose, for example, that the cosmologies in question work like this. In both cases, earth sits at the center of a giant, finite sphere of space. On T1, the sphere is

---

[60] Of course, the philosopher could argue the scientists are ignoring the empirical evidence that they find themselves existing. And more broadly, depending on how we understand the sort of update that different anthropic principles are suggesting, the line between empirical evidence and other sorts of evidence may not be especially clean. Still, the basic intuition that "this philosophical view is suggesting a suspiciously major revision to our naïve empirical worldview" persists regardless.

[61] Though note that betting in anthropics implicates a number of additional issues – see section XIII below.

smaller, and so has more not-at-the-center observers; and on T2, it's bigger, and so has more. In both cases, though, all these non-earth observers can tell that they're not in the center. In this case, SIA doesn't care about the observer count, because our epistemic situation precludes being a not-at-the-center observer. Thus, SIA follows the science: just do the experiment. Of course, not all cosmologies allow us to locate ourselves in this way, so it's possible to make versions of the thought experiment that work: hence the label "nit-pick." But it's a nit-pick that will become relevant in what follows.

My second nit-pick is that pretty clearly, you shouldn't be 100% on a given theory of anthropics. So while it's true that these sorts of credences are implied by SIA, it's not clear that they're implied by a reasonable-person's epistemic relationship to SIA.[62] Thus, for example, if you had 10% credence on SSA, and 90% on SIA, then on a naïve way of incorporating your uncertainty over your anthropic theories, you might end up at 10% on T1 after the non-anthropic evidence starts favoring it, and only 90% on T2. This won't necessarily save you from betting with the scientists, but it's a less extreme distribution overall.[63]

My third nit-pick is that I think it's at least a bit unfair, in a debate about the right credences to have in this scenario, to imagine the philosopher losing all these bets. That is, if SIA is right, then it's not the case that the non-anthropic empirical evidence is the only relevant guide as to what will result from the experiment — the fact that you exist at all, in your epistemic situation, is also itself a massive update. Indeed, if we take this update seriously, then to even end up in a situation in which the non-anthropic empirical evidence favors T1 by a factor of a thousand seems like it might be positing something very weird

[62] See Carl Shulman's comment on Grace (2011).
[63] That said, I don't, here, want to get too far into the question of the right way to assign credences given uncertainty about the right approach to the anthropics – a question that may well get quite complicated.

having happened — something we might expect, naively, to induce the type of uncertainty about our anthropic theory that I just mentioned. And more broadly, to SIA, imagining the philosopher losing these bets is similar to imagining someone betting hard against Bob winning the lottery, and losing twice in a row: by hypothesis, it almost certainly wouldn't happen.[64]

All that said, I don't think these nit-picks, on their own, really take the bite out of the case. The more important point is that SSA gets bitten too.

To see this, return to the version of the case just discussed, in which on both theories, earth is at the center of a giant sphere of space, but on T2, and the sphere and observer count are bigger. Let's say the non-anthropic empirical evidence, here, is 50-50. As mentioned above, now SIA just follows the science. SSA, though, suddenly jumps into the role of presumptuous philosopher.[65] After all, on T2 and SSA, we are a much smaller fraction of the reference class, and it was hence much less likely that we find ourselves in our epistemic position, on earth. Thus, SSA mortgages the house, goes broke betting with the cosmologists, and so on — just like SIA did in the version of the case where we didn't know our location.[66]

Indeed: SSA, famously, can lead to the "Doomsday Argument," which is structurally analogous to the case just given.[67] Thus, suppose that you've narrowed down your picture of the future to two hypotheses: *doom soon*, which says that humanity will go extinct after there have been ~200 billion humans, and *doom later*, which says that humanity will survive

[64] That said, after the first loss, we should be getting uncertain about our model of the lottery as well: e.g., something fishy is going on with Bob…
[65] Or at least, it does if we use a reference class that includes the non-earth observers — more on trying to avoid that below.
[66] See Grace (2011) for more on the parallels here.
[67] See e.g. Leslie (1996) for classic discussion, and Bostrom (2002a) for in-depth analysis.

and flourish long enough for ~200 trillion humans to live instead. On the basis of the available non-anthropic empirical evidence (for example, about the level of extinction risk from nuclear war, pandemics, and so on), you start out with 10% on doom soon, and 90% on doom later. But if you use "humans" as the reference class, then you make a hard SSA update in favor of doom soon, and become virtually certain of it (including mortgaging the house, betting with the scientists, etc) — since in a doom soon world, you are a much larger fraction of the reference class as a whole.[68] Whether this argument actually goes through in the real world, even conditional on SSA, is a further question (it depends, in particular, on what reference class we use, and what other hypotheses are in play).[69] But the bare possibility of making such an argument, on SSA, suggests that un-presumptuousness isn't exactly SSA's strong suit, either.

Is SIA's version of presumptuousness somehow worse? I don't see much reason to think so in principle. In both cases, anthropic reasoning ends up making an important and sometimes extreme difference to how we treat otherwise-live empirical hypotheses. I think it's reasonable to be hesitant about this at the level of overall epistemology, especially given our ongoing confusion about many issues in the vicinity. But as an implication of any given anthropic theory, I think it's to be expected: if your anthropic reasoning can get you to one-in-a-million in *Darkness* or in *Light and Red*, it should be able to do so in real-world analogs as well; and the number of rooms, observers, and so forth in question can get large quickly.

---

[68] The usual doomsday argument appeals to your "birth rank," but I don't think this is necessary: what matters is the number of people in the reference class who aren't in your epistemic situation.

[69] Thus, for example, if you think that the people in the *doom later* world are all brain emulations, but that brain emulations aren't in the reference class, then you can avoid doomsday arguments (Bostrom (2002a, p. 171) expresses interest in this sort of response). But avoiding doomsday arguments by specifically choosing a reference class that avoids them seems to me objectionably ad hoc – especially in light of the problems with reference classes I discuss below.

VII. Fair coins and rolling boulders

So SIA and SSA are both presumptuous in some cases. However: I also think that SSA's brand of presumptuousness is worse. In particular, (a) it involves ~certainty that some fair coins, not yet flipped, will land heads, and (b) it implies that something reminiscent of telekinesis is possible.[70]

Let's start with (a). Consider the following variant on *Light + Red* above, adapted from Bostrom (2007, p. 67):

> *The red-jacketed high-roller*: You wake up in a room with a red jacket. God appears before you. He says: "I created one person with a red jacket: you. Now, if this fair coin comes up tails, I won't create any more people. If it comes up heads, I'll create a million people with blue jackets." What should your credence be that the coin will land heads?

One might think: 50% — after all, it's a fair coin, and you're about to watch it get flipped. But SSA is close to certain that the coin will land heads: after all, if it lands tails, then you would be a tiny fraction of the reference class, and would've been overwhelmingly likely to be a blue jacketed, post-coin-flip person instead. Thus, in effect, SSA treats your existence pre coin-flip, with a red jacket, as an Oracle-like prediction that the coin will land heads. And at least naively, it bets, mortgages the house, and so on accordingly.

Of course, the question of when, exactly, one's credences should align with the objective chances is complicated, and I'm not going to dive into the issue much here.[71] And once can imagine arguing that SIA, too, says weird things about fair coins. After all, SIA is

[70] As above, I'm going to assume in these cases that we're using a reference class that includes the relevant large group of observers.
[71] See, for example, Lewis (1980) and Thomas (2021a) for discussion.

highly confident on tails, in *Darkness* above (though SIA's response here is: that's because I *learned something* from the fact that I exist).

Still, SSA's verdict here seems like a really bad result to me — and in particular, a *worse* result than the *Presumptuous Philosopher*. The strange thing about the *Presumptuous Philosopher* is that anthropic reasoning leads to extreme confidence about *some* empirical hypothesis. The strange thing about the *Red-Jacketed High-Roller* is that it leads to extreme confidence *that a fair coin, not yet a flipped, will land heads*. The latter is a species of the former, but it seems to me substantially more problematic.

But SSA's implications get worse. Consider:

> *Save the puppy*: You wake up in a red jacket. In front of you is a puppy. Next to you is a button that will create a trillion more people, all wearing blue jackets. No one else exists. A giant boulder is rolling inexorably towards the puppy, and it will crush the puppy with very high probability. You want to save the puppy, but you can't reach it. However, you accept SSA, and you understand the power of reference classes. So you make a firm commitment: if the boulder doesn't swerve away from the puppy, you will press the button; otherwise, you won't. Should you now expect the boulder to swerve, and the puppy to live?[72]

This seems like a very strange expectation. Or more specifically: it seems like this type of move – an attempt at what we might think of as "evidential telekinesis" – *won't work*. That puppy is (almost certainly) dead meat. But SSA expects the puppy to live. After all, if the puppy dies, then there will be a trillion extra blue-jacketed people, and you would've been

[72] See Bostrom's (2001) *UN++* and *Lazy Adam* for related examples.

a tiny fraction of the reference class. This seems to me *substantially* more presumptuous than thinking that anthropic reasoning can provide strong evidence about cosmology.

VIII. Does SIA imply telekinesis, too?

Does SIA imply telekinesis, too? After all, SIA updates towards worlds with lots of people in your epistemic situation. Can we use a button that makes lots of those people in particular to gain telekinetic influence?

In a sense: yes. But I don't think SIA's version of this is as bad as SSA's version. Here's the sort of case I have in mind:

> *Save the puppy as SIA*: The boulder is rolling towards the puppy. You set up a machine that will make a trillion copies of you-in-a-sealed-white-room (with your memories) if and only if the boulder swerves. Having set up the machine, you prepare to enter a sealed white room.[73] Should you expect the boulder to swerve, and the puppy to live?

Here, SIA still answers no. To see why, though, recall that the epistemic subjects we want our anthropic principles to apply to are specifically observer-moments, rather than observers-over-time. This distinction is important here, and it's important in some other classic cases, too. Consider, for example, *Sleeping Beauty*, which is basically just a reformulation of *God's coin toss*-type cases, but with person-moments instead:

> *Sleeping Beauty*: Beauty goes to sleep on Sunday night. After she goes to sleep, a fair coin is flipped. If heads, she is woken up once, on Monday. If tails, she is woken up twice: first on Monday, then on Tuesday. However, if tails, Beauty's memories

[73] Below I discuss SIA's verdicts once you're already in the white room, but I want to start with this version.

> are altered on Monday night, such that her awakening on Tuesday is subjectively indistinguishable from her awakening on Monday. When Beauty wakes up, what should her credence be that the coin landed heads?

Here, you need to talk about person-moments to capture Beauty's uncertainty, conditional on tails, about whether it's Monday or Tuesday. With that set-up in place, though, SIA and SSA treat this case in the same way they treat *God's coin toss*.

With the notion of person-moments at the fore, we can see that it's true, in *Save the puppy as SIA,* that on SIA, *once you're in a sealed white room*, you should expect the boulder to have swerved. After all, there are many more *person-moments-in-your-epistemic-situation* in worlds where the boulder swerved than otherwise. But this doesn't mean that *prior* to entering the sealed white room, you should expect swerving. Rather, you should expect the boulder to behave normally.

The dynamic, here, is precisely analogous to the way in which, on Sunday, SIA says that Beauty should be 1/2 on heads; but *once she wakes up*, she should change to 1/3rd. This change can seem counterintuitive, since it can seem like she didn't gain any new information. But that's precisely the intuition that SIA denies. On SIA, when Beauty wakes up, she shouldn't think of herself as Beauty-the-agent-over-time, who was guaranteed to wake up regardless. Rather, she should think of herself as a particular person-moment-in-this-epistemic-situation — a moment that might or might not have existed, and which is more likely to have existed conditional on tails. We can debate whether this is a reasonable way to think, but it's core to the SIA narrative I offered above.

And note, too, that on Wednesday, after the whole experiment is over, Beauty should be *back* at 50% on Heads, just like she was on Sunday. This is because there aren't any

extra person-moments-in-a-Wednesday-like-epistemic-situation conditional on heads vs. tails. This means that you can't use the number of awakenings to e.g. cause Beauty, on Wednesday, to expect to have won the lottery, just by waking her up a zillion times on Monday and Tuesday if she does. And the same holds for *Save the Puppy as SIA*. Yes, you can get the people-in-the-sealed-white-rooms to expect the boulder to have swerved. But if, before letting any of them leave, you kill off all of them except one, or if you make them into Beauty-style awakenings instead of separate people, then the person who leaves the room and re-emerges into the harsh light of this thought experiment should expect (with very high confidence) to see the puppy dead.[74]

That said, it's true that, if you don't do any killing, and instead let *everyone* out of their rooms no matter what, then you and all your copies will expect to find the puppy alive. And thus, from the perspective of the person-moment who *hasn't* yet gone into the room, it's predictable in advance that the person *in the room* (your next person-moment) is going to become extremely confident that something that isn't going to happen (e.g., the swerve) has happened; and when *they* (or more specifically, their next person-moment) emerges into the daylight, they're in for a grisly surprise. On SIA, the reason for this mistake is just that this person-moment-in-the-room has in fact found itself in an extremely unlikely situation — namely, the situation of having been created, despite so few person-moments-in-this-situation getting created. In this sense, your future person-moment-in-a-white-room is like the number 672, who finds itself having been pulled from a bucket of 1-1000 — and who therefore updates, wrongly but reasonably, towards worlds where there were lots of pulls (and hence more chances to pull 672). In worlds with only one pull, *someone* has to make this type of mistake.

---

[74] This, in my opinion, is also the thing to say about Yudkowsky's (2009) "Anthropic Trilemma."

Shouldn't SIA be able to guard against this type of mistake, though? For example, shouldn't you be able to send a message to your likely future self: "don't believe what SIA is telling you; the puppy is almost certainly dead." Well, whether you want to send a message like that, and force your future self to believe it, depends on who you are counting as your future self — or more specifically, whose beliefs you care about making accurate. In particular, if you only care about accuracy of your original self — e.g., the original series of person-moments — rather than the copies, then it's true that you want to propagate forward a "puppy is dead" belief, because the original self ends up almost exclusively in worlds where the puppy is dead. But this move has a side effect: it makes a trillion copies of you (plus the original) wrong, in some much-more-than-one-in-a-trillion number of cases. Thus, if you care about the copies, too, you can't just go writing notes like that casually. Indeed, most of your epistemic influence, if we weight by both probability *and* number of minds-influenced, is funneled towards worlds where the puppy is alive.

That said, once we're bringing questions about which copies you care about, and what sorts of pre-commitments (epistemic and otherwise) you want to make, we're getting into more complicated territory. I'll discuss this territory a bit more in section XIII below. For now, I'm happy to acknowledge that SIA verdicts about this sort of case aren't entirely innocuous. But I think SSA's are worse. In particular, an SSA-agent *actively expects* to be able to use their intentions with respect to the button to save the puppy. Indeed, on evidential decision theory, they will pay to get access to this sort of button.[75] And in

[75] Evidential decision theory chooses the action such that having performed that action would be the best news, whereas causal decision theory chooses the action that has the best causal effects. Thus, if in *Save the Puppy* you construe "form the intention to press the button conditional on the puppy not being saved" as an action, an agent that accepts EDT and SSA will evaluate this action as high expected (evidential) value, since conditional on performing it, your credence in the puppy surviving should be high (whereas conditional on not performing it, your credence in the puppy surviving should be much lower). So given access to the button, you form such an intention. And given the *option* to access the button -- for example, by paying $100

general, they will start celebrating the puppy's imminent survival even before they enter any kind of sealed-white-room. From an SIA-agent's perspective, by contrast, buttons and sealed-white-rooms like this are much less appealing. Exactly what type of not-appealing depends on factors like whether this agent cares about the accuracy of you-copy beliefs, but in general, even if in some cases an SIA-agent ends up expecting telekinesis to have worked, it will generally avoid, or at least not seek out, cases where it forms this belief. An SSA-agent, by contrast, believes in telekinesis ahead of time, and (at least on EDT) goes around looking to use it.

Overall, then, my current view is that (a) SSA is ~as cosmologically presumptuous as SIA, but that (b) SSA endorses stranger stuff, in other cases, in a worse way. On their own, then, I'd be inclined to view the cases thus far as favoring SIA overall. But there's also more to say.

IX. Against reference classes

Let's turn to the issue of reference classes. In this section, I explain my objections to them. In the next section, I talk about why we shouldn't follow Bostrom in invoking them to try to get around the cases above.

What do I object to about reference classes? Well, for one thing, they are mysterious. That is, it's not clear what sort of story about the world undergirds their role in SSA's

---

-- then you expect, conditional on paying, to choose to form the intention above, thereby resulting in a low probability on the puppy's death. Whereas if you don't pay, you don't get access to the button, don't form this intention, and you end up with a high credence that the puppy dies. So conditional on paying to have access to the button, your credence on the puppy dying is low; whereas conditional on not paying to have access to the button, your credence on the puppy dying is high. Thus, if that difference is worth more than $100 to you, you choose to pay.

This sort of reasoning doesn't work on causal decision theory, though. On causal decision theory, either the boulder is going to swerve, or it isn't – and even with access to the button, your intention does not causally affect this (even though your forming the intention, on SSA, changes the probability you should assign to swerving). So access to the button isn't practically useful, to CDT – it's just a method of managing the news.

epistemology. I told a story earlier about God picking some set of people in a world, and randomly "making you one of them," but advocates of SSA don't actually believe such a story. What do they believe, though? What could even make it the case that the "true" reference class is one thing vs. another?

I'm not sure. As far as I can tell, at least for Bostrom the notion of reference class is centrally justified via its utility in getting the answers he wants from various anthropics cases. Indeed, as I'll discuss in the next section, Bostrom demonstrates a lot of willingness to alter the reference class he focuses on in pursuit of those answers. But we are left with very little sense of what constraints — if any — such alterations need, in principle, to obey.

Indeed, in the absence of any such underlying metaphysical picture, we might wonder whether the reference class could be *anything*. Perhaps my reference class consists entirely of Joe, Winston Churchill, the set of 47 pigs that acted in the 1995 comedy-drama *Babe*,[76] five bug-eyed aliens 10^100 light-years away, and a King of France who never existed. When God created this world, he made "me" one of these creatures at random (the relevant King of France happened to not be present in this world). Probably, I was going to be a pig.

What rules out this sort of picture? The natural answer is: its flagrant arbitrariness. But is there some non-arbitrary alternative? We discussed one candidate above: the minimal reference class consisting entirely of "people in your epistemic situation." We saw, though, that this doesn't work: it gives the wrong answers in cases like *God's coin-toss with equal numbers*, and it violates conditionalization as well.

[76] See Chanko (1995). "'There was,' Miller admits reluctantly, 'one animatronic pig'".

If we jettison the minimal reference class, the natural next alternative would be something like the maximal reference class, which I think of as the reference class consisting of all observer-moments. Bostrom, though, rejects this option, because he wants to use various limitations on the reference class to try to avoid various counterintuitive results, like the *Doomsday Argument*, *The Red-Jacketed High Roller*, *Save the Puppy*, and so on.[77] I'll say more about why I don't think this works below. Indeed, my current view is that the most attractive form of SSA embraces the maximal reference class. This is partly because I don't think Bostrom's rejection of it gets him what he wants, but centrally because it feels much less arbitrary than something in between minimal and maximal.

Even for the maximal reference class, though, worries about arbitrariness loom. There are, of course, questions about what counts as an observer-moment. Beyond this, though, if we're really trying to be maximal, we might wonder: why stop with observer-like things? Why not, for example, throw in some unconscious/inanimate things too? Sure, I know that *I'm* an observer-like thing. But the whole point of reference classes is to include things I know I'm not. So why not include rocks, galaxies, electrons? Why not the composite object consisting of the moon and my nose? Why not, for that matter, abstract objects, like the natural numbers? Viewed in this light, "things" seems a more maximal reference class than "observer moments" (and perhaps "things" is itself less-than-fully maximal; do the things have to exist? Can merely possible things count? What about impossible things?). And if "observer-moments" turns out to be less-than-fully maximal, it loses some of its non-arbitrariness appeal.

Suppose that following Bostrom, we reject both the minimal and the maximal reference class. Is there anywhere non-arbitrary we could land in between? One option would be to

[77] See Bostrom (2022a), p. 171.

appeal to some notion of metaphysical essence or modal profile.[78] Thus, we might say, you *couldn't* have been a pig, or an alien, or an electron. And if you *couldn't* have been something, then perhaps God couldn't have randomly made you that type of thing, either. Indeed, it can be tempting to construe the notion of "reference classes" is a manner at least vaguely reminiscent of metaphysical essences or modal profiles (e.g., "but you *couldn't* have been a rock; you're an *observer!*"), even absent an explicit account of the concept at stake. Bostrom, though, seems keen to distance himself from this sort of discourse; and once we start making cosmological predictions on the basis of whether being a brain emulation is compatible with my metaphysical essence or modal profile, one starts to wonder even more about presumptuousness.

Are there other non-arbitrary reference class options, between minimal and maximal? Maybe: humans? But why? Why not: creatures in the genus homo? Why not: primates? Why not: intelligences-at-roughly-human-levels? Why not: people-with-roughly-my-values?[79] I'm not aware of answers, here; and absent a story about what reference classes are, it's hard to say what an answer could look like.

What's more, this untethered quality has real effects on our ability to use SSA to say useful or determinate things. We started to get a flavor of this in the discussion above, when we found it necessary to preface different cases with provisos about who is or isn't in the reference class — e.g., "I'm assuming, here, that God/the puppy/the boulder isn't part of the reference class, but that the people on the other planets/with the blue jackets/in

[78] This connection between metaphysical essences and modal profiles is itself the subject of debate in the literature (see e.g. Fine (1994)), but I won't attempt to wade into that here.

[79] I am assuming, for the sake of this paragraph (though not in the chapter more broadly), that I have enough information about my species, my intelligence level, my values, and so on to rule out scenarios in which some people in my epistemic situation aren't in my reference class, if my reference class were determined by one of these traits. If we don't make an assumption like this, and instead allow the reference class to be determined in a way that places some people in my epistemic situation outside of my reference class, then I expect yet further complications for SSA.

the *doom later* world are." And it becomes even clearer in cases like *God's coin toss with chimpanzees*, in which your credence hinges crucially on whether you count chimps in the jungle as in the reference class or not.[80]

One of Bostrom's main responses to objections like this is to appeal to a kind of partner in guilt with the Bayesian's prior. That is, Bostrom acknowledges that even though we can put *some* constraints on what sorts of reference classes are reasonable, at the end of the day rational people might just disagree about what reference classes to use. But this is plausibly the case with priors, too; and still, we can get to agreement about various types of conclusions, because in cases of strong evidence, a wide variety of reasonable priors will converge on similar conclusions. Perhaps, then, we might hope for something similar from anthropics. That is, some verdicts (e.g., our scientific observations are reliable) will be robust across most reference classes, and others (hopefully: bad ones like the Doomsday Argument, telekinesis, etc) will be less so, and so less objective.

I do think this response helps: seeing reference classes as a mysterious subjective object like priors puts them in somewhat more respectable company. Still, though, I think we should view introducing yet another mysterious subjective object of this kind as a serious disadvantage to a theory — especially when we can't really give an account of what it's supposed to represent.

X. Against reference class epicycles

[80] Indeed, reading over Grace's (2010a) overview of her attempt apply SIA and SSA to reasoning about the Great Filter (that is, the step or steps along the trajectory to space colonization that detectable extraterrestrial life exceedingly rare), I was struck by the contrast been SIA's comparatively crisp verdicts ("SIA increases expectations of larger future filter steps because it favours smaller past filter steps"), vs. the SSA's greater muddle ("SSA can give a variety of results according to reference class choice. Generally it directly increases expectations of both larger future filter steps and smaller past filter steps, but only for those steps between stages of development that are at least partially included in the reference class").

I also want to flag a use of reference classes that I'm especially opposed to: namely, redrawing the lines around the reference class to fit whatever conclusion you want in a given case. Here I want to look at a move Bostrom makes, in an effort to avoid cases like *Save the Puppy*, that has this flavor, for me. I'll argue that this move is problematically epicyclic (and un-Bayesian); and that it doesn't work anyway.

To see the structure of Bostrom's move, recall:

> *God's extreme coin toss with jackets:* God flips a fair coin. If heads, he creates one person with a red jacket. If tails, he creates one person with a red jacket, and a million people with blue jackets.
>
> - *Darkness*: God keeps the lights in all the rooms off. You wake up in darkness and can't see your jacket. What should your credence be on heads?
>
> - *Light+Red*: God keeps the lights in all the rooms on. You wake up and see that you have a red jacket. What should your credence be on heads?

In *Darkness* and *Light + Red,* SIA and SSA (respectively) each give extreme verdicts about the toss of a fair coin. These examples served as the templates for other putatively problematic implications of SIA (the *Presumptuous Philosopher*) and SSA (e.g., the *Doomsday Argument*, *Red-Jacketed High-Roller*, *Save the Puppy*). Bostrom hopes to avoid them both. That is, he hopes to thread a needle that will allow him to be 50% on heads in *Darkness*, *and* 50% on heads in *Light+Red* — *despite* the fact that *Light + Red* is just *Darkness,* plus some information that you didn't know before (namely, that your jacket is red). If Bostrom can succeed, he will have banished both forms of presumptuousness.

How can we reach such a happy state? Bostrom's claim is that your reference class *changes* when God turns the lights on. That is, in *Darkness*, your reference class is

"person-moments in darkness." But in *Light + Red*, your reference class is "person-moments who know they have red jackets." That is, in both cases, your reference class consists entirely of people in your epistemic situation. Thus, as SSA, you don't update away from the prior *in either case*. You start out in *Darkness*, at 50-50. Then, when the light comes on, rather than updating in the way that standard Bayesianism would imply, you re-run SSA's calculation with a new and improved reference class — a reference class that allows you not to think it was unlikely, conditional on tails, that you ended up with a red jacket. After all, on this new reference class, you "essentially" have a red jacket, and know it; you *couldn't* have been someone with a blue jacket (who knows it), granted that you, in the light, have a red. Thus, on tails, your jacket color is no surprise.

Problem solved? I'm skeptical. The immediate objection is that this move doesn't seem very Bayesian. Normally, we think that when you learn new information like "my jacket is red," where this information rules out various tails-world possibilities you had credence on, but no heads-world possibilities, your credence on being in a tails world changes.[81]

A higher-level objection is that it seems pretty clear that Bostrom is making this move specifically in order to give a certain set of answers in a certain set of otherwise problematic cases, and that he would have little interest in it otherwise.[82] Perhaps some philosophers won't be bothered by this. After all, fitting the cases well is an important

---

[81] Bostrom response to this is to appeal to the fact that you're *losing* indexical information (e.g., "I'm a person-moment who doesn't know what their jacket color is") even as you gain new information (e.g., "my jacket is red"). I'm not exactly sure why this is supposed to help; but regardless, you're losing indexical information of this kind all the time. For example, when you see the clock tick forward, you lose the indexical information that "I'm a person-moment at time t1," and even as you gain new information like "it's now 7:01." But we don't think that this warrants violations of conditionalization like the ones Bostrom countenances here.

[82] Indeed, he frames this move as in some sense "optional" — something you can get away with, if you want to avoid both e.g. the *Presumptuous Philosopher and Save the Puppy*, but which you don't, as it were, *have* to make. But the fact that in Bostrom's book you don't "have" to make this move betrays its lack of independent justification: it's not a move you'd come up with on your own, for some other reason. If you *don't* want to make it (for example, because it seems arbitrary, un-Bayesian, and so on) nothing pushes back — except, that is, the cases-you-might-not-like.

desideratum in its own right. But too often, in my opinion, over-focus on this desideratum, relative to theoretical considerations like simplicity and explanatory power, leads to epicyclic contortions of fundamental principles – and Bostrom's version here sets of many alarm bells, for me, in this respect. Indeed, it makes me wonder about what sorts of limits — if any — are meant to constraint how much we can redraw our reference classes, moment to moment, to suit our epistemic whims. If SSA lets us say 50% in both cases, here, what *won't* it let us say? And if our theory can be made to say anything we want, how can we ever learn anything from it? The specter of the reference class's indeterminacy looms ever larger.

My most flat-footed objection, though, is that this particular move doesn't work by Bostrom's own lights. Rather, it runs into the same problems that the minimal reference class does. To see this, consider a version of *God's coin toss with equal numbers*:

> *God's coin toss with equal big numbers*: God flips a fair coin, and creates a million people either way. If heads, he gives them all red jackets. If tails, he gives one of them a red jacket, and the rest blue jackets.
>
> - *Equal Number Darkness*: God keeps all the lights off. You wake up in darkness. What should your credence be on heads?
>
> - *Equal Number Light + Red*: God keeps all the lights on. You wake up and see that you have a red jacket. What should your credence be on heads?

Equal Number Light + Red is very similar to the original *Light + Red*: the only difference is the presence of an extra ~million people with red jackets, conditional on heads. However,

Bostrom is committed (I think, rightly) to saying that in *Equal Number Light + Red*, you should be very confident that the coin landed heads.[83]

But the reference class Bostrom wants to use in the original, non-equal-number *Light + Red* doesn't allow him this confidence in the equal-number version. That is, in *Light + Red*, Bostrom wants to use the reference class "person-moments who know they have red jackets" — that's why he can stay at 50-50, despite all those know-they-have-blue-jackets people in the tails world. But this means that SSA stays at 50-50 in *Equal Number Light + Red,* too: after all, in both cases, people in your epistemic situation are 100% of the reference class. But this is a verdict Bostrom explicitly *doesn't* want.[84]

So overall, I'm skeptical of attempts like Bostrom's to give heads 50% in both *Darkness* and *Light + Red*; and especially skeptical about doing so on the grounds of changing reference classes. And I expect similar issues to apply to other attempts to use reference classes to avoid the SSA's problematic verdicts about cases like *Save the puppy*.

XI. Is killing epistemically different from non-creation?

I'll mention one other category of abstract argument for SIA over SSA, which I find quite compelling. Consider two cases, adapted from Armstrong (2009):

> *Coin toss + killing*: God tosses a fair coin. Either way, he creates ten people in darkness, and gives one of them a red jacket, and the rest blue. Then he waits an

[83] Indeed, as discussed in the footnote 51 above, Bostrom thinks that if you can't say things like that, you can't do science in worlds big enough to contain observers who make all physically possible observations (see Bostrom (2002b, especially section I and section VI) for more).

[84] Indeed, I feel confused by Bostrom's treatment of this issue. After introducing his treatment of the original *Light + Red* on p. 165 of Bostrom (2002a), he goes on, 13 pages to later, to discuss why the minimal reference class fails in cases like *Equal Number Light + Red,* and to suggest that in *Equal Number Light + Red*, the proper reference class to use is wider than "person-moments who know that they have red jackets" (in particular, he discusses the reference class "all person-moments"). But Bostrom surely doesn't mean to suggest that we should use "person-moments who know that they have red jackets" in *Light + Red*, but something wider in *Equal Number Light + Red*. The cases, after all, are basically the same.

> hour. If heads, he then kills the blue-jacketed people. If tails, he kills the red-jacketed person. After the killing in either case, he rings a bell to let everyone know that it's over. You wake up in darkness, sit around for an hour, then hear the bell. What should your credence be that your jacket is red, and hence that the coin landed heads?
>
> *Coin toss + non-creation*: God tosses a fair coin. If heads, he creates one person with a red jacket. If tails, he creates nine people with blue jackets. You wake up in darkness. What should your credence be that your jacket is red, and hence that the coin landed heads?[85]

Here, SIA gives the same answer in each case: 10%. After all, there are many more people in your epistemic situation in tails worlds.

SSA, by contrast, gives different answers in each case.[86] Thus, in *Coin toss + non-creation*, it gives its standard 50% answer: you were (SSA thinks) guaranteed to exist either way. But in *Coin toss + killing*, it switches to agreeing with SIA. In particular, when it first wakes up, but it hasn't yet heard or not heard the bell, it updates against having a red jacket, to 10%: after all, it's an equal-numbers case, and most people have blue jackets. Then, because the chance of death is 50% conditional on either having a blue jacket, or a red jacket, it stays at 10% after hearing the bell: survival is no update.

But are these cases importantly different? Armstrong doesn't think so,[87] and I'm inclined to agree.

---

[85] See also a closely-related version in Dorr (2002), and a related series of cases in Arntzenius (2002)).
[86] Here I'm returning to a version of SSA that keeps the reference class constant.
[87] At least, circa 2009; he's since changed his view (see Armstrong (2011a), for reasons to do with decision theory.

Dorr (2002) makes a similar argument in Sleeping Beauty. Consider a version where Beauty is woken up on both Monday and Tuesday conditional on both heads and tails, but then, if it's heads and Tuesday, she hears a bell after an hour or so. Surely, argues Dorr, Beauty ought to be 50-50 on heads vs. tails prior to hearing-the-bell-or-not, and 25% on each of Heads-Monday, Heads-Tuesday, Tails-Monday, and Tails-Tuesday. Then, after she *doesn't* hear the bell, surely she should cross off "Heads-and-Tuesday," re-normalize, and end up at 1/3rd on heads like an SIA-er. And indeed, this is what SSA *does* do (unless, of course, we mess with the reference classes), *if* Beauty is also woken up in Heads-Tuesday and can hear this type of bell. But if Beauty *isn't* woken up in Heads-Tuesday at all, then suddenly SSA is back to halfing. Does this difference matter? To me it seems like: no.

The dynamic at work in these cases is sometimes called SSA's "sensitivity of outsiders."[88] That is, SSA cares a lot about the existence (or non-existence) of people/person-moments you know that you're not: for example, person-moments who just got killed by God (even though you're alive), or who heard a bell you didn't hear, or who are living as chimpanzees in the jungle while you, a human, participate in strange thought experiments. At bottom, this is because if such people exist (and are in the reference class), their existence makes it less likely that you live in their world, because such a world makes it less likely that you'd be you, and not them.

Indeed, perhaps for some SSA-ers, who hoped to say SIA-like things about various cases, outsiders come as some comfort. This is because (if you use your reference classes right), outsiders can push SSA towards more SIA-like verdicts. Consider, for example, a version of God's coin toss where if heads, he creates one person in a white room, and if tails, two

[88] See discussion in Bostrom (2002b), p. 196.

people in white rooms; but where there are also a million chimps in the jungle either way (and the chimps are in the reference class). In such a case, SSA can get pretty close to thirding: if heads, you had a 1/~1M chance of being in a white room rather than the jungle, and if tails, you had a 2/~1M chance of this, so finding yourself existing in a white room is actually a ~2:1 update in favor of tails. SSA-ers might try to use similar "appeals to outsiders" to try to avoid saying bad things about the doomsday argument. Thus, if there are (finite) tons of observers and they're all in the reference class, the difference between *doom soon* and *doom later* does less to the fraction of people-in-your-reference-class you are.

I think moves like this might well help to alleviate some of SSA's bad results in real-world cases (though we'd have to see if the details check out). But note that they can also be used to give SIA's counter-examples *to* SSA. Thus, in the *Presumptuous Philosopher*, if we add a sufficiently large number of extra observers who we know that we *aren't* to T1 and T2, then the fact that T2 has a trillion times more people-in-our-epistemic-situation makes it the case that in T2, you're a ~trillion times larger fraction of the reference class. So SSA, too, starts mortgaging the house to bet with the scientists.

Beyond this, though, solutions to SSA's problems that involve appealing to the number of outsiders feel, to me, fairly ad hoc. And SSA's bad results in cleaner, more thought-experimental cases (e.g., *Save the Puppy*) will persist.

XII. SSA's solipsism

I'll note one final worry about SSA: it updates strongly towards solipsism. If you were the only thing that exists, it would be *guaranteed* that you are you. Thus, compared to hypotheses where there are tons of people and you just *happen* to be you, solipsism, for

SSA, becomes notably attractive. Indeed, if there are 100 billion+ people in the reference class in non-solipsism worlds, that's a 100 billion+-to-one update in favor of solipsism. Suddenly, your prior credence in solipsism starts to matter quite a bit.

And it's not just other people. Consider your memories. If accurate, they would involve a suspiciously large number of other-person-moments-in-the-reference-class. So SSA is correspondingly dubious about them. And the same, of course, for your future.

I'm not sure that this objection ultimately adds much to the others. But it's a nice illustration of a broader dynamic. Just as SIA updates towards populated worlds, if you don't know who you are, SSA updates towards lonely worlds, if you do. And the solipsist's world is the loneliest of all.

XIII. What about betting?

I've now covered my main objections to SSA. In a moment, I'll turn to SIA's downsides. First, though, I want to briefly explain why I've thus far mostly skipped over a certain category of argument: namely, appeals to what sorts of anthropic approaches lead to the right patterns of betting behavior.

My reason for this is simple: namely, betting in anthropics get complicated very fast. I do think it's important; but it's sufficiently hard to disentangle, and sufficiently far (in my opinion) from the only desiderata, that I decided to focus my analysis elsewhere.

Why is betting in anthropics complicated? Because how you should bet, in a given case, isn't just a function of your credences. It's also a function of things like whether you accept evidential decision theory (EDT) or causal decision theory (CDT) (or something

else),[89] your level of altruism towards other people in your epistemic position, how that altruism expresses itself (average vs. total, bounded vs. unbounded), and the degree to which you go in for acting (and believing) in accordance with pre-commitments you would've made from some prior epistemic position.[90] Cases in anthropics tend to implicate these issues to an unusual degree, and in combination, it's a lot of variables to separate and analyze.

I'll give one example to illustrate some of the complexity here.[91] You might be initially tempted by the following argument for thirding, rather than halfing, in Sleeping Beauty. "Suppose you're a halfer. That means that when you wake up, you'll take (or more specifically, be indifferent to) a bet like: 'I win $10 if heads, I lose $10 if tails.' After all, it's neutral in expectation. But if you take that sort of bet on every waking, then half the time, you'll end up losing $10 *twice:* once on Monday, and once on Tuesday. Thus, the EV of a 'halfer' policy is negative. But if you're a thirder, you'll demand to win *$20* if heads, in order to accept a $10 loss on tails. And the EV of this policy is indeed neutral. So, you should be a thirder."

But this argument doesn't work if Beauty's person-moments accept EDT (and are altruistic towards each other). Suppose you're a halfer person-moment offered the even-odds bet above on each waking. You reason: "It's 50% I'm in a heads world, and 50% I'm in a tails. But if I'm in a tails-world, there's also another version of me, who will be making this same choice, and whose decision is extremely correlated with mine. Thus, if I accept, that other version will accept too, and we'll end up losing twice. Thus, I reject." That is, in

---

[89] As a reminder, EDT says to choose the action such that your choosing it is the best *news*; CDT says to the choose the action with the best causal effects. See Weirich (2020) for more.
[90] See Meacham (2010) for an approach to decision-theory in this broad vicinity; and see Carlsmith (2021a, section IX) for more.
[91] See e.g. Hitchcock (2004), Arntzenius (2002), Briggs (2010), and Yamada (2019) for more discussion of betting in the context of anthropics.

this case, your credences can't be read off directly from the betting odds you'll accept (i.e., the fact that you reject an even-odds bet doesn't indicate that you place something other than 50% credence on each outcome).[92] Is that surprising? It might initially seem that way. But in general, if you're going to take a bet a different number of times conditional on outcome vs. another, the relationship between the betting odds you'll accept and your true credences gets much more complicated than usual. This is similar to the sense in which, even if I am 50-50 on heads vs. tails, I am not indifferent between a 50% chance of taking the bet "win $10 on heads, lose $10 on tails" *conditional on heads* vs. a 50% chance of "win $20 on heads, lose $20 on tails" *conditionals on tail*s. Even though both of the bets are at 1:1 odds (and hence both are neutral in expectation pre-coin-flip), I'd be taking the bigger-stakes bet on the condition that I lose.[93]

Indeed, the EDT-accepting *thirder*, here, actually ends up betting with odds that would suggest a "*fifth-er*" pattern of credence, if you tried to naively read off credences from betting odds (which you shouldn't).[94] That is, if offered a "win twenty if heads, lose ten if tails" bet upon each waking, this sort of Beauty reasons: "1/3rd I'm in a heads world and will win $20. But 2/3rds I'm in a tails world, and am about to take or reject this bet *twice,* thereby losing $20. Thus, I should reject. To accept, the heads payout would need to be $40 instead." And note that this argument applies *both* to SIA, *and* to SSA in the Dorr/Arntzenius "Beauty also wakes up on Heads-Tuesday, but hears a bell in that case" version (since SSA's credences mirror SIA's in that case).[95] That is, every altruistic EDT-er

[92] See Yamada (2019), p. 1249-1251, for more on the relationship between credences and acceptable betting odds.
[93] See Arntzenius (2002) for more on this.
[94] See Yamada (2019), p. 1254, for discussion of this.
[95] Thanks to Paul Christiano and Katja Grace for discussion. See Christiano (2021) for more on "fifth-ing."

bets, sometimes, in a way that would naively (but wrongly) suggest a "fifth-ing" pattern of credence.[96]

Note that in these cases, I've been assuming that Beauty's person-moments are altruistic towards each other. But we need not assume this. We could imagine, instead, versions where the person moments will get to spend whatever money they win on themselves, before the next waking (if there is one), with no regard for the future of Beauty-as-a-whole. Indeed, in analogous cases with different people rather than different person-moments (e.g., God's coin toss), altruism towards the relevant people-in-your-epistemic-position is a lot less of a default – thereby introducing further complications.

Do we need to wade into these complications in order know what beliefs to form in these cases? I'm not sure that we do. In particular, to me it seems possible to separate the question of how to bet from the question of what to believe. Thus, for example, in the EDT halfer case above, it seems reasonable to me to imagine thinking: "I'm 50% on heads, here, but if it's tails, then it's not just me taking this 'win $10 if heads, lose $10 if tails' bet; it's also another copy of me, whose interests I care about. Thus, I will demand $20 if heads instead." You can reason like that, and then step out of your room and continue to expect to see a heads-up coin with the same confidence you normally do after you flip. Maybe this is the wrong sort of expectation, but I don't think your betting behavior, on its own, establishes this.[97]

[96] At least assuming they choose their policy based on the epistemic position they occupy once they wake up, rather than some other epistemic position (e.g., the one they had on Sunday).

[97] Here's another case in this vein: if I know that a coin was flipped in determining whether to poison my sandwich, and I am deciding whether the accept the bet "I win $100 if the sandwich is poisoned, but I lose $100 if it's not," the fact that I reject the bet (because money is worth less to me if I am about to die of sandwich poisoning) need not imply that my credence on the sandwich being poisoned is something other than 50% (even if I typically value money linearly).

More generally, it doesn't feel to me like the type of questions I end up asking, when I think about anthropics, are centrally about betting. Suppose I am wondering "is there an X-type multiverse?" or "are there a zillion zillion copies of me somewhere in the universe?". I feel like I'm just asking a question about what's true, about what kind of world I'm living in — and I'm trying to use anthropics as a guide in figuring it out. I don't feel like I'm asking, centrally, "what kinds of scenarios would make my choices now have the highest stakes?", or "what would a version of myself in some previous epistemic position have pre-committed to believing/acting-like-I-believe?", or something like that. And more generally, in many cases, you can't decide how to bet *until* you have some picture of the truth. That is: anthropics, naively construed, purports to offer you some sort of *evidence* about the *actual world* (that's what makes it so presumptuous). Does our place in history suggest that we'll never make it to the stars?[98] Does the fact that we exist mean that there are probably lots of simulations of us?[99] Can we use earth's evolutionary history as evidence for the frequency of intelligent life?[100] Naively, one answers such questions *first*, then decides what to do about it. And I'm inclined to take the naive project on its face.

XIV. Epistemic pascal's muggings, infinities, and other problems for SIA

Having sketched my objections to SSA (and my provisional take on betting in anthropics), let's return to SIA's problems.

We've already discussed the canonical counter-example to SIA: namely, the *Presumptuous Philosopher*. And as I said earlier, I'm happy to grant that this is bad result. In particular, it seems strange to think that we should upweight scientific hypotheses in proportion to the

[98] See Leslie (1996).
[99] See e.g. Shulman and Bostrom (2012) and Xu and Shulman (2021).
[100] See e.g. Synder-Beattie et al (2021).

number of people-in-our-epistemic-situation they posit or imply. Thus, and especially in light the additional problems I'll discuss below, I agree that we should continue to look for a theory superior to both SIA and SSA, rather than simply forcing ourselves to choose from such an unappetizing menu.

However, for reasons I'll discuss below, I'm not currently optimistic about finding an attractive theory that avoids *Presumptuous-Philosopher*-like conclusions. So I think we should at least be *open* to biting the bullet on the *Presumptuous Philosopher*, despite the discomforts this entails. After all, the *Presumptuous Philosopher* is a very natural extension of reasoning we endorse in other contexts (for example, thirding), and as a matter of philosophical methodology, I think we should be wary of contorting our underlying principles too heavily around a single data-point, especially if the costs of doing so start to approach the costs implied by alternatives like SSA.[101]

That said, the *Presumptuous Philosopher* – especially as canonically stated -- isn't the only problem for SIA. Let's look at a few others, including some more extreme variants on the *Presumptuous Philosopher* structure: namely, variants that focus on (a) especially wacky (and populated) hypotheses, and (b) infinite worlds.

In the original *Presumptuous Philosopher*, the more-observer cosmology was a respectable theory – respectable enough, at least, for the scientists to be interested in testing it. But SIA need not be so conservative in the hypotheses it considers; and once we open the

[101] I find analogies with the repugnant conclusion in population ethics interesting in this respect. It, too, is the canonical counter-example to an otherwise simple and in-many-respects-attractive theory; and in my opinion, some ethicists are too willing to pay extreme theoretical costs in order to avoid it (see Zuber et al (2021) for agreement). Indeed, the repugnant conclusion and the *Presumptuous Philosopher* have an underlying structural similarity as well: both involve adding lots of people each with low-something (welfare, for the repugnant conclusion, and prior probability of existing, for the *Presumptuous Philosopher*) to a world, in a manner that yields a sufficiently large total-something (welfare, for the repugnant conclusion, and probability on that world, for the *Presumptuous Philosopher*). For this and other reasons, I sometimes think of SIA as the epistemic analog of totalism in population ethics.

doors to weirder worlds, it becomes easier to throw lots of people in your epistemic situation into the mix.

The formula is akin to an epistemic version of the "pascal's mugging" suggested by Bostrom (2009). Thus, in an ethical pascal's mugging, the worry is that the mugger can increase the amount of utility at stake in a given world faster than you can decrease your probability on it – thereby causing it to dominate your overall expected utility calculations (though whether this worry is ultimately sound is a different question). In an SIA-like epistemic pascal's mugging, by contrast, the worry is that we can increase the number of people-in-your-epistemic situation that a hypothesis involves faster than you can decrease your prior credence on it – thereby causing it to dominate more standard hypotheses after you make SIA's update towards worlds with more of such people.

Thus, for example, consider the hypothesis that when you next leave the room you're currently in, you'll find yourself in the midst of a giant sea of people just like you, each of whom are emerging from rooms exactly like your own; or that the universe consists entirely of an advanced civilization obsessed with simulating exactly your current experience, over and over, using optimal computational hardware; or that there are a graham's number of "hidden realms," overlapping with this one, each of which are chock-full of people having precisely your experience.[102]

Maybe none of these specific hypotheses, on their own, are going to dominate your credence in practice (there are too many other heavily-populated worlds to compete with). But plausibly, even once we factor in your presumably low prior on worlds (though the details here do matter), the odds ratio that SIA's update puts in their favor will be enough

[102] See Olum (2000, p. 15).

for them to quickly drown out any worlds that *aren't* stuffed to the brim with observers-in-your-epistemic-situation: including, plausibly, most of the everyday worlds we're used to considering.

Is this a problem? Yes, I think it is. But here, the analogy with an ethical pascal's mugging may provide some comfort. That is: we know that pascal's muggings (and related forms of "fanaticism") are problems for expected utility theory.[103] We know that you can make up ridiculous hypotheses involving an even more ridiculous number of lives-to-be-saved. So in a sense, we're already used to this sort of "big number, not-small-enough-probability" problem in other contexts – and perhaps the eventual solutions (if there are solutions) will be similar. In the meantime, though, I don't think we should give up on expected utility theory, or the idea that saving more lives is good (indeed, non-diminishingly good) – these ideas are too useful/compelling. And I'm inclined to think that we should respond to the possibility of making up ridiculous numbers of observers-in-your-epistemic-situation in a similar way: that is, to worry about it, but not to despair just yet.

What about infinite cases? Doesn't SIA become *certain* that the universe is infinite — and in particular, that it's filled with infinitely many observers in our epistemic situation? It depends a bit on the set-up, but my own view is that the best version of SIA probably does become certain of this.[104] Indeed, this sort of certainty seems like a natural extension of the logic of the presumptuous philosopher.[105] And this does seem overconfident. One might think: surely the universe *could be* finite. Finitude, after all, is a live scientific

[103] See Beckstead and Thomas (2021) and Wilkinson (2021) for discussion.
[104] Other options include: becoming undefined on infinite cases, or trying to carve out ad hoc exceptions for infinite cases. See Manley (unpublished) for discussion.
[105] Just as obsession with infinitely high-stakes outcomes seems like a natural extension of vulnerability to a finite pascal's mugging. See Thomas and Beckstead (2021) for more on the close connection here.

hypothesis.[106] And even if the scientists were leaning hard towards an infinite universe, couldn't it have been the case that it was finite instead? And if there would've been observers in such a situation, wouldn't SIA doom them to being infinitely wrong?

I do think this is a problem. I'll note, though, a few responses.

First, SSA can get certain about infinite worlds too: just, in the other direction. That is, SSA becomes certain that we're *not* in an infinite world, once it narrows down its location in that world to any finite population.[107] Maybe this isn't quite as bad as SIA's form of certainty (perhaps because it's hard to know where you are in many infinite worlds; or because it's better to rule out infinite worlds than the rule them in); but it has the same flavor of over-confidence.

Second, as with finite presumptuous philosopher cases, you shouldn't be certain of SIA, and so shouldn't, in practice, be certain of the conclusions it reaches.

Third, if you look at the world from the SIA-like perspective of "I am a particular possible person-in-my-epistemic-situation, who didn't have to exist," and you think of the world as drawing you out of a hat of possible people like that, then it doesn't seem *that* crazy to think that the fact that you got drawn licenses ruling out a merely finite number of draws. In particular: if we assume that the hat of possible people-in-your-epistemic-situation is infinite, then a merely finite number of draws suggests that the probability that you get

[106] See e.g. Sean Carroll's comments on his and Bostrom's (2020) 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."
[107] See Grace (2011) and Manley (unpublished) for discussion. Here I'm assuming that SSA is set up such that the relevant fraction doesn't become undefined.

drawn is zero.[108] Thus, while it's true that in finite worlds, a few sorry SIA-ers with your evidence end up infinitely wrong, *you're* guaranteed not to end up in that situation.

That said, I don't think these responses are especially comforting.[109] And SIA's infinity problems don't stop with certainty that it's in an infinite world. It's also unsure how to reason about *which* infinite world.[110] Suppose, for example, that if heads, God will create an infinite line of people, all with blue jackets except for one red-jacketed person every million people; but if tails, God gives everyone red jackets instead. Now suppose you wake up with a red jacket. What's the probability of heads? In both cases, SIA tries to scale the prior on both worlds by a factor of infinity, and its output becomes undefined.

Now, importantly, SSA has problems with this case, too. That is, rather than scaling the prior on an objective world by *n* (the number of people in your epistemic situation in that world), SSA as I defined it scales the prior on O by $\frac{n}{r}$, where *r* is the number of people in the reference class in that world. But if both *n* and *r* are infinite, this fraction is undefined as well.

One response, in SSA's case, is to appeal to the limiting fraction $\frac{n}{r}$ for the observers contained within an expanding sphere of space-time. And we might look for options like this in SIA's case as well – for example, appeals to the limiting *density* per unit space-time

[108] Assuming it's defined at all.
[109] In particular, the last one appeals to a feature of the SIA narrative I offered that seems unattractive in its own right: namely, that on this narrative (recall that it's not the only one on offer), the most natural prior probability of God drawing you from the hat at all is 0. And positing an infinite number of draws feels like it's far from a straightforward solution.
[110] In conversation, Paul Christiano characterized this to me as a kind of "double whammy." First SIA becomes certain that it's in an infinite world – and then it immediately breaks.

of people in your epistemic situation. Moves like these, though, bring in substantial additional complications and objections.[111]

And note, too, that even if SSA and SIA knew how to come to overall credences on objective worlds with infinitely many people-in-your-epistemic-situation, they would both still need some way of distributing *de se* credence amongst all such people within such a world; and here *Indifference*, the principle I took for granted earlier, leads to problems fast.[112] Indeed, my understanding is that the general question of how to assign any plausible measure to the observers in an infinite universe (the so-called "measure problem") remains, for now, unresolved.[113]

Manley (unpublished) summarizes: "as usual, infinities ruin everything." I think this is a touch pessimistic overall (better solutions than I've discussed might well be available). Regardless, though, I actually think that the "as usual," here, should come as some comfort to SIA-ers. That is: "uh oh: this potentially attractive view, developed in the context of finite cases, says weird/unclear/fanatical things in infinite cases" is a sufficiently common alarm bell (see e.g. expected utility theory, population ethics, and so on) that its ringing with respect to SIA is a weaker update (and as I say, it rings for SSA as well). Indeed, the fact that these problems with SIA — e.g., Pascal's mugging-type cases, infinity issues — are so structurally similar to problems with expected utility theory and totalism seems, to me, some solace. They aren't good problems. But in my opinion, it's good (or at least, respectable) company.

---

[111] For example: they don't work if the relevant limit doesn't exist; they run into problems with relativistic space-times (see Dorr and Arntzenius (2016), p. 28); and they make your credences sensitive to spatio-temporal re-arrangements of the people within the world (e.g., swapping them around with each other, pulling them closer together, and so on), even while holding your other evidence fixed. See Dorr and Arntzenius (2016) for more discussion, focused on SSA-like proportions in particular.

[112] See Weatherson (2005) and Manley (unpublished).

[113] See e.g. the discussion in Wolchover and Byrne (2014): "Few consider the problem to be solved."

What about other objections to SIA? A common one is that you don't learn anything new in Sleeping Beauty (you knew, on Sunday, that you were going to wake up regardless, and you were at ½ on heads then): so why should you update upon waking? But once you're thinking about the case in terms of person-moments, and if you get into the SIA narrative I offered earlier, I think this objection weakens. Yes, *some* person-moment-in-my-epistemic-situation was going to exist either way: but that doesn't mean that "I" was going to exist either way. And my existing, on SIA's story, was more likely conditional on tails.

A different worry about SIA is that it makes bad empirical predictions. For example, if it turned out that the universe is definitely finite, there would be a strong temptation to reject SIA on those grounds (unless we've somehow revised it to get rid of its certainty about infinities). But if that's right, should we reject it now as well? For example: if SIA is right, why *isn't* the world chock full of observers-in-our-epistemic-situation in every possible nook and cranny? Why can I move without bumping into a copy of myself?

I do find this sort of objection worrying. But I think it mostly amounts to a restatement of the *Presumptuous Philosopher* objections above. And as ever, we can offer similar worries about SSA (e.g., "How come I'm not alone in the universe?", "Why don't we see stronger empirical evidence for doom soon?" – though I do think that SIA's worries may be worse, here).

A final objection is that at least in combination with certain values and decision theories, SIA implies inconsistencies between the policy you'd want yourself to adopt *ex ante*, and the behavior you engage in *ex post*. Consider, for example, the "fifth-ing" behavior I discussed above, where an SIA-er who accepts EDT (and who cares about their other person-moments) ends up betting like they have $1/5^{th}$ on heads in Sleeping Beauty, rather than $1/3^{rd}$ (if you were to try reading their credences directly off of their betting odds,

which you shouldn't). On Sunday, when SIA has ½ on heads, it would pre-commit to betting like a third-er instead (since it will make the bet twice in tails worlds), rather than a fifth-er.[114] Does this sort of inconsistency tell against SIA?

I don't think it does – or at least, not much. This is centrally because both EDT and CDT lead to this sort of inconsistency in lots of other non-anthropics cases too. Consider, for example:

> *Parfit's hitchhiker*: You are stranded in the desert without cash, and you'll die if you don't get to the city soon. A selfish man comes along in a car. He is an extremely accurate predictor, and he'll take you to the city if he predicts that once you arrive, you'll go to an ATM, withdraw ten thousand dollars, and give it to him. However, once you get to the city, he'll be powerless to stop you from not paying.[115]

Here, the policy you'd want yourself to adopt ex ante (that is, in the desert) is to pay in the city, because if you have this policy, the man will predict that you'll pay, and he'll save you. But EDT and CDT do not pay in the city – since once you're in the city, paying neither causes nor gives evidence for your being saved. But as a result, EDT and CDT agents die in the desert with very high probability, since the man accurately predicts that they won't pay.

If you're interested in avoiding inconsistencies like this, there are decision theories specifically crafted to do so.[116] And I think decision theory, rather than epistemology, is the place to turn here. In *Parfit's Hitchhiker*, for example, if you want to be the type of agent who pay in the city, you don't need to look for some epistemology that gives

---

[114] That is, on Sunday, a policy of accepting a bet of "$20 conditional on tails, $10 conditional on heads" is neutral in expectation.
[115] See Yudkowsky and Soares (2009, p. 8) for discussion.
[116] See, for example, Levinstein and Soares (2020) and Meacham (2010).

yourself false beliefs, once you're in the city, about what will happen if you don't pay. Rather, you can recognize that you'd get away with not paying, and choose to pay anyway. I'm inclined to extend this principle to anthropics as well. That is, if you don't want to be the type of agent who bets like a fifth-er, then just don't bet that way. But you need not distort your credences in the process. (And even if you don't like this response, SSA is little comfort. After all, as I mentioned above, SSA *also* bets like fifth-er in Sleeping Beauty sometimes – for example, in the Dorr/Arntzenius version of the case where Beauty wakes up on Heads-Tuesday too, but then hears a bell.)

XV. Hold out for Anthropic Theory X?

I've now reviewed a long list of objections to SIA. Some of them are indeed bad; but many apply to SSA as well, and when I bring to mind the other objections to SSA I reviewed earlier (e.g. telekinesis, reference classes, and so on), SIA seems to me superior.

Still, though, I've mostly been focused on comparing these two theories in particular. I haven't tried to survey all available views (see footnote for brief discussions of some candidate alternatives), or to chart the limits of the applicable logical space.[117] So one

---

[117] As mentioned previously, various views (see e.g. Halpern (2006), Meacham (2008), Neal (2006)) attempt to avoid anthropic updating altogether. But this leads to the "double-halfing" behavior I discussed in the text, which I see as incompatible with very basic constraints on Bayesian rationality, and it fails to explain how science is possible in big worlds where at least one observer will make any given observation regardless of its accuracy (this is a key concern of Bostrom (2002b – see footnotes in Section III)). See Manley (unpublished) for discussion of some other problems with it.

One variant of this approach (Neal's "Fully Non-Indexical Conditioning") emphasizes the need to employ a very fine-grained conception of your evidence and of the objective worlds at stake. This results in an approach that behaves much like SIA in cases where no single fine-grained world has multiple people in exactly your epistemic situation. However, it runs into the same double-halfing problems above if that condition doesn't hold. Neal, in response to this, chooses to simply ignore worlds where your epistemic situation is duplicated exactly, but as Arntzenius and Dorr (2016) point out, this seems unjustified.

Some (relatively niche) approaches to anthropic use a prior over observer moments that reflects the ease with which they can be described by an (arbitrarily chosen) universal turing machine. See Carlsmith (2021a) and (2021b) for more discussion of the advantages and disadvantages here.

might reasonably wonder: given the disadvantages of both views just discussed, shouldn't we look for some alternative? And even if an exhaustive survey and analysis reveals that a superior alternative has yet to be articulated, shouldn't we hold out hope for one? After all, we do not yet have official "impossibility proofs" of the type we have in population ethics, to the effect that there is no anthropic theory that will satisfy all of Y constraints we hoped to satisfy.[118] So why would we settle for less?

I do think it's reasonable to keep looking for views superior to both SIA and SSA. But I also want to sound a note of pessimism about exactly how much satisfaction to expect such a process to yield, especially with respect to problematic results like the *Presumptuous Philosopher*. In particular, even if SIA and SSA sound like two very specific theories, to which there are presumably many viable alternatives, their *verdicts about particular cases* seem to exhaust many of the most plausible options about those cases. But yet, it is *precisely these verdicts*, applied in structurally identical contexts, that lead to some of their worst results.

Consider, for example, Sleeping Beauty. What are you going to say in Sleeping Beauty, if not ½, or 1/3?[119] Suppose that you're like me, and you want say a third, perhaps because you're moved by basic, compelling, and fairly-theory-neutral arguments like the

---

Armstrong's (2011a) "Anthropic decision theory" attempts to dissolve the need for anthropic updating by focusing on what sort of policy you'd pre-commit to prior to such updates, and then sticking with that. I think this is an interesting program (and one I'd like to explore more), but as I indicated in the text, I also think we can separate questions of epistemology and decision-making; and naively, Armstrong's approach seems to me to leave the epistemic questions unanswered.

I'll also emphasize again that I'm not treating the specific narratives I gave about SIA and SSA in the text as definitional; rather, what matters is the quantitative mechanics of how they update. In this sense, the equivalent principles in Manley (unpublished) and Isaacs et al (2021) are the same. And principles like *Weighting* in Thomas (2021a, p. 21), which lead to SIA-like results (at least in cases where objective chances are available), wouldn't qualify, on my set-up, as competitors to SIA (my understanding is that Thomas treats his principle as a competitor to SIA in virtue of imputing to SIA a commitment to my own existence not being a priori – but such a commitment is not an essential feature of SIA's quantitative mechanics).

[118] Thanks to Nick Beckstead, again, for suggesting this consideration. That said, I think the sorts of considerations discussed in this section suggest that we're not necessarily all that far away from such proofs.

[119] Let's leave aside "fifthing" for now, along with moves that attempt to incorporate your uncertainty about which theory of anthropics is correct.

Dorr/Arntzenius argument that "you should be 1/4th if you wake up on both days no matter what, and then if you learn that you're not Heads-Tuesday you should end up a thirder." So, you craft your candidate Theory X to say a third. But now make it a zillion wakings if tails instead. It's the same case: there's no magic about the number "a zillion." (Or at least, denying "no magic about the number a zillion" -- for example, by updating less and less with each additional person -- is itself a source of implausibility).[120]

Indeed, as I tried to emphasize above, the *Presumptuous Philosopher* and its variants are basically more science-flavored versions of a zillion-wakings Sleeping Beauty. Yes, there are a few candidate differences: for example, the prior is set by an objective frequency in Sleeping Beauty vs. some unspecified build-up of empirical evidence about fundamental reality in a more scientific case (though it's not clear, to me, why this difference would matter). And perhaps there are further differences to bring out as well. But ultimately, the basic thing that thirding does, regardless of its justification, is update towards worlds where there are more people in your epistemic situation. And this is also the basic thing that many of the most prominent objections to SIA get so worried about.

Thus, to go further: make it an infinite number of wakings, if tails.[121] Uh oh: for thirders, it really seems like this should be an update towards tails, relative to a merely zillion-waking world. After all, thirders updated towards a zillion-wakings world in virtue of its larger number of wakings; but the number of wakings in the infinite world is much (infinitely) larger again. Indeed, this sort of logic leads very naturally to being more confident in the

[120] This is the route taken by "UDASSA" – see Carlsmith (2021b) for discussion.
[121] See Jäger (2021), section 4, and Huemer (2021) for related discussion – though in the context of my preferred focus on observer-moments, I generally expect talk of "reincarnation" and "immortality" to mislead.

infinite wakings tails-worlds than in a heads world with any finite number of wakings. But assuming a non-zero prior on infinite worlds, that sounds a lot like a route certainty.

Obviously, there's more to say here, especially about infinite cases. My main point is just that the gap between "thing we want to say" and "thing we don't want to say" isn't necessarily very theory laden. Rather, it looks a lot like we like/want a given type of result in one case, and then we hate/don't want *that same type of result*, in some other more extreme but structurally similar version. This dynamic currently inclines me towards pessimism about finding an especially satisfying Anthropic Theory X, at least of a standard kind, that avoids both SIA most problematic implications, while also capturing the cases that, in my opinion at least, it gets right (e.g., Sleeping Beauty).

XVI. Implications

I'll close with a brief discussion of practical implications. What would SIA, if true, say about our real-world situation?

One classic implication, mentioned above, is skepticism about the traditional Doomsday Argument. This would be good news. Unfortunately, though, SIA may suggest its own sort of Doomsday Argument – at least, with respect to futures where humanity goes on to become the sort of civilization that does easily-observable things to the universe, like settling it on large scales. That is, faced with our failure to observe intelligent life (and thus, with the need to posit some step or steps on the way from lifeless matter to large-scale space settlement that makes detectable extraterrestrial life exceedingly rare), SIA plausibly updates towards most life killing itself (or otherwise failing to reach some threshold of detectability) *after* reaching our stage, rather than very rarely reaching our stage at all,

because this would make it more likely that there is lots of life out there at our stage, and hence more observers like us.[122]

Shulman and Xu (2021), however, argue that SIA updates *more* strongly towards being in a simulation run by a very powerful civilization devoting lots of resources to simulations of people at our stage – since hypotheses like these result in much higher populations of people like us, relative to more conventional cosmologies.[123] Perhaps this sort of update would be good news relative to doom soon – but it would hardly be a return to normalcy.

My own view, though, is that focusing on these sorts of updates fails to take seriously enough the *Presumptuous Philosopher* and infinity problems discussed above. That is, even for purely finite cases, worlds where advanced civilizations simulate people at our stage seem unlikely to be the worlds most populated-in-expectation-with-people-like-us, even after weighting according to pre-anthropic priors (consider, for example, the "graham's number of hidden realms" worlds above; worlds where the advanced civilizations are obsessed with simulating this exactly person-moment in particular rather than our-stage civilizations more generally; and so on). What's more, as I discussed above, SIA plausibly becomes certain that we're in an infinite world; and in order for these updates to occur in infinite worlds in a manner analogous to how they work in finite worlds, we need to take for granted some method of saying that there are "more" observers like us in one vs. the other. But we don't (or at least, I don't) have such a method (the "limiting density of observers-like-us per unit space-time" is one possibility here, but as I discussed above, it has problems).

[122] See Grace (2010) and Olson and Ord (2021).
[123] See also Shulman and Bostrom (2012).

That is, my own view is that the most salient implication of accepting SIA is (a) becoming certain that we live in an infinite universe, and then (b) not knowing how to reason about which.[124] So this leaves me quite unsure about what to take away from SIA as a whole – and inclined towards caution in applying it to real-world cases.

It's an uncomfortable situation – and one, indeed, that can tempt one to ignore this whole anthropics business and seek a return to more scientifically-respectable normalcy, even without endorsing some explicit alternative like SSA. But for all their strangeness, I don't think ignoring these issues is the right response, either. In particular, anthropics — at least, naively construed — purports to identify and make use of a form of *evidence* about the world: namely, for SIA, the evidence we get from the fact that we exist; and for SSA, the evidence that we get from the fact that we exist as these people in particular, as opposed to others in the reference class. This form of evidence is often overlooked, but on both of these views, it can end up a very powerful clue as to what's going on (hence, presumptuousness — and views that aren't presumptuous in this way struggle to make basic updates/conditionalizations in cases like *God's coin toss with equal numbers*). Neglecting anthropics as a category of consideration therefore risks missing out on centrally important information — including information it might be hard to get otherwise (for example, information about great filters, simulations, multiverses, and so on). And even if we don't see any immediate uses for this information, it seems useful to have on hand.

What's more, doing anthropics *badly* has costs. You can end up confused about the doomsday argument, for example, or about the fine-tuning of the universe. At the very

[124] This is quite analogous to the way in which accepting totalism about population ethics plausibly implies (a) becoming obsessed with cases where your actions affect infinite amounts of utility, but then (b) not knowing how to choose between actions of this kind. See Chapter 3 for more on this.

least, then, we need some sort of anthropic hygiene. And the line between "avoid basic errors" and "make important updates" isn't especially clear.

Overall, then: I currently think SIA is better than SSA. SIA still has problems, though, and I'm not sure how to apply it to the real world. We should try to figure out a better theory (or a better version of SIA -- the need to handle infinite cases seems especially pressing), and perhaps there is one out there already. In the meantime, we should tread carefully, but stay interested in understanding the implications of the theories we have.[125]

[125] This essay owes an especially large amount to discussion with Katja Grace, and to her work on anthropics. Thanks, as well, to Amanda Askell, Nick Beckstead, Paul Christiano, Tom Davidson, Carl Shulman, Bastian Stern, and Ben Weinstein-Raun for discussion.

Chapter 2

Simulation arguments

I. Introduction

Call someone who lives in a computer simulation a "sim," and someone who does not, a "non-sim." Some argue that we should have high confidence that at least one of the following is true: either it's not the case that most people with experiences broadly similar to our own (for example, experiences of living on something like 21st century earth) are sims, or we are sims.[126]

This essay formulates and defends what I see as the strongest version of this argument, and then explores some of the complexities and uncertainties that it leads to. I begin by describing the original argument in Bostrom (2003), and by offering a more general set of conditions that an argument like Bostrom's should satisfy. I then distinguish between two versions of such an argument – a Type 1 version, which rests on various empirical claims (in particular, about the computational power available to advanced civilizations); and a Type 2 version, which does not.[127] Bostrom's is a Type 1 argument.[128] But Type 1 arguments face objections – in particular, "selective skepticism" objections (why accept the relevant

[126] Here I am assuming an internalist conception of experiences, on which you and your brain-in-a-vat (BIV) equivalent have the same experiences, regardless of the nature of your environment. More on this assumption below. Importantly, the argument is not specific to one specific class of observer, like "observers with 21st century experiences." Rather, as I discuss below, it applies to *any* class of observer that satisfies certain conditions. For convenience, though, I focus initially on "observers with 21st century experiences" in particular.

[127] Chalmers (2022) offers something more like a Type 2 version. As he formulates it, though, Chalmers's argument doesn't say enough to require us to make any interesting modifications to our common-sense view of the world (more below). My own formulation is heavily influenced by the formulation and set-up in Thomas (2021) – though Thomas (2021) is focused on a different (and even more revisionary-if-sound) argument than Bostrom's. I discuss Thomas's argument in section XI below.

[128] Or at least, the text of the paper itself strongly suggests a Type 1 interpretation. Below I discuss whether it's possible to interpret it in Type 2 terms instead.

empirical claims about computational power, but not other common-sensical claims that would provide evidence that we're not sims?), and "self-undermining" objections (if we're sims, why think that our evidence supports the relevant empirical claims, applied to the reality simulating us?).

These objections may be answerable. But in my opinion, they distract unnecessarily from the core of what makes simulation arguments interesting. Loosely stated, this core (as I formulate it) consist of two claims: (1) conditional *only* on having e.g. 21st-century-experiences and on living in a world where most people with such experiences are sims, our credence on being sims should be high (I call this claim "*Observer-class indifference*"); and (2) adding in the rest of our evidence, while continuing to condition on most people with such experiences being sims, shouldn't change this credence (I call this claim "*Admissibility*"). Type 2 arguments focus on these core claims directly, and they treat questions about the likelihood of various empirical claims as secondary (though obviously, important to our credences overall). For this reason, I suggest, Type 2 arguments are superior to Type 1 arguments, and less vulnerable to the objections above. More importantly, though, I think they have real force – force I attempt to elucidate.

The second part of the paper examines some of the complexities and uncertainties that arise once we start to take Type 2 arguments seriously. In particular;

- *How should we adjust our overall credences – including our credence on being sims -- in light of these arguments?* Here I describe a number of options, all of which have problems. I then focus in particular on an argument from Thomas (2021), to the effect that we should be basically certain that we're sims, because conditional on being non-sims, the expected ratio of sims to non-sims is high. I suggest that we need not accept this conclusion, but that it points at the need for caution in trying to preserve common-

sensical credences *conditional on being non-sims*, if we accept Type 2 simulation arguments.

- *To what range of sims and scenarios do Type 2 arguments apply?* Here I suggest that these arguments apply to a wider (and stranger) range of cases than the literature has focused on. For example, I suggest that they prevent you from giving non-trivial credence to being an early-history non-sim human, living in a world where most people with early-history experiences are simulated squid-people with tentacle arms (and that this is true even though your evidence is incompatible with being a squid-person). Charting the full limits of the Type 2 argument's applicability, though, looks challenging.

I close by briefly mentioning a few other outstanding uncertainties – about infinite worlds, about the implications of anthropic principles like the Self-Indication Assumption ("SIA") and the Self-Sampling Assumption ("SSA") for simulation arguments,[129] and about the practical implications of taking such arguments seriously.

II. Bostrom's argument

Bostrom's (2003) simulation argument is the most prominent in the literature. He introduces it with two central assumptions: first, that suitably sophisticated sims can have conscious experiences like ours; and second, that technologically advanced civilizations (Bostrom calls these "post-human") would have enough computational power to run enormously many such sims with a trivial portion of their overall resources.[130] Bostrom justifies the first

[129] See Bostrom (2002a) for an overview of these principles.
[130] It's a little bit ambiguous, in Bostrom, whether he means to only talk about technologically advanced civilizations with human ancestors, but I think that's the most natural reading of the text and context (for example, in his (2008) FAQ, he talks separately about the possibility that we are being simulated by "aliens" with ancestors who are very unlike us; and his argument focuses on scenarios on which we either ancestor simulations in particular, or non-sims).

assumption by appeal to what he calls the "substrate independence" thesis in the philosophy of mind, and the second assumption via appeal to additional arguments about the limits of computation, the resources available to advanced civilizations, and the computation required to simulate a human-like mind and the necessary environment.

Bostrom then goes on to argue, in light of these assumptions, that at least one of following three propositions must be true:

1. The fraction of civilizations at our current stage that eventually become technologically mature is ~0.
2. The fraction of technologically mature civilizations that devote a non-negligible fraction of their resources to running simulations of their pre-post-human history (hereafter: "ancestor simulations") is ~0.
3. The fraction of observers with human-type experiences who are simulated (call this f_{sim}) is ~1 (where "human-type" means "pre-post-human").

(Bostrom's argument for this disjunction is not airtight, but it won't be my focus here.)[131]

[131] Bostrom's argument for this disjunction proceeds via an equation. Let f_p be the fraction of human-level technological civilizations that survive to technological maturity, f_i be the fraction of technologically mature civilizations interested in running ancestor simulations, let *H* be the average number of individuals that lived in a civilization before it reaches a pre-post-human stage, and let *N* be the average number of ancestor simulations run by a technologically mature civilization interested in running ancestor simulations. Bostrom suggests that the fraction of pre-post-human observers who are sims (f_{sim}) is given by the equation:

$$f_{sim} = \frac{f_p f_i NH}{f_p f_i NH + H}$$

The idea here is that each civilization contributes an average of H non-sims, and an average of $f_p f_i NH$ sims (e.g., an $f_p f_i$ fraction are interested in and capable of running sims, and the interested and capable ones create an average of *N***H* sims), to the overall population of pre-post-human observers. The *H* cancels out, and *N*, Bostrom has argued, is likely very large, so one of f_p or f_i needs to be very small for the fraction to be less than ~1.

But this calculation isn't airtight. For one thing, as Chalmers (2022, appendices) points out, in principle post-human civilizations could also create many *non-sims* with pre-post-human experiences (for example, non-sims on terraformed planets, non-sim brains in vats, etc) – non-sims that this equation wouldn't capture. What's more, as Bostrom and Kulczycki (2011) point out, it could be that *H*, the average number of people that live

Then, in the final step of the argument, Bostrom argues that if we condition on 3, our credence that we're sims (he calls this hypothesis "SIM") should be ~1. He makes this step on the basis of what he calls a "bland indifference principle" (BIP) which states that:

> "…if we knew that a fraction x of all observers with human-type experiences live in simulations, and we don't have any information that indicate [sic] that our own particular experiences are any more or less likely than other human-type experiences to have been implemented in vivo rather than in machina, then our credence that we are in a simulation should equal x:
>
> Cr(SIM | f_{sim} = x) = x" (p. 7)

Thus, concludes Bostrom, we should have high credence that at least one of 1, 2, or SIM is true.[132] In the original paper, he suggests splitting our credence evenly between them; and in a 2008 FAQ, he assigns SIM "something like in 20%-region, perhaps, maybe" (though in more recent interviews he explicitly "punts" on probabilities).

III. Set up

in a pre-post-human stage of a civilization, is much larger for civilizations that do not end up creating ancestor sims than for the ones that do. Thus, for example, if there are two civilizations, A and B, and A has a pre-post-human population of 10 people, and goes on to create 1000 ancestor sims of 10 people each (so, 10,000 sim people total), but B has a pre-post-human population of a billion people, and it goes on to create no ancestor sims, then the equation above gives the wrong result for f_{sim} (f_p is ½, f_i is 1, N is 1000, so the equation says that f_{sim} should be 500/501, but actually it's ~1/100,000 – i.e., 10,000/(10,000+1B+10)). Bostrom and Kulczycki (2011) discuss "patches" meant to address this issue. For my purposes, though, and especially for Type 2 arguments, the details of how we calculate f_{sim} are not central. Rather, what matters is the probability of the more basic disjunction: either it's not the case that f_{sim} = ~1, or we're sims.

[132] Bostrom typically phrases this conclusion as: either 1, 2, or we're almost certainly sims. Chalmers (2022, appendices) argues that strictly speaking this doesn't follow: if A is likely given B, that doesn't mean that either B is false, or A is likely. My own view is that it's fairly clear what Bostrom means – namely, that our overall credences should be such that, conditional on both 1 and 2 being false, our credence on SIM is very high. To avoid confusion, though, I'll generally to state this sort of disjunctive conclusion in a different form: namely, that we should have high credence on the disjunction of 1, 2, or "we're sims" (rather than on 1, 2, or "we're probably sims").

I think this argument contains a very important core of truth, but that Bostrom's framing leads to a variety of unnecessary problems. To get at this truth, and to illuminate these problems, let's do a bit more set up.

Following Bostrom and Chalmers (2022), I am not going to assume that if we are sims, we are systematically misguided in all of our everyday beliefs. It might well be true, for example, that if I am a sim, it's still the case that I have hands, because "I have hands" is made true by my simulated hands. And I'll generally assume that sims can be humans, people, observers, and so on, just like non-sims can.

However, I will assume that some sims are wrong about some beliefs that have more specific simulation-relevant implications. For example, I'll assume that if I'm a sim, the beliefs expressed by e.g. "I'm not a sim," "I have two unsimulated hands," and "No one created the universe I see around me," are wrong. And if I'm in a simulation where e.g. the stars are fake (e.g., they aren't simulated in any detail, or they aren't simulated at all when no one is looking), or where the universe I see around me is less than 10 billion years old, or where my memories have been tampered with by the simulators, then many of my more mundane beliefs are false as well.

Also, and importantly: when I talk about experiences, I'll be assuming an internalist conception of experiences – that is, a conception on which me and my brain-in-the-vat equivalent have the same experiences, even if our environments differ substantially. This isn't meant to be a substantive philosophical claim about the nature of experience – rather, I'm simply *stipulating* that when I talk about experience, I'm talking about whatever appearance-like thing it is that you and your brain-in-the-vat equivalent share. Readers uncomfortable with using the word "experience" for this are free to substitute an alternative.

I'll approach the questions here from a Bayesian perspective.[133] In particular, I'll assume that we come to these questions equipped with a prior probability distribution, which I'll denote Pr, over both *de dicto* hypotheses (that is, roughly, hypotheses about the nature of the objective world) and *de se* hypotheses (that is, roughly, hypotheses about my *location* within an objection world – e.g., which person I am, where and when I exist, etc).[134] I'll denote the *de se* proposition that I am a sim as "*i*S", the *de se* proposition that I am a non-sim as "*i*NS", and the *de se* proposition that I have total evidence E as "*i*E." And I'll assume that my overall probability on a given proposition *p*, given my total evidence, is the conditional probability $\Pr(p \mid i\text{E})$.[135]

As a generalization of Bostrom's "human-type," I'm also going to make use of the notion of an "observer class," which will denote a set of observers such that "I am an O" or "*i*O" is true and a part of my evidence (I call this condition *Membership* below), but which can also include people who do not have my evidence.[136] To get a flavor for this notion, consider the following case.

> *Sims with random numbers*: You wake up in a white room, with the number 3 written on your hand. A sign in front of you reads. "I, God, created nine sims, and one

[133] In particular, my approach and terminology are both heavily influenced by the approach in Thomas (2021), which I see as the most rigorous in the literature.
[134] See Lewis (1979) for a classic discussion. More specifically, we can think of an objective world as a fully specific *de dicto* hypothesis, and we can think of a centered world as a triple $\langle w, s, t \rangle$ where *w* is an objective world, *s* is a subject in that world, and *t* is a time.
[135] Here I do not mean to assume any particular views about the relationship between your evidence and your knowledge.
[136] Thomas (2021) calls this a "reference class," but this terminology calls to mind Bostrom's own use of "reference class" in the context of his work on the "Self-Sampling Assumption" (SSA) in e.g. his (2002a); and I've argued in the previous chapter that this sort of reference class is problematically mysterious. But my use of observer classes will be importantly different. In particular, Bostrom's use of reference classes in the context of SSA requires that there be one particular (and worryingly arbitrary) reference class that governs the sorts of updates SSA makes about the probability of living in a given world (or at least, the simple version requires this; we can imagine more complicated versions, where e.g. it's vague what the reference class is). But observer classes do not have this problem, because we do not have to fixate on a particular observer class. Rather, we can simply say (as I do below) that the constraints imposed by the simulation argument apply to *any* observer class satisfying certain conditions.

> non-sim, all in white rooms, all with identical signs. Then, for each observer, I drew a number (without replacement) out of a hat containing the numbers 1-10, and wrote it on their hand. No one else exists, other than me, the nine sims, and the one non-sim."

Let's assume, for the moment, that the truth of the sign's claims is part of your evidence (questions about this will become important later). Now consider the observer class "people who wake up in white rooms, seeing signs like this one." *i*O, for this observer class, is part of your evidence, as is the fact that no one else in this observer class has your evidence – you see a 3 on your hand, whereas they do not. Following Bostrom, I'll denote the fraction of sims in the observer class as f_{sim}, and I'll call a world where $f_{sim} = \sim 1$ a "high fraction" world.[137]

Because *i*O is weaker than *i*E, we can imagine conditioning the prior solely on *i*O and on f_{sim} being some fraction x, while ignoring the rest of our evidence. Thus, for example, in the case above, we can imagine forgetting about the number on your hand, and simply asking: "Conditional on being a person who wakes up in a white room seeing a sign like this one, and conditional on 90% of such people being sims, what's the probability that you're a sim?". The intuitive case for Bostrom's "bland indifference principle" (BIP) above begins with the idea that at the very least, in this sort of circumstance, your answer should be 90% (formally: $\Pr(i\text{S} \mid i\text{O} \text{ and } f_{sim} = x) = x$). Some might dispute this, but I won't do so here.[138]

Importantly, though, for Bostrom's full BIP to hold, it also needs to be the case that adding in the rest of your evidence doesn't make a meaningful difference to this probability (formally, and assuming "no difference" instead of "no meaningful difference" for the sake

---

[137] When the specific ratio of sims to non-sims matters, I also sometimes focus directly on that. But usually talking about the fraction being ~1 is sufficient.
[138] See Weatherson (2005) for some relevant worries.

of simplicity, $\Pr(\text{S} \mid \text{E and } f_{sim} = x) = \Pr(\text{S} \mid \text{O and } f_{sim} = x)$). Let's say that an observer class is "admissible" if this further condition holds.[139] This allows us to capture Bostrom's condition that "we don't have any information that indicate [sic] that our own particular experiences are any more or less likely than other human-type experiences to have been implemented in vivo rather than in machina."

Thus, in the case above, the observer class "people who wake up in white rooms, seeing signs like this one" is admissible: granted that you're a member of this observer class, adding in the rest of your evidence – including the fact that you have a 3 on your hand – does not alter your probability on being sim (because you're equally likely to draw a 3, conditional on being a sim vs. a non-sim). If you condition on the truth of the sign's claims, then, it seems very plausible that you should have high overall credence on being a sim. Bostrom wants to say that our own epistemic position conditional on $f_{sim} = \sim 1$ is in some sense analogous.

In my opinion, one of the most unnecessarily confusing parts of Bostrom's argument is that it focuses on one very specific observer class (namely, human-type observers) and one specific type of simulation (namely, ancestor simulations) – a choice that can seem arbitrary, which fails to address the question of how far reasoning of this broad type extends, and which raises questions about whether the argument will fall victim to the types of problems that plague appeals to "reference classes" in other contexts – for example, anthropic reasoning.[140] In fact, I think, the best formulation of the argument does not single out a particular observer class; rather, it applies to *any* observer class that satisfies a certain set of conditions, namely:

---

[139] I borrow the term "admissibility" from Thomas (2021), though he also includes a further condition in his definition – namely, that the expected ratio of sims to non sims, conditional on being a non-sim, doesn't alter as you move from conditioning on being in the observer class to incorporating all of your evidence. But I don't think this is necessary for my version of the argument.

[140] See discussion in the footnotes at the beginning of this section.

- *Membership*: The fact that you're in O is part of your evidence.[141] (Formally: $\Pr(i\mathrm{O} \mid i\mathrm{E}) = 1$)
- *Observer-class indifference*: Conditional only on being in O and on the fraction of sims in O being x, your credence on being a sim should be x. (Formally: $\Pr(i\mathrm{S} \mid i\mathrm{O} \text{ and } f_{\text{sim}} = x) = x$)
- *Admissibility:* This credence shouldn't change once you incorporate the rest of your evidence. (Formally: $\Pr(i\mathrm{S} \mid i\mathrm{E} \text{ and } f_{\text{sim}} = x) = \Pr(i\mathrm{S} \mid i\mathrm{O} \text{ and } f_{\text{sim}} = x)$)

As I see it, the core upshot of the most interesting version of the simulation argument comes from the following constraint on your credences:

> *Core constraint*: For *any* observer class that satisfies *Membership*, *Observer-class indifference*, and *Admissibility*, you cannot assign non-trivial probability to being a non-sim in a world where almost everyone in that observer class is a sim (that is, to the conjunction of iNS and $f_{\text{sim}} = \sim 1$).

As a purely formal matter, *Core constraint* looks good to me. By *Observer class indifference*, conditional only on iO and $f_{\text{sim}} = \sim 1$, your credence on iS should be ~1. So by *Admissibility*, conditional on *all* your evidence and on $f_{\text{sim}} = \sim 1$, your credence on iS should be ~1 as well.[142] Thus, there's no room left for non-trivial credence on the conjunction of iNS and $f_{\text{sim}} = \sim 1$.

---

[141] We can also formulate versions of the argument that start only with the claim that being in O is merely high probability conditional on your evidence; but for simplicity, I'll just treat your membership in O as certain. Note, though, that this certainty needs to remain compatible with whatever conception of evidence we adopt (such that, e.g., if you say that your evidence just consists in your experiences, then the observer class needs to be defined in terms of your experiences). For this reason, I'll generally focus on observer classes defined in terms of experiences in particular.

[142] Indeed, strictly speaking we do not need *Membership* as a condition. I include it, though, because the fact that you are a member of relevant observer class is central to the intuitions that simulation arguments draw on, and because I think it's easier to think about the step from *Observer-class indifference* to *Admissibility* (and in

Consider, for example, the hypothesis that the ratio R of sims to non-sims in the observer class (where $R = \frac{sims\ in\ the\ observer\ class}{non\text{-}sims\ in\ the\ observer\ class}$) is 999,999. For whatever credence you have on R = 999,999, conditional on your evidence, only one millionth of that credence can go to being a non-sim; or put another way, whatever credence you have on the conjunction of *i*NS and R = 999,999, you need to have 999,999 times more on the conjunction of *i*S and R = 999,999. Thus, the *maximum* credence you can put on *i*NS and R=999,999, conditional on your evidence, is one in a million -- and this requires certainty to that R=999,999. If you're only at, say, .2 on R=999,999, then the maximum credence you can put on *i*NS and R=999,999 is one in five million (that is, a millionth of .2). And if $R=10^9$, or 10^{15}, the constraint in question is all the stricter.

On its own, I see *Core constraint* as uncontroversial. But it's also, on its own, relatively uninteresting. In particular, we haven't yet said anything about which observer classes satisfy all of *Membership*, *Observer-class indifference*, and *Admissibility*, or about whether those observer classes are such that we have reason to take the hypothesis that f_{sim} = ~1 at all seriously. Thus, for example, consider the observer class "observers with experiences exactly identical to your own." You might well grant that this sort of observer class satisfies all three of these conditions, and thus that you cannot place meaningful credence on being a non-sim in a world where most people with exactly your experiences are sims. But you don't have any obvious motivation for taking seriously the hypothesis that there are a large number of observers having *exactly* your experiences, most of whom are sims – a hypothesis much more exotic than e.g. the hypothesis that technologically mature civilizations create lots of

particular, about the question of whether, conditional on *i*O and on f_{sim} = ~1, *i*E is more likely conditional on being a non-sim vs. a sim) if your membership in the observer class remains constant.

ancestor simulations (or other simulations) more generally.[143] Even a highly detailed ancestor simulation, after all, would presumably not replicate a historical non-sim's experience *exactly* – a point that Bostrom concedes.[144] So you can plausibly accept *Core constraint* about an observer class like "observers with experiences exactly identical to your own," without making very meaningful alterations to your everyday worldview overall.[145]

An interesting simulation argument should do more than this. In particular, it should wield *Core constraint* in a way that forces us to substantively revise some aspect of our everyday beliefs. Bostrom, I think, is trying to do this, but the most natural interpretation of what he's doing also adds additional structure and ambition to the reasoning involved – structure and ambition that I think muddles the impact of his core insight, and which opens him up to objections he doesn't have to worry about.

IV. Type 1 and Type 2 arguments

---

[143] Some anthropic principles, like the Self-Indication Assumption ("SIA"), update towards worlds where there are a lot of people with exactly your experiences (see Bostrom (2002a)); and simulations seem like a salient way for there to be a lot of those. But I am setting aside this issue for the moment.

[144] As does Chalmers (2022); see p. 97.

[145] This is my central objection to the formulation of the argument in Chalmers (2022), which is otherwise quite similar to my own. Chalmers defines a "sim sign" as a feature that raises the probability that a creature is a sim (for example, seeing glitches in physical reality), and a "non-sim sign" as a feature that raises the probability that a creature is not a sim (for example, the size and complexity of our universe, assuming that such universes are less likely to be simulated). He then defines a "humanlike" being as a being with "roughly the same major sim signs and nonsim signs as humans," and a "sim blocker" as something that prevents the creation of enough humanlike sims to ensure that most humanlike beings will be sims (p. 97). (Thus, for Chalmers, Bostrom's 1 and 2 above are just examples of sim blockers, as are possibilities like "sims aren't conscious" and "sims take too much computational power" – the denial of which Bostrom builds into his argument as assumptions.)

Equipped with these definitions, and assuming something like the BIP above, Chalmers then argues that:

I. "If there are no sim blockers, most humanlike beings are sims.
II. If most humanlike beings are sims, we are probably sims.
III. So: if there are no sim blockers, we are probably sims" (p. 98).

Here, "human-like" is functioning in a manner similar to "satisfying *Membership*, *Observer-class indifference*, and *Admissibility*"; and in this sense, Chalmer's argument is actually just a purely formal statement of *Core constraint* above. But the purely formal argument leaves open what sorts of observers count as human-like in the relevant sense (despite the fact that the text elsewhere suggests that Chalmers has a particular sort of observer class in mind); and so it doesn't, on its own, say enough to require us to make any major revisions to our everyday worldview.

To see this, let's focus, for the moment, on the sort of observer class Bostrom focuses on – namely, humans having the experience of living in the early history of their civilization, prior to some sort of technological maturity (call these "early-seeming people") – and on the specific type of simulation that Bostrom focuses on – namely, ancestor simulations. I'll discuss other observer classes, and other simulations, later.

I want to distinguish between two formulations of a Bostrom-like argument – what I'll call a Type 1 version, and a more minimal, Type 2 version. I'll suggest that the Type 2 version is superior.

The core difference between a Type 1 and a Type 2 argument is that the former, but not the latter, treats some set of empirical claims as likely, given our evidence. I'm interested, in particular, in the set of claims about physics, neuroscience, cosmology, and computer science that Bostrom uses to argue for the claim that "Posthuman civilizations would have enough computing power to run hugely many ancestor-simulations even while using only a tiny fraction of their resources for that purpose" (p. 6). The precise content of these claims does not matter much for our purposes, but for concreteness, they include claims like:

- A realistic simulation of the experience generated by a human can be created using approximately 10^{14}-10^{17} operations per second.[146]
- "The main computational cost in creating simulations that are indistinguishable from physical reality for human minds in the simulation resides in simulating organic brains down to the neuronal or sub-neuronal level."

[146] See p. 4. This is the estimate for human brain computation that Bostrom uses in his calculation of the overall computational cost of simulating all of human history thus far.

- "We can use 10^{33}-10^{36} operations as a rough estimate" of the cost of a realistic simulation of human history (this is assuming that detailed simulation of the environment is not required).
- "A rough approximation of the computational power of a planetary-mass computer is 10^{42} operations per second."[147]

Let's call this set of claims *Comp* (for "computer power"). A Type 1 version of Bostrom's argument treats *Comp* as highly likely, given our evidence (that is, it accepts that Pr(*Comp*|*i*E) = ~1).[148] It runs as follows:

A Type 1 version of a Bostrom-like argument:

I. *Comp* is likely, given our evidence. (Formally: Pr(*Comp*|*i*E) = ~1)

II. Conditional on *Comp*, at least one of the following is true: 1 (very few civilizations reach technological maturity), 2 (very few technologically mature civilizations use much of their resources running ancestor sims), or most early-seeming people are sims. (Formally: Pr(1, 2, or f_{sim} = ~1 | *i*E and *Comp*) = 1)

III. The observer-class "early-seeming people" satisfies *Membership*, *Observer-class indifference*, and *Admissibility*.

IV. Thus, conditional on most early-seeming people being sims, we should have high credence on being sims. (Formally: Pr(*i*S | *i*E and f_{sim} = ~1) = ~1).

V. Thus, it's very likely that at least one of the following is true: 1, 2, or *i*S (Formally: Pr(1, 2, or *i*S|*i*E) = ~1).[149]

---

[147] These quotes are all from section III, p. 3-6.
[148] It also accepts that Pr(Sims can be conscious|*i*E) = ~1, but for simplicity I'm going to pass over this in what follows.
[149] As discussed earlier, Bostrom's formulation of the conclusion is that at least one of 1, 2, or "we're probably sims" is true; but I've altered the formulation here to avoid confusions about conditional and unconditional probabilities. I'm also generally assuming, for simplicity, that ~1*~1 = ~1.

In a more visual form, the argument's structure looks like:

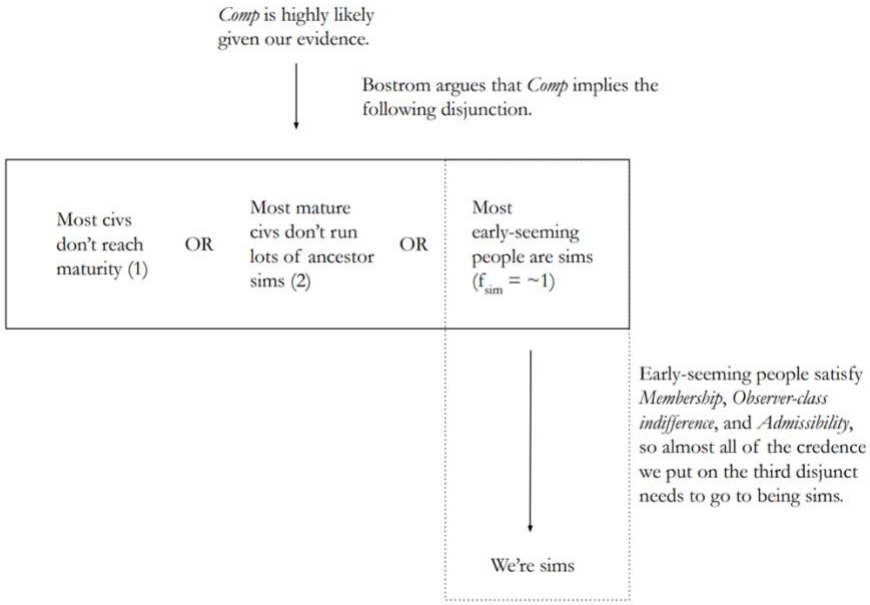

Figure 1: A Type 1 version of a Bostrom-like argument

A Type 2 formulation, by contrast, has less structure. It runs as follows:

A Type 2 version of a Bostrom-like argument:

VI. The observer-class "early-seeming people" satisfies *Membership*, *Observer-class indifference*, and *Admissibility*.

VII. Therefore, it's very likely that one of the following is true: it's not the case that most early-seeming people are sims, or we're sims. (Formally: $\Pr(f_{sim} \neq \sim 1 \text{ or } i\mathcal{S} \mid i\mathcal{E}) = \sim 1$)

That is, in picture form:

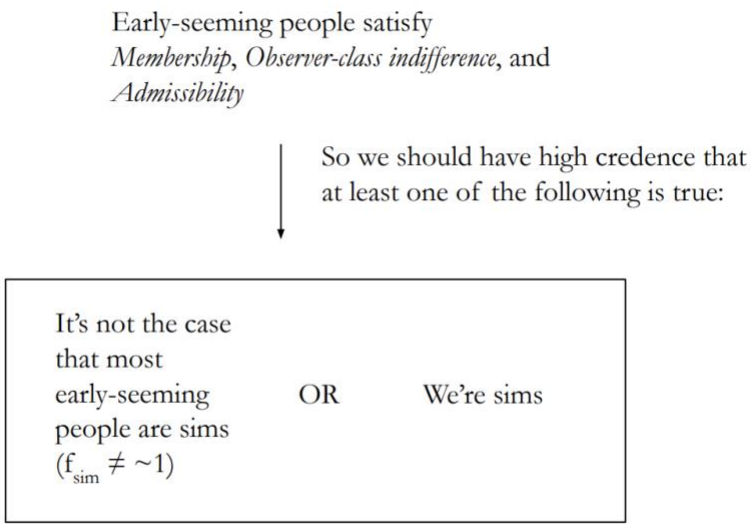

Figure 2: A Type 2 version of a Bostrom-like argument

Bostrom's is a Type 1 argument – or at least, the text of the paper strongly suggests a Type 1 interpretation. In particular, Bostrom (2005b) writes that "The simulation argument relies crucially on non-obvious empirical premises about future technological abilities" (p. 95), and the structure of the paper closely mirrors the Type 1 structure above. And indeed, many in the literature – e.g. Birch (2013), Besnard (2004), Garfinkel (unpublished), Tegmark (2014) – have read Bostrom in a roughly Type 1 way.[150]

Type 1 arguments, though, face objections. I'll focus on two: what I'll call the "selective skepticism" objection and the "self-undermining" objection. These objections aren't

[150] There are alternative interpretations available, on which Bostrom *starts* by assuming that *Comp* is likely given our evidence, but then shows that we are forced to either revise that assumption, or to accept his disjunction. In my opinion, this interpretation is a worse fit with the paper, with the framing of its conclusion, and with the level of confidence in *Comp* that he displays overall, but I don't want to focus, here, on exegetical questions about how Bostrom is most accurately and/or charitably interpreted (and I think it's possible that at the time of writing the paper, Bostrom, himself, hadn't fully worked out precisely what sort of argument he was trying to run).

necessarily fatal to a Type 1 argument, but they can get confusing – and the confusion they create, I'll suggest, isn't necessary. We should just focus on Type 2 arguments instead.

V. Selective skepticism

The selective skepticism objection to Type 1 arguments runs as follows.[151] Consider some further set of claims which, if you include them in your evidence, cause *Admissibility* to fail; let's call claims of this form "admissibility blockers." Clear examples here might include: "I have two unsimulated hands," "no ancestor sims will see exactly the books on my bookshelf," and "if there are any sims, they're all in my future" – since if any of these claims are included in your evidence, then even conditional on most early-seeming people being sims, your evidence rules out being a sim, and thus *Admissibility* fails.

Admissibility blockers need not be directly simulation oriented. Thus, for example, Bostrom's calculations of the computational costs of running ancestor sims assume that simulating the necessary environment is a negligible portion of the overall computational burden – but he proceeds with this assumption only after first acknowledging that the environment in question may need to be simulated at only a very low level of resolution, with much left out, for the project to be computationally realistic:

> "Simulating the entire universe down to the quantum level is obviously infeasible, unless radically new physics is discovered. But in order to get a realistic simulation of human experience, much less is needed – only whatever is required to ensure that the simulated humans, interacting in normal human ways with their simulated environment, don't notice any irregularities. The microscopic structure of the inside of the Earth can be safely omitted. Distant astronomical objects can have highly compressed representations: verisimilitude need extend to the narrow band of properties that we can observe from our planet or solar system spacecraft … Should any error occur, the director could easily edit the states of any brains that have become aware of an anomaly before it spoils the simulation" (p. 5).

[151] The term "selective skepticism" comes from Birch (2013). Garfinkel (unpublished) makes a similar argument; and I see Thomas's (2021) discussion of how to choose the right observer class as wrestling with some similar tensions.

Here, Bostrom suggests that the ancestor simulations he has in mind are centrally what are sometimes called "short-cut simulations" – that is, simulations that do not fully reproduce the empirical dynamics of the universe of the early-history civilization, but which instead only do so to the extent required to fool the inhabitants of the sim.[152] In this sense, if all you know is that you're an early-seeming person in a world where almost all early-seeming people are ancestor sims of the sort Bostrom has in mind, but then you learn that e.g. the stars you see in your telescopes are real, or that the microscopic structure of the inside of your planet exists, or that the universe you see around you contains quantum phenomena that would be extremely computationally expensive to simulate, this is a strong update towards being a non-sim – and in this sense, even such mundane, everyday scientific claims would be admissibility blockers as well.[153]

The selective skepticism objection argues that:

> *Parity of evidence*: Some admissibility blockers are on evidential footing that is comparable to or stronger than the claims in *Comp*.

That is, faced with a Type 1 argument, *Parity of evidence* suggests that the overall situation looks like this:

---

[152] See Chalmers (2022), p. 94-96, for more discussion. In my opinion, this dimension of Bostrom's discussion is underemphasized. That is, in my experience, casual readers of Bostrom often end up assuming that if they are sims, they live in a simulated world where their conventional scientific worldview is at least true *of that world*. Actually, though, the sorts of simulations that Bostrom has in mind are much more revisionary – e.g., they involve fake stars, fake scientific experiments, possible edited-memories if anyone ever notices something amiss, and so on.

[153] Whether they would block *Admissibility* enough to leave you overall confident that you're a non-sim in a world with lots of ancestor sims is a further question (e.g., there could be some very computationally expensive ancestor simulations, that do in fact incorporate stars, quantum phenomena, etc). But it's not central to the present dialectic.

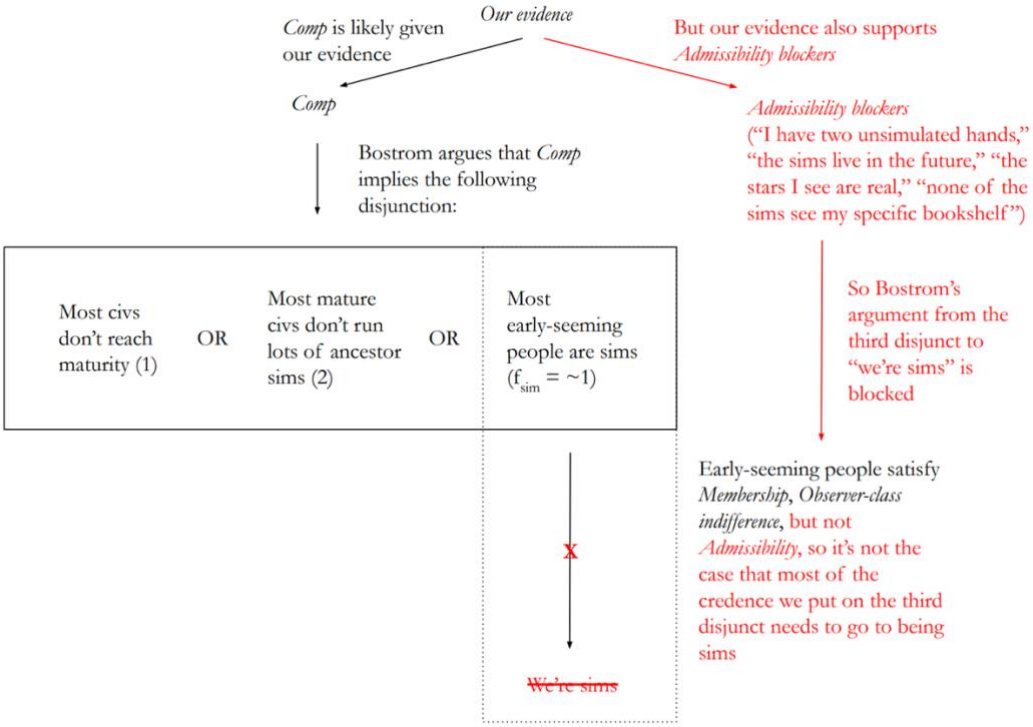

Figure 3: Selective skepticism objections to a Bostrom-like Type 1 argument

Now, by everyday standards, it does indeed seem that claims like "the stars I see are real" are on competitive (indeed, actively superior) evidential footing relative to claims like "A rough approximation of the computational power of a planetary-mass computer is 10^{42} operations per second." Whether the same is true of claims like "all the sims live in the future" or "none of the sims will see my specific bookshelf" is less clear, but we can idealize the real-world case to make evidential parity more plausible. Thus, consider:

> *Imperfect ancestor sims*: You are a mediocre office worker living in a pre-post-human civilization, who doesn't tend to think about big picture topics much. Rather, you get most of your information on them from the stable global government's official scientific authorities, who thus far have never (to your knowledge) led you astray.

One day, you're sitting by your bookshelf and watching TV, when a broadcast from the official scientific authorities comes on. They say:

"People of Earth: we have a few announcements:

- Announcement 1: Our super-duper forecasters are saying that it's 99.9% likely that we're going to make it to technological maturity, that this government is going to remain stable until then, and that whatever intentions we commit to now will be enacted later.
- Announcement 2: The universe is finite, and we're the only life that exists anywhere or that will ever develop on its own. Also, the physics of the universe we see around us involves quantum phenomena that would be extremely computationally expensive to simulate, the stars we see in our telescopes are real, and the physical microstructure of the earth continues to exist when no one is looking.
- Announcement 3: It'll take around 10^{36} operations to run a detailed simulation of all of our cognitive history – an 'ancestor simulation.' (That is, assuming we take lots of shortcuts on the environment and don't simulate quantum phenomena, stars, galaxies, or the earth's physical microstructure – if we had to do that, running ancestor sims would be out of the question.) And we're going to have tons of planet-sized computers that can run at least 10^{42} operations *per second*, so running tons of detailed ancestor simulations is going to be extremely easy in the future.
- Announcement 4: In light of this, we've decided to make the following binding commitment: come technological maturity, we're going to run a billion such ancestor simulations (but that's it: no more, no less). However,

we're not going to be able to recreate our history *exactly*. Rather, lots of details are going to be forever lost. For example, none of the sims are going to have the same books on their shelves as you do.

- Announcement 5: We're going to turn off the ancestor simulations before they reach technological maturity – so even though they will likely *think* that they're going to run simulations themselves (and their global governments will make announcements to this effect), they actually won't."

Obviously, this case differs from our own situation in a number of respects: for example, the purported stability of the government, the purported finiteness of the world, the weirdly specific intentions with respect to future simulations, and so on. Structurally, though, it's sufficiently similar to our own situation that it seems like Bostrom's argument, if it works, should apply here as well.

In this case, though, *Comp* (here given in Announcement 3) is just one amongst many government announcements – and the other announcements are admissibility blockers. If you believe the whole of the government's story, then, you end up confident that you live in world where $f_{sim} = \sim 1$, but where you're a non-sim, which is the sort of thing Bostrom is trying to rule out.

Applying a Type 1 argument to this case, then, requires believing *some* of the government's story (namely, the *Comp* parts), but not the rest. But why would you do that? They come, after all, from the same source – aren't their credentials similar? And we might say the same about the real-world case as well. Whence your confidence in the computational power of a planetary-mass computer, if not from the same sources that gave you confidence that the stars you see are real? Thus the charge of "selective skepticism."

VI. Self-undermining

The self-undermining objection is related but distinct.[154] It runs as follows. Consider:

> *Sim ignorance*: Conditional on being sims, it's not the case that our evidence strongly supports *Comp*. (Quasi-formally: Pr(*Comp*|*i*S and *i*E) is a good bit lower than 1).

Sim ignorance seems intuitively plausible, especially once we bring to mind that for Bostrom's argument to work, the claims in *Comp* need to specifically apply to the level of reality simulating us – a place that we, if we're sims, have never seen, touched, or been to.[155]

But *Sim ignorance* is in tension with the structure of a Type 1 argument – or at least, with the credence on *i*S that the argument is supposed to argue for. That is, a Type 1 argument works by arguing that *Comp* is very likely given our evidence, then arguing that our credence *on Comp* needs to be divided between 1, 2, and f_{sim} = ~1, and then arguing that the portion of *that credence* that goes to f_{sim} = ~1 needs to go almost entirely to *i*S as well. Thus, the credence that ends up on *i*S, as a result of the argument, is specifically credence on the conjunction of *Comp* and *i*S. That is, to the extent a Type 1 argument tells us to put credence on being sims, it tells us to put credence specifically on being sims in worlds where *Comp* is true of the level simulating us. But if that's *all* the credence on being sims that we end up with, then *Sim ignorance* fails: conditional on our evidence and on being sims, *Comp* is treated as certain.

---

[154] See Tegmark (2014), p. 349 and Besnard (2004). Birch (2013) also gestures at something like a self-undermining objection in section 3.

[155] See e.g. Tegmark (2014): "I think the logical mistake happens at the very first step: if you're willing to assume that you're simulated, then as emphasized by Phillip Helbig, the computational resources of your own (simulated) universe are irrelevant: what matters are the computational resources in the universe where the simulation is taking place, about which you know essentially nothing" (p. 349).

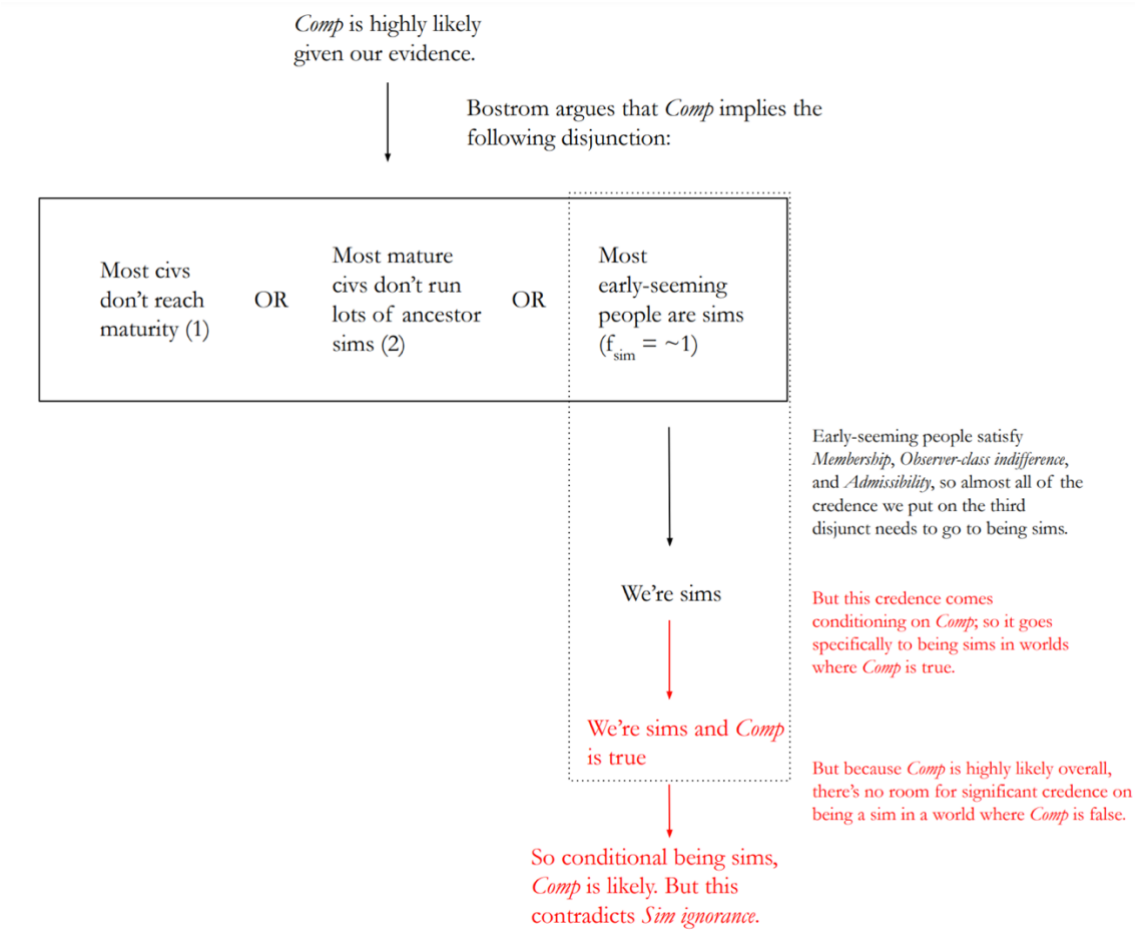

Figure 4: A self-undermining objection to a Bostrom-like Type 1 argument

What if we had some other credence on being sims as well, independent of the Type 1 argument? The tension with *Sim ignorance* still arises, though less cleanly. Suppose, for example, Pr(*Comp*|*i*E) = .99, and that all of your remaining .01 on not-*Comp* goes to being a sim. The degree to which you are allowed to be non-confident in *Comp*, conditional on being a sim and on your evidence, is now determined by your credence on being a sim, conditional on *Comp* and on your evidence. Thus, if Pr(*i*S|*Comp* and *i*E) is, say, 1/3rd, as Bostrom originally suggests, then Pr(*Comp*|*i*S and *i*E) is still 33/34, which is quite a bit higher than the intuition behind *Sim Ignorance* suggests. Indeed, the only way to drive Pr(*Comp*|*i*S and *i*E) down below, say, .5 is to put Pr(*i*S|*Comp* and *i*E) at less than 1/99, which is quite a bit lower than Bostrom wants to go.

VII. Arguing for *Admissibility* directly

Both selective skepticism and self-undermining objections, then, accuse Type 1 arguments of some kind of problematically confident relationship to *Comp*. Selective skepticism objections say that Type 1 arguments aren't justified in treating *Comp* as so likely, while ignoring other common-sensical admissibility blockers; and self-undermining objections say that Type 1 arguments aren't justified in treating *Comp* as so likely, conditional on us being sims.

Now, advocates of Type 1 arguments may have replies to these objections. Perhaps, for example, Type 1 arguments could argue against *Parity of evidence*, on the grounds that hypotheses to the effect that $f_{sim} = \sim 1$ undermine or defeat our evidence for claims like "the stars I see are real," but not our evidence for claims like "a rough approximation of the computational power of a planetary-mass computer (in unsimulated reality) is 10^{42} operations per second."[156] And perhaps they could reject *Sim ignorance*, despite its initial intuitive appeal, and say that actually, if we're sims, we're most likely to be sims who are right about claims like *Comp*.[157]

Below I'll discuss replies like this in a bit more detail. But I don't think we need to rest the viability of the simulation argument, in general, on the viability of replies like this, because we don't have to treat *Comp* as highly likely. Rather, we can run a Type 2 argument instead: one that just claims, directly, that *Membership*, *Observer-class indifference*, and *Admissibility* holds

[156] Some of the comments in Bostrom (2005b), for example, seems to me to suggest this sort of response.
[157] An alternative response to the self-undermining objection, offered by Chalmers (2022, appendices), is to try to run the argument by saying that *either Comp* is true, in which case an argument based on *Comp* can go through, *or* it's false, in which case we should have high credence on being sims (because the most likely way our evidence for *Comp* would be misleading is via our being sims). But this is no longer a Type 1 argument, because it no longer treats *Comp* as likely. It's also not clear that Chalmers is right that the most likely way for *Comp* to be false is if we're sims; and this response, on its own, doesn't clearly handle selective skepticism objections, which would still apply to the portion of your credence devoted to *Comp*.

for the observer class "early-seeming people," and thereby concludes, per *Core constraint*, that we can only have extremely low probability on the conjunction of iNS and $f_{sim} = \sim 1$ – regardless of our views on *Comp*.

Membership isn't in dispute, and I'm happy to grant *Observer-class indifference*, as are many of Bostrom's objectors (if *all you knew* is that you're an early-seeming person, and that x is the fraction of early-seeming people that are sims, it seems to me very reasonable to put x credence on being a sim). So the key claim here, as I see it, is that the observer class "early-seeming people" satisfies *Admissibility*: that is, recall, that if you had *all* your evidence and you knew that x is the fraction of early-seeming people that are sims, your credence on being a sim should be the same as if you only knew that you're an early-seeming person and that x is the fraction of early-seeming people that are sims (i.e. $\Pr(i\text{S} \mid i\text{E} \text{ and } f_{sim} = x) = \Pr(i\text{S} \mid i\text{O} \text{ and } f_{sim} = x)$). Let's look more closely at this claim now.

Admissibility falls out of *Membership* and *Observer-class indifference* if we also grant the following:

> *No update*: Conditional on being in the observer class, on the fraction of sims in the observer class being x, and on being a sim, the probability of your evidence is the same as it would be conditional on being in the observer class, on the fraction of sims in the observer class being x, and on being a non-sim. (Formally: $\Pr(i\text{E} \mid i\text{S} \text{ and } i\text{O} \text{ and } f_{sim} = x) = \Pr(i\text{E} \mid i\text{NS} \text{ and } i\text{O} \text{ and } f_{sim} = x)$

That is, once you know that you're in the observer class and that the fraction of sims in the observer class is x, the rest of your evidence doesn't tell you anything about whether you're a sim or a non-sim – rather, it's equally likely either way.[158] The aim here is to get at

[158] See also the discussion of Bayes factors in Thomas (2021, p. 6).

the more general analog of the sort of reasoning you do in *Sims with random numbers* about the evidence you get from the "3" you see written on your hand. In particular: because (on God's story at least, which we grant for the purposes of the case) you're equally likely to have ended up with a "3" written on your hand conditional on being a sim vs. a non-sim, adding in the fact that you have a "3" written on your hand doesn't update your probability on being a sim vs. a non-sim, once you've taken into account the fact that 90% of people who wake up in white rooms seeing the sort of sign you're seeing are sims. The claim is that the same sort of dynamic applies – at least roughly – to the rest of our evidence, once we update on being early-seeming people and on the fact that most early-seeming people are sims. That is, *Membership*, *Observer-class indifference*, and *No update* hold for the class early-seeming people, so *Admissibility* holds as well.

Is this true? To me it seems quite plausible – especially if we adopt a simple internalist conception of evidence on which two observers with the same experiences (e.g., you and your brain-in-the-vat equivalent) have the same evidence. Let's start by working with this conception, and then discuss what happens if we complicate it.

VIII. No update from your experiences

Let's call your full set of experiences Q, and let *i*Q denote the proposition that you have experiences Q. Suppose that we grant:

> *Simple internalism*: Your evidence is constituted by your experiences. (Formally: *i*E = *i*Q)

In that case, to get *Admissibility*, we only need a weaker principle, namely:

> *No update from your experiences*: Conditional on being in the observer class, on the fraction of sims in the observer class being *x*, and on being a sim, the probability

> of having your *experiences* is the same as it would be conditional on being in the observer class, on the fraction of sims in the observer class being *x*, and on being a non-sim. (Formally: $\Pr(i\text{Q} \mid i\text{S}$ and $i\text{O}$ and $f_{\text{sim}} = x) = \Pr(i\text{Q} \mid i\text{NS}$ and $i\text{O}$ and $f_{\text{sim}} = x)$

Conditional on *Membership* and *Observer-class indifference*, this principle gets you to high credence on being a sim, conditional only being in the observer class, most people in the observer class being sims, and having your experiences in particular. And then *Simple internalism* says that having your experiences in particular is all the evidence you've got. So conditional on all your evidence and on most people in the observer class being sims, you're still at high credence on being a sim.

Should we accept *No update from your experiences*, applied to the observer class "early-seeming people"? To me it seems plausible that we should, at least in fairly idealized cases like *Imperfect ancestor sims*. Thus, e.g., it's not as though the books on my particular bookshelf are any more likely, on priors, to show up on the bookshelf of a sim vs. a non-sim; rather, they're just a set of 21st century books, with nothing, on priors, especially sim-y or non-sim-y about them. In this sense, I should treat my seeing these books in particular the same way I treat seeing the number "3," in particular, in *Sims with random numbers*. And the same, plausibly, can be said of all my experiences, once we condition on $i\text{O}$ and $f_{\text{sim}} = \sim 1$ – that is, that these experiences are not as any particular indication of sim-hood vs. non-sim-hood.

We can restate the intuition here, and the eventual upshot, in terms of your absolute priors over having different experiences in different scenarios. Thus, consider a more general version of the description of the world offered by the government in *Imperfect ancestor sims*. On this description, there are *n* early-seeming non-sim people on a non-simulated planet

(call this planet 0), and a billion sets of *n* early-seeming sim people, each on a different simulated planet (call these planets 1-1,000,000,000), all of whom hear announcements from their government like announcements 1-5 above, and all of whom have fake stars, simulated hands, and no sims in their future. Let's call any world that fits this description a "Z world," and let's call the proposition that you have experiences Q on planet *y* in a Z world iZQ_y.

One way of drawing out the intuition in favor of *No update from your experiences*, applied to a case like *Imperfect ancestor sims*, is to note that on priors, and conditional on living in a Z-world, it seems equally likely that you have experiences Q on any one of these planets. That is, $Pr(iZQ_0) = Pr(iZQ_1) = Pr(iZQ_2)$ … If your experiences Q are, for example, hearing a particular set of government officials announce that you aren't a sim but that there are lots of sims in the future, these experiences simply aren't any more likely to show up on the non-sim planet vs. any given sim planet. By hypothesis, in Z-worlds, *all the governments* (sim and non-sim) announce that their listeners are non-sims but that there are lots of sims in the future. Thus, finding yourself with experiences Q *can't* simultaneously update you towards high probability on living in a Type Z world, *and* high probability on being a non-sim – all of your credence on Type Z worlds needs to stay split equally between each of iZQ_n, and there are a more than a billion of these, all mutually incompatible, in play. And if $iE = iQ$, then your experiences are all the evidence you have. So your posterior credence on iZQ_0 (the only Z-type scenario where you're a non-sim) is (dramatically) capped.

Overall, then, *No update from your experiences* looks pretty good to me, at least in cases like *Imperfect ancestor sims*.[159] So if we accept *Simple internalism*, then *No update* follows, as does *Admissibility*.

IX. Reject *No update*?

To my mind, the most philosophically interesting way of denying *Admissibility* is to grant *No update from your experiences*, but to deny *No update*. Thus, for example, in a case like *Imperfect ancestor sims*, and conditional on living in a Z-type world, you can concede that your experiences are equally likely to occur on the non-sim planet vs. any given sim planet (that is, $Pr(iZQ_0) = Pr(iZQ_1) = Pr(iZQ_2)\ldots$), but deny that you would have the same evidence, if you were having those experiences as a sim vs. a non-sim. In particular, you might say that if you're a non-sim, then one or more of the admissibility blockers we discussed above (e.g., "I have two unsimulated hands," "the stars I see are real," "all the sims live in the future," "none of the sims see my bookshelf," etc) are either included in your evidence, or supported by it in a way that simple internalism cannot account for. And indeed, in the real world, we tend to treat claims like "the stars I see are real" like they are on quite solid evidential footing. Type 2 arguments, then, plausibly require denying this sort of common-sense, at least if we're also going to put substantive credence on most early-seeming people being sims in whose mouth such claims are false.

In response to objections of this kind, Bostrom and Chalmers both argue that if you learn that you live in e.g. a Z world, this makes it the case that claims like "the stars are real" or "I have two unsimulated hands" are no longer a part of your evidence or strongly supported

[159] Here I'm assuming your experiences are reasonably typical of an early-seeming person. If you're Donald Trump, it's a somewhat more complicated story (see Chalmer's (2022) on "sim signs"), but I won't try to get into that here.

by your evidence, even if they might have been in other circumstances.[160] And this seems plausible to me. Maybe I can normally be confident that I am not an envatted brain. But if I learn that the aliens are going around envatting many people's brains while they're sleeping – enough, indeed, that most human brains are envatted -- then this confidence needs to alter.

Now, strictly, this is not enough for the present purpose, because you may not have learned that you live in a Z world – rather, you may only have some positive credence *c* that you live in a Z world.[161] But if we accept the claim that *if* you condition on living in a Z world, then you should have high credence on being a sim (just as, if you condition on the aliens envatting almost everyone, you should have high credence on being a BIV), then it seems like whatever overall credence *c* you put on living in a Z world, conditional on your evidence, should mostly go to being a sim – and our conception of evidence will need to accommodate this.

Wading too deeply into these waters, though, is beyond my purpose here.[162] I'm happy to grant that there is theoretical daylight between *No update from your experiences* and *No update*, that various epistemologies may accept the former but resist the latter, and that the simulation argument works most smoothly against the backdrop of epistemologies that are

---

[160] Bostrom (2005): "I would claim that given [$f_{sim} = \sim 1$], we have grounds for concluding that we are in just such a special circumstance in which illusions are ubiquitous and in which we should distrust our senses in regard to one particular (narrowly circumscribed) set of facts, namely, facts that have to do with how we are physically implemented… Thus I would maintain that externalist epistemology, of any reasonable stripe, should regard [$f_{sim} = \sim 1$] as implying a case such that if we knew on theoretical grounds that this was the actual case, then we should not take our perception of two hands as giving us strong reason to think that they are two non-simulated hands." (p 95-6). Similarly, Chalmers (2022) writes in the online appendices to Reality+: "Even most externalists allow that perceptual evidence (e.g. seeing a zebra) can be defeated by other evidence (e.g. knowing that most zoos contain holograms). When we grant that 90% of beings with evidence like ours are sims, this in effect overwhelms any evidence provided by our being nonsims, so that we should be 90% confident that we are sims. An externalist of this sort can endorse the key indifference principles that we have been working with. I think that reflection on the cases we have discussed recommends this view" (p. 12). Thomas (2021) also discusses this sort of response to externalist-flavored objections.

[161] This is a point made by Thomas (2021).

[162] See Thomas (2021) for some additional discussion.

fairly happy to grant the latter once the former is in place (*Simple internalism* is an especially clear example).[163]

That said, I do want to note that especially once you grant *No update from your experiences*, Type-2-style reasoning starts to take on, at least for me, pretty strong intuitive force, such that I start to feel like we should be actively *seeking* epistemic principles that allow us to capture this force, rather than looking for ways to resist it. For me this sort of intuition is especially vivid in the case where the sims have genuinely indistinguishable experiences. Thus, consider:

> *Indistinguishable sims:* You are a simulation scientist. You've been working on a technology that will scan a non-sim's body and brain, and then create sims with experiences subjectively indistinguishable from those of the scanned non-sim. The scanner operates by continuously scanning anyone who is in a certain white room

---

[163] If we open ourselves to doubting beliefs like "I have two unsimulated hands" and "the stars I see in my telescope are real," though, does the simulation argument retain its dialectical interest relative to discussions of more standard skeptical threats in the literature? Yes. Simulation arguments rest specifically on claims about how to relate epistemically to hypotheses where most observers of a certain type are in a skeptical scenario, and where, at least naively, we have some empirical reason to take seriously a hypothesis of this form. More standard skeptical discussions have neither of these features. And while simulation arguments require that on priors, and conditional on living in a Z world with experiences Q, you're equally likely to live on any of the planets (e.g. $Pr(iZQ_0) = Pr(iZQ_1) = Pr(iZQ_2)\ldots$), they do not include any comparable constraints relating your prior on being a brain in a vat with experiences Q (call this *i*BIVQ) to your prior on having experiences Q in worlds with no BIVs at all. Even after having experiences Q, then (and even accepting *Simple internalism*), a hypothesis like *i*BIVQ can remain arbitrarily less likely than a hypothesis like "I'm a non-sim in a Z world."

That said, I think it's an interesting question whether assigning substantive credence to being sims, on the basis of Type 2 simulation arguments, should lead us to assign substantive credence to other skeptical scenarios as well. One argument for this might appeal to an intuition like:

> *Rough wackiness parity*: Conditional on being in a skeptical scenario as wacky as being in a simulation (call this *i*W), I should have substantive credence on being in a wacky non-sim skeptical scenario (*i*WNS).

Prior to considering simulation arguments, this would've seemed to me quite plausible: either my situation is "normal," or it's wacky – but if it's wacky, it could be wacky in tons of different ways, most of which I'm probably not considering. And *Rough wackiness parity* would require us to raise our credence on other skeptical scenarios a lot (assuming it was very low before), if we raise our credence on being sims. But *Rough wackiness parity* does not seem like an especially problematic intuition to give up – in light of simulation arguments, we might just have pretty strong evidence that if we're in a wacky skeptical scenario, it's probably a simulation. That said, Type 2 arguments don't take a stand on this issue.

> in your lab, such that it can recreate any of the experiences that occurred while inside. Inside this room you've placed a red button, with a sign on it that says: "If you are a non-sim, this button will create a billion sims with experiences exactly like yours, facing a button and a sign that look just like this one. If you're a sim, though, pressing the button won't actually create any new sims – that would take too much computational power." You enter the white room. You are currently planning to press the button.

If I were in this case, I would feel intuitively uncomfortable resting easy with the belief that I'm a non-sim about to press the button.[164] And an epistemic procedure that licenses such confidence would lead the sims, at least, very much astray – sims that I would be actively expecting to share the world with (not some purely hypothetical set of sims, living in a possible world I have no reason to put much credence on). And I would feel this same discomfort if, say, I thought the button was going to be pressed with only 10% probability (say, if a ten-sided dice came up 1) – that is, I would feel like I couldn't put 10% on the button getting pressed, while also putting ~100% on being a non-sim.

And if we grant such discomfort, should it make a difference whether the experiences in question are *exactly* the same? Consider:

> *Sims with different light speckles*: You're in the same set-up as above, except that the scanner is *slightly* imperfect. In particular: it can't exactly reproduce, in the sims, the specific patterns of random light speckles in the visual field of the non-sim.

[164] And indeed, various skeptics of simulation arguments – for example, Birch (2013) and Garfinkel (unpublished) – seem sympathetic to the logic in cases like this, where the experiences of the sims and non-sims are genuinely indistinguishable.

> Rather, the sims see their own, distinct random patterns, which the non-sim never saw.

I would feel the same discomfort, here, about thinking that I'm a non-sim with a billion sims in my future. And not feeling such discomfort would suggest a strange discontinuity between the two cases. Suppose, for example, that as you're striding confidently towards the button in *Sims with different light speckles,* you notice a little note from one of your grad students pinned to the scanner, which says: "I fixed the scanner! Now it perfectly captures the non-sim's light speckles." Should that note really cause meaningful change to your willingness to believe that you're a non-sim about to create a billion sims? I'm skeptical. So granted inability to rest easy with "I'm a non-sim in a sim-filled world" in cases like *Indistinguishable sims*, it seems like the same inability should apply in *Sims with different light speckles* as well. Structurally, though, *Sims with different light speckles* looks very similar to *Imperfect ancestor sims*. So it seems, to me, like the same sort of dynamic should apply to all of these cases. Type 2 arguments capture it.

For this reason, even in the absence of a worked-out epistemology that tells us whether or not to make the transition from *No update from your experiences* to *No update* and thus to *Admissibility*, Type 2 arguments seem to me independently forceful and attractive. That is, to me it seems quite plausible that even for fairly inclusive observer classes like "early-seeming people," *Membership*, *Observer-class Indifference*, and *Admissibility* will hold, and thus, that the upshot of *Core constraint* will apply: you can't put substantive credence on the conjunction of *i*NS and $f_{sim} = \sim 1$.

X. Where should we end up overall?

What happens, though, if we start taking this seriously? For the remainder of this paper, I want to examine some of the complexities and uncertainties that arise if we accept the basic logic of Type 2 arguments. In particular:

- How should we adjust our overall credences – including our credence on being sims -- in light of Type 2 arguments?
- To what range of cases and observer classes do Type 2 arguments apply?

I don't have confident answers about how to handle the issues I'll discuss, but I'll try, where possible, to at least point at some interesting constraints that our responses must respect.

Let's say that prior to considering the simulation argument in its entirety (but after reflecting on cosmology, the current landscape of existential risks, the possible motivations for running simulations of early-seeming people, and the existing empirical evidence about the computational power available to technologically mature civilizations), you had the following pattern of credences:[165]

Starting distribution

a) *i*NS: ~100%

b) *i*S: ~0%

c) *Comp*: 99%

d) "The stars I see are real": ~100%

e) "Most early-seeming people are sims" (i.e., f_{sim} = ~1): 20%

[165] Here I am attempting to set aside more empirical questions about the credences that a standard scientifically-informed worldview would place on e.g. Bostrom's 1, 2, and f_{sim} = ~1, and to focus on the more philosophical questions about how to *adjust* those credences in light of Type 2 arguments. That said, the distinction here is loose, and may not be sustainable.

Core constraint tells us that either (e) or (a) (or both) needs to shrink dramatically (and if (a) shrinks, then (b) must grow). And on a Type 1 argument, (c) would need to stay high as well – but we no longer need to include this constraint. Still, though: where should we end up?

I really don't know. Bostrom doesn't claim to either, but his implicit answer in the paper appears to be that you should keep (c) and (e) roughly fixed, and adjust everything else accordingly. Thus, you might get (with changes bolded):

Bostrom-like distribution

a) *i*NS: **~80%**

b) *i*S: **~20%**

c) *Comp*: 99%

d) "The stars I see are real": **~80%**[166]

e) "Most early-seeming people are sims" (i.e., f_{sim} = ~1): 20%

We can think of this sort of adjustment as attempting to preserve as much as possible of your previous picture of the objective world, while allowing that conditional on (e), your confidence in your location within that world (and in some aspects of your previous picture of the objective – for example, whether the books on your shelves appear on an early-history non-sim planet, or on a sim planet) becomes undermined – and with it, your confidence in claims like "the stars I see are real." And this is, indeed, a salient possible end-state: albeit, one that requires rejecting the idea that you should be least as confident

[166] For simplicity I'm here assuming that you should have high probability on the stars you see being fake, conditional on being a sim. But this, too, isn't obvious.

in (d) as in (c) (call this "*Stars ≥ Comp*" – it's one way of cashing out *Parity of evidence* above), and that conditional on being a sim, *Comp* is not likely (i.e., *Sim ignorance*).

It's not the only possible end-state, though. You could, for example, simply shrink your credence on (e) to roughly zero, thereby preserving *Sim ignorance*, *Stars ≥ Comp*, and your original conviction that you're not a sim. That is (again, with changes from the original distribution bolded):

Apparently-f_s-is-small distribution

a) *i*NS: ~100%

b) *i*S: ~0%

c) *Comp*: 99%

d) "The stars I see are real": ~100%

e) "Most early-seeming people are sims" (i.e., f_{sim} = ~1): **~0%**

The problem with this move, though, is that it smacks of a kind of dogmatic and anti-empirical attempt to avoid putting substantive credence on being a sim – akin to concluding, in *Indistinguishable sims*, that apparently the "create the sims" button is going to malfunction, or that the dice isn't going to come up "1"; otherwise, after all, you'd probably be a sim, and you're confident that you're not.

Alternatively, if you wanted to try to preserve *Sim ignorance*, but you're OK putting substantive credence on being a sim, you could use some portion of your credence (below I use 80%) on a Bostrom-like distribution, and put the rest on being a sim in a scenario where *Comp* is false. Thus, for example, the following distribution is compatible with giving less 50% credence to *Comp*, conditional on *i*S.

Trying-to-capture-Sim-ignorance distribution

a) *i*NS: **~64%**

b) *i*S: **~36%** (and for more than half of this credence, *Comp* is false)

c) *Comp*: **~80%**

d) "The stars I see are real": **~64%**

e) "Most early-seeming people are sims" (i.e., f_{sim} = ~1): **~16%**

These are just a few examples of possible moves we could make, here – and they can be combined.

XI. Are we almost certainly sims?

In thinking about moves like this, I want to flag a certain sort of aspiration that we need to be very careful with, in making these adjustments: that is, the aspiration to preserve our credence on f_{sim} = ~1 *conditional on being non-sims* (in the original distribution above, this was 20%). To see this, let's look briefly at a certain way of extending the simulation argument, offered by Thomas (2021) – one that threatens to leave us almost *certain* that we're sims.

Thomas's argument runs as follows. Call the ratio of sims to non-sims in the observer class R, and let's assume that the observer class satisfies *Membership*, *Observer-class indifference*, and *Admissibility*. Type 2 arguments show that, conditional on R being very high, your credence in being a sim should be very high, too – but as just discussed, it doesn't tell you what credence you should have on R being high. But now suppose you condition on being a non-sim. What's the *expected* ratio of sims to non-sims in the observer class? If there's even a small (but non-trivial) possibility that the ratio is enormous, then the expected ratio will be very large as well. But this, in conjunction with the rest of the simulation argument's logic, entails extreme confidence, overall, that you're a sim.

To see why, consider a toy version of the basic dynamic, involving only the following four hypotheses:

i. You're a non-sim and R = 0 (there are no sims in the observer class).

ii. You're a non-sim and R = 1,000,000,000.

iii. You're a sim and R = 1,000,000,000.

iv. You're a sim in some other situation.[167]

Suppose that conditional on being a non-sim, (i) and (ii) are the only possibilities, and you think that (i) is 99x more likely than (ii). But if *Membership*, *Observer class indifference*, and *Admissibility* hold, then (iii) needs to be 1,000,000,000x more likely than (ii) (because conditional on R = 1,000,000,000, your odds on being a sim need to be 1,000,000,000:1). So (iii) needs to be 10,000,000x more likely than (i) and (ii) combined. So whatever your credence on (iv), your credence on being a non-sim has to be less than 1 in 10,000,000.

More generally, whenever you give c credence to being a non-sim in a world where the ratio of sims to non-sims is R, you are committing yourself to giving R*c credence to being a sim in such a world (this is why, according to *Core constraint*, c has to be less than 1/R overall). So if c is much more than a 1/R fraction of your total credence on iNS, then iNS gets swamped by iS.[168]

The key premise in this argument is the claim that conditional on being a non-sim, the expected ratio of sims to non-sims in the observer class is high. And this premise looks initially plausible. After all, we're used to assuming that we're non-sims, and the empirical

[167] In an unpublished version of the paper, Thomas presents an example like this.

[168] Note that there's no asymmetry between sims and non-sims, here. Thus, if conditional on being a sim, the expected ratio of non-sims to sims is very high, then you are similarly committed to ~certainty that you're a non-sim. This means that you can't have a high expected ratio of sims to non-sims, conditional on being non-sims, *and* a high expected ratio of non-sims to sims, conditional on being sims. Thanks to Hilary Greaves for raising questions about this.

facts Bostrom points to (i.e., *Comp*) make it hard to rule out futures where our descendants create very large numbers of sims in various plausibly admissible observer classes – even if we're substantially more skeptical of Bostrom's empirical case than Bostrom is.

But when you step back, Thomas's conclusion seems quite strange. Consider, for example, the following case:

> *Sims seem extremely unlikely*: You live on 21st century earth. Modern science says that the universe is almost certainly finite and entirely devoid of life, except for humans. What's more, it appears to you that simulations are prohibitively computationally expensive to run, even for very advanced civilizations. The brain appears to work via quantum microtubules that each have their own libertarian free will. Also, your stable global government has made a binding commitment to never run any sims, ever, and the universal human consensus, professed by all babies as soon as they can think, is that running sims would be a moral and epistemic horror. Also, there is a giant asteroid heading towards earth which will very likely kill everyone. Still, the super-duper forecasters – all of whom take for granted that we are not sims – give a one-in-a-billion probability to the hypothesis that all this anti-sim evidence is misleading, and that humanity will one day reach technological maturity and run a billion-billion ancestor simulations of this time in history.

In such a case, are we supposed to conclude that actually, despite all appearances to the contrary, we are overwhelmingly likely to be sims? This is the conclusion that would fall out of Thomas's argument, in combination with the credences that the super-duper forecasters give conditional on being non-sims. But it seems a strange lesson. Is it really so

hard to keep some credence on being a non-sim, in a world where it appears overwhelmingly likely that no sims will ever be created?

I think that this is indeed the conclusion you should reach, *if* you accept the credences that the super-duper forecasters give conditional on being non-sims. But you need not accept such credences. And indeed, if you don't, you can explain why: namely, that the super-duper forecasters *aren't taking simulation arguments into account.*[169] After all, as the previous section made clear, such arguments generally require that we *revise* something about how we would have otherwise apportioned our credences – and a very salient revision (indeed, one more made more salient by the implication that Thomas's argument highlights) is to *drastically* reduce our credence in a high ratio of sims to non-sims, conditional on being non-sims.

Thus, suppose that in the context of (i)-(iv) above, and prior to considering simulation arguments, you would have assigned ~99% credence to (i), ~1% to (ii), and negligible credence to being a sim. Simulation arguments require that your credence on (ii) shrink dramatically – in particular, to something less than one in a billion. But nothing requires that in making this revision, the *ratio* of your credence on (i) and your credence on (ii) must stay even roughly constant. If it does, then Thomas is right that being a non-sim goes out the window. But as discussed in the previous section, there are other options available – for example, you could take only the portion of your credence that was previously on

[169] Does this mean that they have incoherent credences? Not necessarily. It could be that they have very strange priors – and in particular, priors that radically privilege worlds where they are not sims, even conditional on R being high. Either way, though, this failure does indeed compromise their "super-duper"-ness in some sense – albeit, in a manner that would plausibly apply to many real-world super-forecasters as well (assuming that real-world super-forecasters, too, would put substantively higher probability on R being high than on being sims themselves).

(ii), and give almost all of it to (iii) instead (this is the move made by the Bostrom-like distribution above).

It's true that this route requires disagreeing with the credence the super-duper forecasters place on $R = 10^{18}$, conditional on being a non-sim. But if you accept the simulation argument's logic (as Thomas's argument does), you already knew that you were going to end up disagreeing with the super-duper forecasters *somehow* – since according to such logic, such forecasters *also* place a much-too-high *unconditional* probability of being non-sims in an $R = 10^{18}$ world (namely, ~one in a billion, where ~one in a billion-*billion* is the maximum permitted). And if you're going to disagree with their unconditional probability no matter what, it's not clear why preserving their conditional probability would be a priority.

That said, I do think Thomas's argument points to another option for an end-state probability distribution: that is, if we really want to preserve our original credences *conditional on being non-sims*, then we can do so. For example, we can say:

Preserve your credences conditional on being a non-sim

a) *i*NS: **~0% (but where Pr(f_{sim} = ~1 | *i*NS and *i*E) = 20%)**

b) *i*S: **~100%**

c) *Comp*: **~100%**

d) "The stars I see are real": **~0%**

e) "Most early-seeming people are sims" (i.e., f_{sim} = ~1): **~100%**

Or to put it another way, Thomas's argument points at a very substantive additional constraint on the overall credences you need to end up with after considering simulation arguments: for all observer classes that satisfy *Membership*, *Observer class independence*, and

Admissibility, *either* the expected ratio of sims to non-sims in those classes needs to be low conditional on being a non-sim, *or* you have to be basically certain that you're a sim. And naively, a high expected R conditional on being a non-sim seems very reasonable; as does placing at least some substantive credence on being a non-sim. But so does not being basically certain that you're a sim. So something has to give.

In general, it's not at all clear to me how we should adjust our credences, overall, in light of Type 2 arguments – but it seems to me a fruitful topic of further investigation.

XII. To what range of sims and observer classes does the argument apply?

Let's turn to another thorny issue.

I said at the beginning of the paper that the focus on early-seeming people as an observer class is arbitrary: the argument applies to *any* observer class that satisfies *Membership*, *Observer-class Indifference*, and *Admissibility*. That is, for *all* observer classes of this type, you shouldn't have non-trivial credence on the conjunction of $f_{sim} = \sim 1$ and *i*NS.

When we begin to explore what the implications of this are, though, we end up in territory stranger than Bostrom's original argument engages with. Suppose, for example, that we do not require that members of the observer class be human. Rather, let's use the observer-class "early-seeming creatures" – that is, any creature such that it seems to that creature that it lives in the early-history of its civilization, prior to technological maturity. And now consider the following case:

> *Squid sims*: The situation is just like *Imperfect Ancestor Sims*, except that according to the announcements you hear, the government authorities do not decide to run any ancestor simulations. Indeed, they decide that they will never, ever simulate any humans. Rather, they decide to run a billion simulations of *squid people* with tentacle

> arms, living in squid civilizations that haven't yet reached technological maturity – and not to run any other sims. What's more, while the authorities expect the simulated squids to plan on running simulations of some yet-different animal civilization (and to have global governments that make announcements to this effect), they are going to turn off the squid sims before such sims make it to technological maturity.[170]

Are you rationally permitted, in this sort of case, to believe the government's entire story – that is, that you are an early-seeming non-sim human, with a billion early-seeming sim squid-civilizations in your future? For the observer class "early-seeming creatures," this story would entail the conjunction of f_{sim} = ~1 and *i*NS, so in order to believe the government's whole story, we'd need to say that this observer class fails one of *Membership*, *Observer-class indifference*, or *Admissibility*. And since *Membership* and *Observer-class indifference* both look good, the question is whether *Admissibility* holds.

Now, in my experience, many people's intuitive reaction in this case is that *Admissibility* does not hold, and that it's fine to believe the government's entire story.[171] And perhaps, ultimately, there will be a way of justifying this intuition. Personally, though, I'm skeptical: my own best guess is that a Type 2 argument blocks believing the government's whole story here, as well.

To see this, it's important to understand that the question *isn't whether you're a squid sim*. Obviously, you're not a squid sim, because you're not a squid – your experiences of

---

[170] This case, along with the general methodology of comparing cases like *Squid Sims* with cases like *Imperfect Ancestor Sims*, is inspired by Garfinkel's (unpublished) discussion, which has generally been very influential on my own thinking. His version of *Squid Sims* involves it seeming to us like humanity is solidly on track to run a vast number of simulations exclusively of the actor Charlie Chaplin.

[171] And indeed, one suspects that if Bostrom had focused on a case like this off the bat, the argument would've gotten less traction.

human hands, rather than tentacle arms, make this clear. The question, rather, is whether, conditional only being an early-seeming creature and on living in a world where almost all of the early-seeming creatures are sims, the rest of your evidence is much more likely conditional on being a non-sim vs. being a sim (that is, whether *No update* holds). And in many worlds of this type, it's not the case that all the non-sims are humans, and all the sims are squids. For example, in some worlds, non-sims of some other species – for example, *lions* – create lots of sim *humans*, who then *think* that they'll go on to create lots of sim squids (and whose governments make announcements to this effect), even though actually, they won't.

Here's an intuition pump that might be helpful. Consider the set of worlds that fit the following schema, for some set of three animal types A, B, and C (e.g. lions, humans, and squids). On planet 0, there is one set of *n* early-seeming non-sim A-animals, whose government says "all the early-seeming A-animals are non-sims (so you're non-sims), but there are a billion sets of *n* early-seeming sim B-animals in the future"; and then, in the future of planet 0, and on each of planets 1-1,000,000,000, there are *n* early-seeming sim B-animals, all of whose governments say "all the early-seeming B-animals are non-sims (so you're non-sims), but there are a billion sets of *n* early-seeming sim C-animals in the future" (when actually, no animals of type C exist). Call any world that fits this description – for some values of A, B, and C – a V-world.

Conditional on living in a V-world and on having the experiences described in *Squid sims* – e.g., looking down at human hands, listening to a government announcement about future squid sims – is it any more likely that you live on Planet 0 vs. Planet 1? I don't see why it would be. Humans, after all, are not more likely, on priors, to play the role of a type A animal, in a V-world, vs. a type B animal; and the same holds for squids with the respect to

type B and type C. That is, "humans simulate squids, who wrongly think they'll simulate some other type of animal" is no more likely, on priors, than "some type of animal simulates humans, who wrongly think that they'll simulate squids." And conditional on it being humans on any given planet, your experiences in particular seem equally likely. So to me it seems plausible that your credence in living on Planet 0 should be equal – or at least, roughly equal – to your credence in living on Planet 1, and same for Planet 2, 3, and so forth. Thus, conditional on living in a V-world and having your experiences, the usual Type 2 logic applies, and if we accept it in the previous cases (e.g., Z worlds above), then it seems like you can't have gotten strong evidence that you live on Planet 0 in particular. But the world you hear your government describing, in *Squid sims*, is effectively a V-world where you live on Planet 0. So you can't have strong evidence for the government's whole story.

Indeed, the sense in which "all the sims are squids" is not good evidence that you're a non-sim, in this case, seems to me closely analogous to the sense in which claims like "none of the sims will see my particular books" or "none of the sims will see my particular light speckles" aren't good evidence that you're a non-sim in cases like *Imperfect ancestor sims* and *Sims with different light speckles*. That is, in all of these cases, you were initially tempted to posit an objective description of the world that rules out your being a sim, given your experiences (i.e., where your books/light speckles are only seen by a non-sim). And the key point isn't that this description is true, but that you can't tell whether you're a sim in that sort of world. Rather, it's that you can't have strong evidence for that description being true, given reasonable priors; rather, if you're giving credence to that description, you need to be giving credence to other alternative descriptions as well, in which your experiences are had by sims instead.

In fact, *Sims with different light speckles*, *Imperfect ancestor sims*, and *Squid sims* seem to me sufficiently analogous, structurally, that my best guess is that they stand or fall together – that is, that Type 2 simulation arguments either work in all these cases, or in none of them. My best guess is: all of them – and I think that if you say "none of them," then you should probably start resting easy with the belief that you're a non-sim in *Indistinguishable sims* as well (though this seems to me quite strange), given its similarity to *Sims with different light speckles*. That said, I haven't tried to exhaust all of the possible disanalogies here, and it's possible that there's some other line to be drawn (ideally, in my opinion, between *Imperfect ancestor sims* and *Squid sims*, as I do think that applying Type 2 arguments to *Squid sims* is somewhat counterintuitive, and it would be nice to explain why).

In the meantime, though, we might wonder: once we're applying Type 2 arguments to *Squid sims*, how far, exactly, will they go? Consider, for example:

> *Sims with little tags*: A case like *Imperfect ancestor sims*, but the stable global government also announces that it's going to put a little tag in the visual field of all the future sims, which says "you are a sim" – but where such sims have 21st-century-like experiences beyond this. You have no tag in your visual field.

Are you allowed to put substantial credence on the government's announcements being true – including the bit where you're a non-sim? That depends on whether *Admissibility* holds for any observer classes that you and these sims-with-tags are both a part of – for example, the observer class "people who have 21st-century-like experiences, whether or not they have little tags." This observer class, too, plausibly satisfies *Observer class indifference* – so the main question is whether, once you condition on being in this observer class and on $f_{sim} = \sim 1$, the rest of your evidence – e.g., listening to your global government announce a plan to simulate lots of sims with little tags, not having a little tag yourself – is

any more or less likely, conditional on being a sim vs. a non-sim. But just as, in *Squid sims*, you can't rest easy with "all the sims are squids" (since the experience of the government announcing that all the sims are squids seems similarly likely conditional on being a sim vs. a non-sim, given *i*O and $f_{sim} = \sim 1$), neither can you rest easy with "all the sims have little tags." Rather, the question is whether government announcements to the effect that all the sims have little tags, combined with the absence of little tags in your experience, are a strong sign you're a non-sim, given *only* that you have 21st-century-experiences, and that almost everyone with such experiences is a sim. And this seems to me quite unclear.

And this same unclarity applies in a whole panoply of more exotic cases – for example, ones in which the government announces that e.g. it will only run sims of people on cooking shows; or of people in extremely violent and entertaining video games; or of people for whom running sims seems impossible; or of universes governed by physical laws dramatically different from our own; or where the government announces that humans will never run sims, but the lizard people the next planet over are getting ready to run lots of ancestor sims.[172] Importantly: in all of these cases, we need to take care not to conflate the frequency of our own experiences among sims vs. non-sims, *in the world that the government posits*, with the likelihood, *on priors*, of our having those experiences given that we're sims vs. non-sims, conditional only on *i*O and $f_{sim} = \sim 1$. After all, our priors, here, cannot be set by the frequencies *within the world* – because which world obtains is precisely the question. And this makes *Admissibility* correspondingly hard to reason about.

What's more, though, *Admissibility* isn't actually required. Rather, the argument continues to work, in a roughly similar way, even if your particular experiences are a strong update

[172] This last one requires that we relax the "we're alone in the universe" aspect of the government's announcements.

towards being a non-sim, after you condition on *i*O and f_{sim} = ~1. Thus, for example, suppose you are convinced that experiences as boring and mediocre as yours are massively more likely to be had by a non-sim with 21st century experiences than a sim with 21st century experiences, conditional on *i*O and f_{sim} = ~1 (perhaps because you think entertainment value by far the most likely explanation of someone running sims). Still: *how much* more likely? If, for example, *i*O and f_{sim} = ~1 leaves you at a billion to one on *i*S, then even if your boring experiences are a million to one update towards *i*NS, you'll still be at a thousand to one on *i*S at the end of the day. So if the ratio of sims to non-sims is large enough, *Admissibility* can fail very badly, and you'll still lose your ability to place much credence on the conjunction of *i*NS and f_{sim} = ~1.[173]

So overall, I find it hard to think about what classes of sims the simulation argument's restrictions cover. Indeed, at present, I tend to view with suspicion any hypothesis that combines *i*NS, *i*O and f_{sim} = ~1 for some observer class I'm a member of – regardless of questions about *Admissibility*. So the idea that I'm a non-sim, but that sims of *any kind* substantially outnumber the early-history people, takes a serious hit.

XIII. Wrapping up

In the last few sections, I've discussed a number of complications and uncertainties that arise if we start to take Type 2 arguments seriously. In particular: we need to find an overall credence assignment that respects *Core constraint* for all observer classes that satisfies *Membership*, *Observer-class indifference*, and *Admissibility* (a set that may be quite large,

[173] See Thomas (unpublished), p. 11, who makes this point in the context of his own formulation of the argument, discussed above. Note, though, that *Admissibility* needs to fail in extreme ways in everyday cases in order for us to have justified confidence in pretty basic beliefs. Thus, for example, conditional only on being one of 7.7 billion people on earth right now, and the fraction of those who don't live on my block being ~1, in order for my experiences to make me justifiably confident I live on that block, they need to be 7.7 *billion* times more likely conditional on living on that block vs. conditional on not doing so. But strong evidence like this is actually quite common (see Xu (2021)).

and which seems challenging to analyze), while also navigating the pressure created by principles like *Stars ≥ Comp* and *Sim ignorance*, and by arguments like Thomas's (2021) for high expected ratios of sims to non-sims, conditional on being a non-sim.

In closing, I'll also briefly note a few other issues. First: simulation arguments – including Type 2 versions -- work best in finite worlds. Indeed, Bostrom "deliberately sets aside" infinite worlds in his original paper. In a later FAQ, he suggests that handle them, we might appeal to the limiting fraction of sims vs. non-sims in expanding hyperspheres – but this sort of approach faces significant problems.[174] Granted, finding a way to talk sensibly about the fraction of observers with X experiences in infinite worlds is a problem for cosmologists more generally, but the existence of this problem adds an open question about how to best apply simulation-argument-style reasoning to infinite cosmologies.[175]

Second, I haven't been discussing the bearing that anthropic principles like the Self-Indication Assumption ("SIA") and the Self-Sampling Assumption ("SSA") might have on hypotheses involving simulations – but pretty clearly, there are implications. SIA updates towards worlds where there are more observers with evidence like yours; whereas SSA updates towards worlds where the observers with evidence like yours are a larger fraction of some reference class.[176] Notably, though, sim-filled worlds plausibly involve more

---

[174] See Dorr and Arnztenius (2017) for discussion.

[175] A related issue, here, is what our epistemic relationship should be to the idea that we are "freak observers": i.e., observers generated by random fluctuations in a sufficiently big world, which make all possible sets of observations some very large number of times. Like sims, freak observers fall out of various seemingly-plausible empirical claims, but positing them can quickly lead to sharing the world with some very large number of observers making observations similar to your own, except in a strange skeptical setting – thereby prompting the concern that either such empirical claims are false, or you're overwhelmingly likely to be in such a skeptical setting. I won't try to delve into this topic here, but I do want note that while it's possible to point to various differences between the arguments (notably, for example, the vast majority of freak observers will disintegrate in the next moment – but this isn't true of sims), to me it seems likely that some aspects of the reasoning involved will stand or fall together (see Crawford (2013) for more on this). And to the extent we find it harder to take seriously the hypothesis that we are freak observers vs. the hypothesis that we are sims (I do), this suggests either that there is some lurking confusion underlying *both* arguments, or that additional revisions to our naïve picture of our situation are in order.

[176] See Bostrom (2002a) for an introduction.

observers in epistemic situations like yours – thereby prompting SIA to update towards them (and especially: towards worlds obsessed with simulating you in particular). But they also involve lots of civilizations reaching technological maturity; and if the average non-sim population in a technologically mature civilization outnumbers the average sim population (plausibly, after all, most of the resources go to the actual *citizens* of the civilization itself, rather than to running sims), and most of the non-sim population has technological-maturity-indicating experiences quite unlike our own, then on SSA, finding yourself without technological-maturity-indicating experiences is a large update *against* worlds like this – since conditional on living in such a world, you should've expected to be a non-sim member of technologically mature civilization, rather than either a sim, or a non-sim very early on in history.[177] And I expect other implications of SIA and SSA for debates about sims as well, beyond these examples.

Finally, I haven't, here, tried to tackle the practical implications that fall out of the discussion – and in particular, that would fall out of starting to give serious credence to being a sim. Some writers on this topic have explored the possibility that, for example, research into whether we live in a simulation itself risks causing the simulators to turn us off;[178] that the simulation argument should make you act more selfishly;[179] and that it should make you act on shorter time horizons and with less concern for the long-term future.[180] Topics like these seem to me well worth further investigation.

All in all, then, there's still a lot to work out. Still, as far as I can tell, the basic thrust of the simulation argument has real philosophical force and interest – especially when interpreted

[177] Or, let's assume, a non-sim in some weirder situation – i.e., a fake, terraformed non-sim world created by an advanced civilization. And of course, as with all conclusions drawn from SSA, this one depends on the reference class.
[178] See Greene (2020).
[179] See Hanson (2001).
[180] See Tomasik (2016).

in the Type 2 manner I've argued for here (that is, as not resting on the likelihood of any particular set of empirical claims). Perhaps it does not, ultimately, work – but I don't think its failures are at all obvious, and I expect that teasing them out, if they exist, will itself be an instructive exercise. After all, serious arguments for such dramatic re-orientations in our basic understanding of our existential predicament do not come along every day. We should pay attention when they do.

And whether we buy simulation arguments or not, they are a reminder that the world we see and take for granted is only a part of the world; and that in principle, our overall existential situation could in fact be many different ways, not all of which we are accustomed to considering. Ultimately, we need priors – and indeed, capacious ones, adequate to include worlds that are bigger and stranger than what we take to be normal (is it so normal, when you step back and look?). Dealing well with such worlds is a delicate art. But it's one that simulation arguments, whether sound or not, remind us to learn.[181]

[181] Thanks to Katja Grace for extensive discussion of the issues in this essay (and for written comments on a later version); to Ben Garfinkel, whose work on this topic has been especially influential on my own thinking; to Hilary Greaves, for written comments on multiple versions; and to Teruji Thomas, for discussion of his own work and for comments and discussion on part of an earlier draft as part of the confirmation of status process. And thanks to Paul Christiano, Owen Cotton-Barratt, Cate Hall, Ketan Ramakrishnan, and Carl Shulman for discussion as well.

Chapter 3

Infinite ethics and the utilitarian dream

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.[182] 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

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

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, spatio-temporal 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 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.[183] 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

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

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.[184] 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 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.[185] So exerting infinite causal influence probably needs new science. Maybe we learn to make computers that perform infinite amounts of computation,[186] or

[184] 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).

[185] 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.

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

baby universes with infinite space-times.[187] 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.[188] 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.[189] 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

[187] See Guth (1996).
[188] 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.
[189] 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).

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 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.[190]

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 data-point – 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.

---

[190] 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.

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 on 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:

> *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.[191] 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

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

plausibly, average utilitarians are going to have similar troubles.[192] Bostrom calls this "infinitarian paralysis," and he treats it as a devastating problem.[193]

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.[194] 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.[195]

Bostrom considers a response of this type, but he argues that it faces two problems.[196] 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.[197] But this doesn't worry me much: I'm fine with merely aspiring to help people (rather than to

---

[192] Though we'd need to say more about how we're doing the averaging; see the section on average views below for more.
[193] 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).
[194] I haven't tried to work through the details here, partly because I see average utilitarianism as independently implausible.
[195] See MacAskill (2015): "It's not the size of the bucket that matters, but the size of the drop" (p. 25).
[196] See section 3.2 on "the causal approach."
[197] 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).

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.[198] 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.[199] 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 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

[198] 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.

[199] 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).

Lrisky: value V with probability e; value 0 otherwise

Lsafe: value v with probability 1" (p. 5).

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.[200] 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.[201] After all, *any* non-zero credence on an infinite payoff is enough to bite

[200] 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.
[201] See Bostrom (2011), p. 32.

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?).[202] Thus, no matter where she is, no matter what she has seen, the infinite fanatic never gives finite things any intrinsic attention.[203] When she kisses her children, or prevents a genocide, she does it for some infinite prize, however improbable.[204]

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.[205] 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.*

---

[202] 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.

[203] 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 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.

[204] 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.

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

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.[206]

What's more, though, one salient response to finite fanaticism is to just bite the bullet, at least in sufficiently thought-experimental cases;[207] 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 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

[206] See Beckstead and Thomas (2021) for discussion of these options.
[207] See Wilkinson (2021c) for advocacy.

(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.[208]

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

---

[208] 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.

> *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.[209] 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.)

To see the conflict between *Infinite Agent-Based Pareto* and *Agent-Based Anonymity*, consider the following example.[210] 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	a_1	a_2	a_3	a_4	a_5	a_6	a_7
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.[211] So by *Agent-Based Anonymity*, w1 and w2 are equally good. Contradiction.

Here's another example.[212] 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,

---

[209] 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."
[210] This example is adapted from Van Liedekerke (1995).
[211] 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.
[212] This example is adapted from Hamkins and Montero (1999).

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}	a_0	a_1	a_2	a_3 …
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*).

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

Agents	a_1	a_2	a_3	a_4	a_5	a_6	a_7
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}	b_0	b_1	b_2	b_3 …
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).[213] But letting go of *Agent-Based Anonymity* has strong intuitive costs.

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

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.[214] And note that we can get contradictions between *Something-Based Pareto* and *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).[215]

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

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

[215] 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.

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 <1, 1, 1, 0, 1, 1, 1, 0 …>

60% on zone of suffering, plus an immortal lizard (always at 1) on the side.

l2: 80% on <1, -2, 3, -4, 5 … >

20% on zone of happiness, plus four immortal 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.

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 -∞. So which is it? If you're God, and you can create this world, should you?[216]

Or consider a world where the welfare levels are: 1, -1/2, 1/3, -1/4, 1/5, and so on.[217] Depending on the order you use, these can sum to *any welfare level you want*.[218] 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

---

[216] 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.

[217] 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.

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

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. 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 $<l_1, l_2, l_3 \ldots l_i \ldots>$, and for a discount rate α between 0 and 1 (non-inclusive), could we say that the value of *w* is $\sum_{i=1}^{\infty} \alpha^{i-1} l_i$?[219] This would allow us to say that a world of <1, 1, 1, 1…> is better than a world of <2, 2, 2, 2…>, 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., $1\times.5^0 + 1\times.5^1 + 1\times.5^2 \ldots = 1 + .5 + .25\ldots$), and the value of the 2s world is 4 (i.e., $2\times.5^0 + 2\times.5^1 + 2\times.5^2 \ldots = 2 + 1 + .5\ldots$).

Approaches like this, applied to locations increasingly distant in time from the decision-maker, are common in economics.[220] 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

[219] This is a slightly simplified version of the discounted utilitarian rule discussed by Lauwers (2014), p. 6.
[220] See Cowen and Parfit (1992) for discussion.

value that are in some sense "simpler" to describe, for some definition of "simpler," should be given ethical priority.[221]

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 extreme costs to do this.[222]

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, <0, 1, 0, 0…>, total discounted value .5 (0 + .5 + 0 + 0…) – 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 <1, 0, 0, -1, -1, -1…> and a total discounted value of .75 (1 + 0 + 0 – .125 – .0625 …).[223] But causing infinite suffering for the sake of such a re-arrangement seems (at least to me)

[221] 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.

[222] 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.

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

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.[224] And even if we set such reasons aside,[225] 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.[226] 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…); and so, too, is it infinite in worlds where the utility at each location is a million times larger (1M + 1M + 1M…). 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 <2, 2, 2, 2, …> is better than <1, 1, 1, 1, …>, one option for capturing this would be to assess the value of an infinite world via the limit $\lim_{n\to\infty} \frac{\text{total welfare of the locations counted so far}}{\text{number of locations counted so far}}$, 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:

---

[224] 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.

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

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

- 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.[227]

- Even if we fix an ordering, it's indifferent to many ways of helping infinitely many locations, like moving from <1, 2, 3, 4..> to <2, 3, 4, 5,…> (limiting average: ∞ in both cases).

- It becomes undefined whenever you end up adding locations in an order like <1, -3, 5, -7…>, 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. <∞, -∞, -∞, -∞, -∞, -∞ … > vs. <-∞, ∞, ∞, ∞, ∞, ∞…>).

- 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).[228]

---

[227] 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.
[228] See Parfit (1984) for canonical discussion.

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*.[229] I won't examine this proposal in detail here, but here's a brief description.[230] Hyperreal numbers are extensions 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 <3, 3, 3, …>); 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, <1, 1, 1…> would be worse than <2, 2, 2,…>, 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

[229] 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.
[230] See Arntzenius (2014), section 5, for an especially clear introduction.

hyperreal corresponding to the sum of the utility as you proceed along the ordering. Thus, the world <1, 1, 1…> would correspond to the hyperreal <1, 2, 3 …>; and if you bumped up the first person to 2 (such that the world is now <2, 1, 1…>), that would change its corresponding hyperreal to <2, 3, 4, …> -- 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.[231] Thus, suppose that the locations are times, and that the seconds in the world have utilities <1, 1, 1 …>, then your world gets better depending on whether you use one-second time intervals (hyperreal: <1, 2, 3…>), two-second time intervals (hyperreal: <2, 4, 6 …>), three-second time intervals (hyperreal: <3, 6, 9…>), and so on.[232]

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:

- The world <1, -2, 1, 1, -2, 1, 1 …> can be made better, worse, or equal to an empty world (since its corresponding hyperreal, <1, -1, 0, 1, -1, 0…>, can be made

[231] See Arntzenius (2014), p. 49.
[232] 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.

to equal <1, 1, 1 …>, <-1, -1, -1…>, or <0, 0, 0…> 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 <2, 2, -2, 2, -2 …> is either twice or four times as good as a single person at 1 (its corresponding hyperreal is <2, 4, 2, 4, 2 …>, 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).[233]

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.[234]

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

[233] See Bostrom (2011),p. 23, and Arntzenius (2014), p. 50-1, for more on these objections.
[234] 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.

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 <1, 1, 1, 1, …> vs. <2, 2, 2, 2, …>, the utility inside any expansion is bigger in the 2 world. And similarly, in <1, 2, 3, 4 …> vs. <2, 3, 4, 5>, 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*).[235] Pinning down "uniform expansion" also 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.[236]

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.

---

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

[236] 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).

Thus, suppose you're choosing between the following lotteries, all for with the same locations of value:

l3: 50% on <1, 1, 1, 1…>

50% on <1, 2, 3, 4…>

l4: 50% on <-1, 0, -1, 0…>

50% on <1, 4, 9, 16…>

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 <1, 1.5, 2, 2.5 …>, and l4 is equivalent to <0, 2, 4, 8 …>. The latter is better according to Vallentyne-Kagan, so Arntzenius says to choose it.[237]

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 Dystopias to the world.[238] 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

[237] See Wilkinson (2021) for similar themes.
[238] Thanks to Amanda Askell, Hayden Wilkinson, and Ketan Ramakrishnan for discussion.

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.

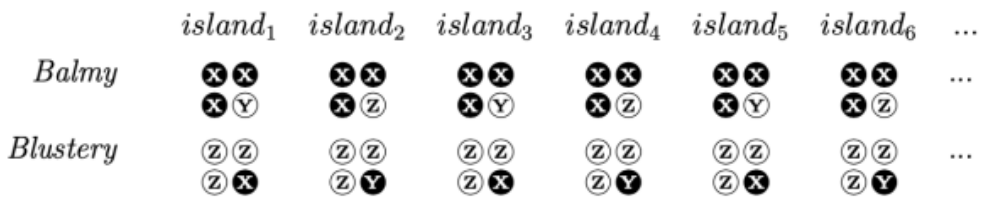

Figure 31: Balmy and Blustery

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.[239]

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 <…-2, -1, 0, 1, 2 …>, 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 <…-2, -1, 0, 1, 2 …>. 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.

[239] 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 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).

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.[240]

Will this problem arise in practice? Arntzenius (2014) and Wilkinson (2021) seem to think not.[241] But I disagree.[242] 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.

---

[240] 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.

[241] 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."

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

This seems like a general problem for any approach infinite ethics that requires identity or counterpart relations between spatio-temporal locations across worlds.[243]

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.[244] 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. *Value: ∞.*
2. *Negative infinities*. Worlds with infinite pain, and finite pleasure. *Value: -∞.*
3. *Mixed infinities*. Worlds with infinite pleasure *and* infinite pain. *Value: 0* (the good and bad infinities cancel).[245]
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

[243] See e.g. Wilkinson (2022) and Easwaran (2021).
[244] 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.
[245] 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.

break the tie (though whether ties will come up very often in practice is a further question – I'm skeptical).[246]

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.
- *Hell*: Infinite people being tortured for eternity (with no pleasure whatsoever).[247]

On the four types view:

---

[246] 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).

[247] 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).

- *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.[248]

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

[248] Thanks to Amanda Askell for discussion of the sense in which utilitarians really care about the *amount* of utility.

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,[249] 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.[250]

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

[249] 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.

[250] 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.

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 counter-example 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.[251]

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

[251] 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.

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.[252] But notably, various prominent views in infinite ethics (notably, expansionist 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

[252] 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).

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.[253] 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.[254]

We can see the outlines of a broader pattern. A certain kind of reductionist impulse in finite ethics often tries to ignore structure.[255] 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.[256]

---

253 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).

254 See Askell (2018), p. 198, for more on this.

255 See e.g. Chappell (2011) on "value atomism."

256 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.[257] 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*."

[257] 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.

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.[258] 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 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.[259] 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

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

[259] 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."

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;[260] some population ethicists have seen its difficulties, even in finite cases, as arguments for meta-ethical anti-realism;[261] and some deontologists see the difficulty of generating plausible systematizations of our deontological intuitions in a similar light.[262] 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" and some doubt about whether the project of rationalist ethics is in good order – a connection that seems worth understanding.[263]

XVIII. Infinities in practice

[260] 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.

[261] 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).

[262] See Ramakrishnan (unpublished).

[263] 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.

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.[264] After all, if

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

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-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.[265] 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.[266]

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

[265] 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.
[266] See e.g. Ord's (2020, Chapter 8) discussion of the "long reflection."

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.[267] 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).[268] So relative to a longtermism based on simple totalism, 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).[269]

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

[267] See e.g. Bostrom (2003b) for a classic example.
[268] Thanks to Nick Beckstead for suggesting this point.
[269] We can speculate about the value of resources growing superlinearly, but note that this applies in the finite case, too.

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, over-confidence, 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.[270]

[270] 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.

Conclusion

This thesis has examined three main issues: SIA vs. SSA, simulation arguments, and infinite ethics. I've argued that SIA is better than SSA (though it still has problems); that simulation arguments can be formulated in a forceful way (albeit, one that leaves many questions unanswered); and that infinities puncture the dream of a simple, bullet-biting utilitarianism (and that they cause many other problems in normative ethics, and possibly meta-ethics, as well).

All three of these issues have the potential to substantially complicate the empirical and normative case for longtermism, especially in combination. SIA and simulation arguments both suggest that our future may be smaller than we would naively think (because most life that reaches our stage of development fails to reach technological maturity, or because we live in a simulation), and that our universe and our causal influence may be bigger – indeed, infinite (SIA is certain that we live in an infinite universe; and living in a simulation makes it more plausible that the physics of unsimulated reality allows for infinite causal influence). In infinite universes, though, SIA starts to break down, as does the sort of indifference-based reasoning underlying simulation arguments; and the sort of aggregative utilitarian reasoning that most naturally undergirds longtermism starts to break down as well – or at least, it starts having to make hard choices about which seemingly-core commitments to give up. What's more, that same aggregative utilitarian reasoning plausibly suggests that infinite impacts swamp the sorts of astronomical but finite impacts that the standard case for longtermism focuses on. In this sense, just as SIA becomes certain that we live in an infinite world, but then doesn't know how to reason about which one, aggregative utilitarianism becomes obsessed with infinite impacts, but then doesn't know

how to prioritize between them. In both cases, infinite scale hijacks our attention, but then leaves us confused.

Of course, there's much more to say about all of these issues; and we may, ultimately, be able to salvage some or all of longtermism's logic and/or practical upshot. A full litigation of the implications of these issues for longtermism, though, is outside the scope of this thesis. My aim, here, has been more modest: to clarify some of the underlying philosophical dynamics at stake in these issues, and thereby to lay the groundwork for further and more thorough investigation of where those dynamics lead.

As I noted in the introduction, I find these dynamics interesting partly because they seem to me notably continuous with the sort of philosophical aesthetic that often motivates longtermism itself, even as they plausibly lead to stranger and more confusing places. It's an aesthetic willing to accept that big numbers can lead to big changes in belief and action; that philosophical reasoning, more broadly, can justify serious departures from our everyday moral and empirical common sense; that Cotra's "train to crazy town" is worth riding. Many of the most common objections to longtermism are, implicitly or explicitly, objections to this sort of aesthetic as well – and for this reason, they can sometimes leave longtermists in the grip of this aesthetic uncompelled. Whether these objections are sound or not is a further question; but regardless, it seems to me especially worthwhile to examine objections that attack longtermism from the other direction – that is, objections that accuse longtermists not of taking a certain kind of quantitatively-oriented philosophy too far, but of not taking it far enough.

The topics I have examined aren't the only topics with the potential to do this. To take just one other example: to me it seems plausible that once we start giving substantive credence to non-causal decision theories like evidential decision theory – on which,

roughly, you should choose the action such that your performing that action would be the best *news*, rather than the action that would have the best causal effects – the focus of an effective altruist's moral attention should shift away from the causal consequences of our actions for *our* long-term future, and towards the non-causal consequences of our actions elsewhere in sufficiently big universes/multiverses, since the expected number of moral patients affected by non-causal action across a big universe/multiverse dwarfs the number causally downstream from our particular location.[271] Indeed, depending on how you normalize across the value at stake on evidential vs. causal decision theory, and on how you treat decision-theoretic uncertainty more broadly, this conclusion seems plausibly robust even to low credences on big universes, and on non-causal decision theories, because so much more value is stake conditional on the ability to a-causally influence a large universe (see MacAskill et al (2021) for some discussion).

As above, I'm not going to try to litigate this issue here.[272] Rather, I mostly want to note that it, too, seems to me an objection to longtermism continuous with the sort of philosophical aesthetic I gestured at above (it accepts strange, philosophically motivated revisions to common sense; it turns our attention to larger-scale impacts even in the face of low probabilities; and so on). And I expect there to be more objections of this type as well. Indeed, to me it seems plausible that one of the ways philosophers can most effectively contribute to the intellectual project of effective altruism is by searching out new and game-changing extensions of this sort of philosophical aesthetic, and by grappling with some of the confusions that arise as we push it to its limits.[273] As I noted in the introduction, effective altruism has uniquely strong roots in analytic philosophy; its

[271] See Oesterheld (2017) for a lengthy investigation of some of the dynamics, here.
[272] Though see Carlsmith (2021a) for some discussion.
[273] See Bostrom (2014) on "Crucial Considerations" for more in this vein.

practitioners treat philosophical reasoning as a direct (and in some cases, decisive) input to their decision-making; and it is centrally in virtue of such reasoning that billions of dollars and thousands of careers are now being directed towards longtermist causes in particular. In effect, that is, effective altruism has already ridden the crazy train fairly far – and by riding it further, philosophers can break new, fertile, and possibly decision-relevant ground.

That said, the question of how, ultimately, to relate to the crazy train remains open. Indeed, I think that one of the key benefits of investigating where the crazy train leads -- and especially, where it breaks down -- is the cultivation of humility about the project it represents. It's one thing if one feels that the verdict of the most rigorous and quantitative philosophy (this is how the aesthetic I have in mind would style itself – accurately or no) is clear, albeit uncomfortable – in that case, it becomes tempting to say that one should simply accept the discomforts in question (in the language of the third chapter, to "bite all the bullets"), rather than to compromise on quantitative rigor. And such a response is made easier when the bullets at stake are confined to those put forward by arguments like Bostrom's "Astronomical Waste" – e.g., that very small reductions in existential risk can massively outweigh more familiar forms of altruism. As the bullets become even stranger, though (massive updates towards hidden realms filled to the brim with copies of this specific experience? any probability of an infinite impact outweighs a guarantee of any finite impact? high probabilities that the microstructure of the earth stops existing when no one is looking at it?), hesitations may increase. And once the tools and assumptions of "the most rigorous, quantitative philosophy" start to actively fail in your hands (as they often do, I've suggested, in infinite cases, at least when naively applied), or to lead into a morass of uncertainty, your confidence in them may be shaken still further. In particular, such a breakdown can cancel some backdrop sense that there is a privileged default one

need only be rigorous enough to accept. One is left, instead, with a sense that there is no clear road ahead – though there is, perhaps more temptingly now, a road back.

Relating gracefully and reasonably to this sort of intellectual and practical predicament seems to me a delicate dance – and I don't have any sort of universal prescription for dancing it well. One might think that at the least, one should avoid getting off the crazy train at some arbitrary stop: one should ride to the end, or one should never ride at all. On reflection, though, the line has neither a clear beginning (all of our everyday moral and empirical common sense is shot through with philosophical assumptions; and what seems common-sensical to you and your peer group would often seem revisionary to many others further away in time or space) nor, I have suggested, a clear end.

However, it's also not the case that we need to "get off" at a single discrete stop – rather, we can have differing mental and practical attitudes to different conclusions reached along the way. In particular, my own current view is that we should devote serious intellectual energy to pushing the crazy train as far it goes; but that we should be more cautious about actively incorporating the considerations it exposes into high-stakes decision-making, especially where doing so runs actively counter to more common-sensical values and worldviews, and/or where the considerations in question still have a flavor of brittle cleverness rather than direct and resonant plausibility. Indeed, in practice, many longtermists already do something like this – that is, they focus on and advocate for longtermist interventions (e.g., decreasing the risk of engineered pandemics, preventing war between great powers, making sure that a transition to a world with advanced AI goes well) that look at least *good* on more near-termist grounds too; and they don't emphasize a need for extreme sacrifices in the present for the sake of the future.[274] Sometimes this sort

[274] See MacAskill (2022) for a representative example of this sort of presentation of longtermism.

of moderation is instrumentally motivated, but I think it's also, often, an expression of longtermist considerations influencing action *conditional* on passing various sorts of intuitive tests for robustness to and resonance with a richer set of heuristics, values, and epistemic gut-checks. To the extent we start riding the crazy train past longtermism, a similar approach seems to me appropriate – and perhaps, all the more important.

Indeed, in this spirit, to me it seems plausible that riding the crazy train past longtermism will often, nevertheless, preserve one's interest in an intermediate goal of many longtermists: namely, for humanity to reach a state of much greater wisdom and empowerment, in which it will become possible to understand all of these issues at a much deeper level, and to act on their implications more effectively as well. After all, if we are, indeed, at a very early stage of understanding where these considerations point, the value of information seems high; and to the extent they point towards cosmically-oriented projects (creating baby-universes, acausal interaction with civilizations across the multiverse, etc), greater civilizational empowerment seems likely to be useful as well.

In this sense, we might think of our uncertainty about these domains as suggesting what we might call "wisdom longtermism," as opposed to "welfare longtermism" – that is, a form of longtermism especially focused on the instrumental value of the wisdom that we hope the future will provide, rather than on the welfare of future generations in particular. Standard presentations of longtermism incorporate both of these considerations to different degrees;[275] but I think that topics like the ones discussed in this thesis emphasize the "wisdom" component, and they caution us against surprise if this sort of wisdom turns

[275] See e.g. Ord (2020) and MacAskill (2022) on the "long reflection."

humanity's moral attention to something other than the welfare of the sentient creatures in our lightcone.

That said, while the approach to the crazy train I've suggested here – that is, curiosity about where it leads, caution about incorporating its verdicts too hastily into our decision-making, emphasis on reaching a future of greater wisdom and empowerment – seems like a plausible path forward to me, it is not risk free, and I think we should be wary of nodding sagely at words like "caution" and "wisdom" in a way that masks the real trade-offs at stake. Sometimes, for example, "caution" about accepting a strange-but-true conclusion leads to action coming far too late; sometimes the right thing to do is neither intuitively resonant nor robust to alternative worldviews; sometimes there is no time to seek greater wisdom. Indeed, if we live in a simulation, or if the vast majority of civilizations at our stage of development fail to reach technological maturity, the idea that we can punt our uncertainties to some future set of substantially wiser and more technologically-empowered descendants may lead us astray, since regardless of what we do, there may be no such descendants to take on the inquiry.

As ever, then, there is no total safety; no way to avoid making bets – and, sometimes, losing them. This is true of common sense, true of "rigor" and its associated aesthetics, true of any stop along the crazy train, and of any combination. To me, topics like the ones discussed in this thesis are especially useful in complicating the notion that one particular stop – namely, the sort of longtermism associated with Bostrom's astronomical waste argument, and related variants -- offers some privileged type of safety, associated with "just following logic where it leads," "biting the bullets," and so on. But these topics don't, in themselves, offer the comforts of an alternative path we can be confident in. Indeed, I

suspect that such comforts are too much to ask for, and that we should learn to live without them.

Work cited

Alexander, Scott (2022). "Book Review: What We Owe The Future." *Astral Codex Ten* https://astralcodexten.substack.com/p/book-review-what-we-owe-the-future

Armstrong, Stuart (2009). "Avoiding doomsday; a 'proof' of the self-indication assumption." *LessWrong* https://www.lesswrong.com/posts/5A9x74mgCwJwSg4sN/avoiding-doomsday-a-proof-of-the-self-indication-assumption

Armstrong, Stuart (2011a). "Anthropic decision theory." *arXiv* https://arxiv.org/pdf/1110.6437.pdf

Armstrong, Stuart (2011b). "Dead men tell tales: falling out of love with SIA." *LessWrong* https://www.lesswrong.com/posts/LnearFbA4thE646tR/dead-men-tell-tales-falling-out-of-love-with-sia

Arntzenius, Frank (2002). "Reflections on Sleeping Beauty." *Analysis* 62, no. 1: 53-62

Arntzenius, Frank (2003). "Some Problems for Conditionalization and Reflection." *The Journal of Philosophy* 100, no. 7: 356-370

Arntzenius, Frank (2014). "Utilitarianism, decision theory and eternity." *Philosophical Perspectives* 28, no. 1: 31–58

Arntzenius, Frank and Dorr, Cian (2017). "Self-locating Priors and Cosmological Measures." *The Philosophy of Cosmology*, eds. Khalil Chamchan, John Barrow, Simon Saunders, and Joe Silk. Cambridge University Press*:* 396-428

Arrhenius, Gustaf (2011). "The Impossibility of a Satisfactory Population Ethics." *Descriptive and Normative Approaches to Human Behavior, Advanced Series on Mathematical Psychology*, eds. Hans Colonius and Ehtibar N. Dzhafarov. World Scientific Publishing Company: 1–26

Askell, Amanda (2018). *Pareto Principles in Infinite Ethics*. [PhD dissertation, New York University] https://askell.io/files/Askell-PhD-Thesis.pdf

Beckstead, Nick (2013). *On the Overwhelming Importance of Shaping the Far Future*. [PhD dissertation, Rutgers University - New Brunswick] https://drive.google.com/file/d/0B8P94pg6WYCIc0lXSUVYS1BnMkE/view?resourcekey=0-nk6wM1QIPl0qWVh2z9FG4Q

Beckstead, Nick and Thomas, Teruji (2021). "A paradox for tiny possibilities and enormous values." *Global Priorities Institute* (GPI Working Paper No. 7) https://globalprioritiesinstitute.org/wp-content/uploads/Beckstead-Thomas-A-Paradox-for-Tiny-Probabilities-and-Enormous-Values-Version-2.pdf

Besnard, Fabien (2004). "Refutations of the Simulation Argument." *Bienvenue sur Mathéphysique !* http://fabien.besnard.pagesperso-orange.fr/pdfrefut.pdf

Bollard, Lewis (2019). "Will Companies Make Good on Cage-Free Pledges?" *Open Philanthropy Project Farm Animal Welfare Newsletter* https://us14.campaign-archive.com/?u=66df320da8400b581cbc1b539&id=01387dda51&fbclid=IwAR2kAiKh02gmX0AKP5ICAaX02Oh8E7KZgT8OoOwpDoLk6Dyyafsfj0pLL_0

Bostrom, Nick (2001). "The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe." https://anthropic-principle.com/preprints/cau/paradoxes.

Bostrom, Nick (2002a). *Anthropic Bias*. Routledge.

Bostrom, Nick (2002b). "Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation." *The Journal of Philosophy* 99, no. 12: 607-623

Bostrom, Nick (2003a). "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53, no. 211: 243-255

Bostrom, Nick (2003b). "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15, no. 3: 308-314

Bostrom, Nick and Ćirković, Milan M. (2003). "The Doomsday Argument and the Self-Indication Assumption: Reply to Olum." *The Philosophical Quarterly* 53, no. 210: 83-91

Bostrom, Nick (2005). "The Simulation Argument: Reply to Weatherson." *The Philosophical Quarterly* 55, no 218: 90-97

Bostrom, Nick, (2006). "Quantity of experience: brain-duplication and degrees of consciousness." *Minds and Machines* 16, no. 2: 185-200

Bostrom, Nick (2007). "Sleeping Beauty and Self-Location: A Hybrid Model." *Synthese* 157, no 1: 59-78

Bostrom, Nick (2009). "Pascal's mugging." *Analysis* 69, no. 3: 443-445

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

Bostrom, Nick and Kulczycki, Marcin (2011). "A Patch for the Simulation Argument." *Analysis* 71, no 1: 54-61

Bostrom, Nick (2013). "Existential Risk Prevention as Global Priority." *Global Policy* 4, no 1: 15-31

Bostrom, Nick (2014). "Crucial Considerations and Wise Philanthropy." [Speech audio recording] *Radio Bostrom* https://radiobostrom.com/12/crucial-considerations-and-wise-philanthropy

Briggs, Rachael (2010). "Putting a Value on Beauty," in *Oxford Studies in Epistemology, Volume 3*, eds. Tamar Szabo Gendler and John Hawthorne, Oxford University Press: 3-34.

Campbell, Tim (2021). "Personal Identity and Impersonal Ethics." *Principles and Persons: The Legacy of Derek Parfit*, eds. Jeff McMahan, Tim Campbell, James Goodrich, and Ketan Ramakrishnan. Oxford University Press: 55-84

Carlsmith, Joseph (2021a). "Can you control the past?" *LessWrong* https://www.lesswrong.com/posts/PcfHSSAMNFMgdqFyB/can-you-control-the-past

Carlsmith, Joseph (2021b). "On the Universal Distribution." LessWrong https://www.lesswrong.com/posts/XiWKmFkpGbDTcsSu4/on-the-universal-distribution

Carlsmith, Joseph (2021c). "Anthropics and the Universal Distribution." *LessWrong* https://www.lesswrong.com/posts/Hcc9fopx7sRexYhhi/anthropics-and-the-universal-distribution

Carlsmith, Joseph (2021d). "SIA > SSA, part 2: Telekinesis, reference classes, and other scandals." *LessWrong* https://www.lesswrong.com/posts/GJdymoviRywpBMXqc/sia-greater-than-ssa-part-2-telekinesis-reference-classes

Carroll, Sean M. (2020). "Why Boltzmann Brains Are Bad." *Current Controversies in Philosophy of Science*, eds. Shame Dasgupta and Brad Weslake. Routledge: 7-20

Carroll, Sean and Bostrom, Nick (2020). "Nick Bostrom on Anthropic Selection and Living in a Simulation." [Podcast recording] *Mindscape* (Episode 111) https://www.preposterousuniverse.com/111-nick-bostrom-on-anthropic-selection-and-living-in-a-simulation/

Chalmers, David (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton & Company.

Chanko, Kenneth M. (1995). "Real pigs steal the scene in 'Babe.'" *Entertainment Weekly* https://ew.com/article/1995/08/18/real-pigs-steal-scene-babe/

Chappell, Richard Yetter (2011). "Value Holism." *PhilPapers* https://philpapers.org/archive/CHAVH.pdf

Chichilnisky, Graciela (1996). "An axiomatic approach to sustainable development." *Social Choice and Welfare* 13, no. 2: 231-257

Christiano, Paul (2011). "The Absolute Self-Selection Assumption." *LessWrong* https://www.lesswrong.com/posts/QmWNbCRMgRBcMK6RK/the-absolute-self-selection-assumption

Christiano, Paul (2021). "EDT with updating double counts." *The sideways view* https://sideways-view.com/2021/10/12/edt-with-updating-double-counts/.

Ćirković, Milan M. (2004). "Is Many Likelier than Few? A Critical Assessment of the Self-indication Assumption." *Epistemology* 27, no. 2: 265-298*a*

Conway, John H. (2000). *On Numbers and Games* (2nd ed.). A K Peters / CRC Press.

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

Cowen, Tyler and Parfit, Derek (1992). "Against the Social Discount Rate." *Justice Between Age Groups and Generations*, eds. Peter Laslett and James Fishkin. Yale University Press: 144-161

Crawford, Lyle (2013). "Freak Observers and the Simulation Argument." *Ratio* 26, no. 3: 250-264

Dorr, Cian (2002). "Sleeping Beauty: in defense of Elga." *Analysis* 62, no. 4: 292-296

Easwaran, Kenny (2021). "A New Method For Value Aggregation." *Proceedings of the Aristotelian Society* 121, no. 3:, 299–326.

Elga, Adam (2004). "Defeating dr. evil with self-locating belief." *Philosophy and Phenomenological Research* 69, no. 2: 383-396

Enoch, David (2011). *Taking Morality Seriously: A Defense of Robust Realism* (1st ed.). Oxford University Press

Garfinkel, Ben (unpublished) "Objections to the simulation argument." https://docs.google.com/document/d/1TSvjJL1uJGtREaKGlau9A_oA-Fj6ny2ieRpzhfbKfqA/edit

Garrabrant, Scott (2014). "Preferences without Existence." *LessWrong https://www.lesswrong.com/posts/NvwJMQvfu9hbBdG6d/preferences-without-existence*

Guth, Alan (1997). "A Universe In Your Backyard." *Third Culture: Beyond the Scientific Revolution*, eds. John Brockton. Simon & Schuster: 276-286

Grace, Katja (2010a). "Anthropic principles agree on bigger future filters." *Meteuphoric https://meteuphoric.com/2010/11/02/anthropic-principles-agree-on-bigger-future-filters/*

Grace, Katja (2010b). "SIA doomsday: The filter is ahead." *Meteuphoric* https://meteuphoric.com/2010/03/23/sia-doomsday-the-filter-is-ahead/

Grace, Katja (2011). "The Unpresumptuous Philosopher." *Meteuphoric* https://meteuphoric.com/2011/01/10/the-unpresumptuous-philosopher/

Greaves, Hilary and MacAskill, William (2021). "The case for strong longtermism." *Global Priorities Institute* (GPI Working Paper No. 5) https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf

Greaves, Hilary and Ord, Toby (2017). "Moral Uncertainty About Population Axiology." *Journal of Ethics and Social Philosophy* 12, no. 2: 135-167

Greene, Preston (2020). "The Termination Risks of Simulation Science." *Erkenntnis* 85, no. 2: 489-509

Halpern, Joseph (2006). "Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems." *Oxford Studies in Epistemology* 1

Hanson, Robin (2001). "How To Live In A Simulation." *Journal of Evolution and Technology* 7, no. 1

Hájek, Alan (2017). "Pascal's Wager." *Stanford Encyclopedia of Philosophy Archive* https://plato.stanford.edu/archives/fall2017/entries/pascal-wager/

Hamkins, Joel David and Montero, Barbara (2000). "With Infinite Utility, More Needn't Be Better." *Australasian Journal of Philosophy* 78, no. 2: 231-240

Hanson, Robin (1998). "The Great Filter – Are We Almost Past It?" https://mason.gmu.edu/~rhanson/greatfilter.html

Hedden, Brian (2015). "Time-Slice Rationality." *Mind* 124, no. 494: 449-491

Hitchcock, Christopher (2004). "Beauty and the bets." *Synthese* 139, no. 3: 405-420

Holt, Jim (2021). "The Power of Catastrophic Thinking." *The New York Review* https://www.nybooks.com/articles/2021/02/25/power-catastrophic-thinking-toby-ord-precipice/

Huemer, Michael (2021). "Existence is Evidence of Immortality." *Noûs* 55: 128-151

Isaacs, Yooav, John Hawthorne, and Jeffrey Sanford Russell (2021). "Multiple Universes and Self-Locating Evidence." *PhilArchive* https://philarchive.org/archive/ISAMUA-2

Jäger, Jens (2021). "Immortal Beauty: Does Existence Confirm Reincarnation?" *Australasian Journal of Philosophy* 100, no. 4: 798-807

Jonsson, Adam and Voorneveld, Mark (2018). "The limit of discounted utilitarianism." *Theoretical Economics* 13, no. 1: 19-37

Karnofksy, Holden (2018). "Update on Cause Prioritization at Open Philanthropy." *Open Philanthropy* https://www.openphilanthropy.org/research/update-on-cause-prioritization-at-open-philanthropy/

Lauwers, Luc (2010). "Ordering infinite utility streams comes at the cost of a non-Ramsey set." *Journal of Mathematical Economics* 46, no. 1: 32-37

Lauwers, Luc (2016). "The Axiomatic Approach to the Ranking of Infinite Streams." *The Economics of the Global Environment,* eds. Graciela Chichilinsky and Armon Rezai. Springer: 231-255

Leslie, John (1996). *The End of the World: The Science and Ethics of Human Extinction* (1st ed.). Routledge.

Lewis, David (1979). "Attitudes *De Dicto* and *De Se*." *The Philosophical Review* 88, no. 4: 513-543

Lewis, David K. (1980). "A subjectivist's guide to objective chance." *Studies in Inductive Logic and Probability* 2, ed. Richard C. Jeffrey: 263-293

Levinstein, Benjamin A. and Soares, Nate (2020). "Cheating Death in Damascus." *Journal of Philosophy* 117, no. 5: 237-266

MacAskill, William (2014). "Replaceability, Career Choice, and Making a Difference." *Ethical Theory and Moral Practice* 17, no. 2: 269-283

MacAskill, William (2015). *Doing Good Better: Effective Altruism and How You Can Make A Difference.* Random House.

MacAskill, William (2019). "The Definition of Effective Altruism." *Effective Altruism: Philosophical Issues*, eds. Hilary Greaves and Theron Pummer. Oxford University Press: 10-28

MacAskill, William, Aron Vallinder, Carl Shulman, Caspar Österheld, and Johannes Treutlein (2021). "The Evidentialist's Wager." *Journal of Philosophy* 118, no. 6: 320-342

Macaskill, William (2022). *What We Owe The Future.* Basic Books.

Manley, David (unpublished). "On being a random sample." *U-M Personal WWW Server* http://www-personal.umich.edu/~dmanley/manley/David_Manley_files/On%20Being%20a%20Random%20Sample.pdf

Matthews, Dylan (2015). "I spent a weekend at Google talking with nerds about charity. I came away … worried." *Vox* https://www.vox.com/2015/8/10/9124145/effective-altruism-global-ai

Matthews, Dylan (2022). "How effective altruism went from a niche movement to a billion-dollar force." *Vox* https://www.vox.com/future-perfect/2022/8/8/23150496/effective-altruism-sam-bankman-fried-dustin-moskovitz-billionaire-philanthropy-crytocurrency

McMahan, Jeff (2013). "Causing People to Exist and Saving People's Lives." *The Journal of Ethics* 17, no. 1/2: 5-35

Meacham, Christopher J. G. (2008). "Sleeping beauty and the dynamics of de se beliefs." *Philosophical Studies* 138, no. 2: 245-269

Meacham, Christopher J. G. (2010). "Binding and its Consequences." *Philosophical Studies* 149, no.1: 49-71

Meacham, Christopher J. G. (2016). "Ur-Priors, Conditionalization, and Ur-Prior Conditionalization." *Ergo: An Open Access Journal of Philosophy* 3, no. 17

Mogensen, Andreas L. (2020). "Moral demands and the far future." *Philosophy and Phenomenological Research* 103, no. 3: 567-585

Monton, Bradley (2019). "How to Avoid Maximizing Expected Utility." *Philosophers' Imprint* 19, no. 18

Neal, Radford (2006). "Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning." *arXiv* https://arxiv.org/pdf/math/0608592.pdf

Nover, Harris and Hájek, Alan (2004). "Vexing Expectations." *Mind* 113, no. 450: 237-249

Nozick, Robert (1969). "Newcomb's problem and two principles of choice." *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher: 114-146

Oesterheld, Caspar (2017). "Multiverse-wide Cooperation via Correlated Decision Making." *Center on Long-Term Risk* https://longtermrisk.org/files/Multiverse-wide-Cooperation-via-Correlated-Decision-Making.pdf

Olson, Jay S. and Ord, Toby (2021). "Implications of a search for intergalactic civilizations on prior estimates of human survival and travel speed." *arXiv* https://arxiv.org/pdf/2106.13348.pdf

Olum, Ken D. (2002). "The Doomsday Argument and the Number of Possible Observers." *The Philosophical Quarterly* 52, no. 207: 164-184

Ord, Toby (2002). *Hypercomputation: Computing more than the Turing machine.* [PhD dissertation, University of Melbourne] https://arxiv.org/pdf/math/0209332.pdf

Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity.* Hachette Books

Parfit, Derek (1984). *Reasons and Persons* (1st ed.). Oxford University Press

Peterson, Martin (2019). "The St. Petersburg Paradox." *Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/entries/paradox-stpetersburg/

Ramakrishnan, Ketan (2021). "Deontology over time." [Virtual event] *University Center for Human Values* (Ira W. DeCamp Bioethics Seminar) https://uchv.princeton.edu/events/ketan-ramakrishnan-university-oxford-and-yale-law-school-deontology-over-time

Ramsey, Frank P. (1928). "A Mathematical Theory of Saving." *The Economic Journal* 38, no. 152: 543-559

Russell, Jeffrey Sanford (2021). "On two arguments for Fanaticism." *Global Priorities Institute* (GPI Working Paper No. 17) https://globalprioritiesinstitute.org/wp-content/uploads/Jeffrey-Sanford-Russell_On-two-arguments-for-Fanaticism.pdf

Russell, Jeffrey Sanford (2022). "Problems for Impartiality." [Lecture] *Global Priorities Institute* (Parfit Memorial Lecture 2022) https://globalprioritiesinstitute.org/parfit-memorial-lecture-jeffrey-sanford-russell-problems-for-impartiality/

Samuel, Sigal (2022). "Effective altruism's most controversial idea." *Vox* https://www.vox.com/future-perfect/23298870/effective-altruism-longtermism-will-macaskill-future

Shulman, Carl and Bostrom, Nick (2012). "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." *Journal of Consciousness Studies* 19, no. 7-8: 103-130

Sidgwick, Henry (1907). *The Methods of Ethics* (7th ed.). Macmillan and Company.

Singer, Peter (2021). "The Hinge of History." *Project Syndicate* https://www.project-syndicate.org/commentary/ethical-implications-of-focusing-on-extinction-risk-by-peter-singer-2021-10?barrier=accesspaylog

Smith, Nicholas J. J. (2014). "Is Evaluative Compositionality a Requirement of Rationality?" *Mind* 123, no. 490: 457-502

Snyder-Beattie, Andrew et al (2021). "The Timing of Evolutionary Transitions Suggests Intelligent Life is Rare." *Astrobiology* 21, no. 3: 265-278

Tarsney, Christian and Thomas, Teruji (2020). "Non-additive axiologies in large worlds." *Global Priorities Institute* (GPI Working Paper No. 9) https://globalprioritiesinstitute.org/christian-tarsney-and-teruji-thomas-non-additive-axiologies-in-large-worlds/

Tegmark, Max (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality.* Knopf.

Thomas, Teruji (2021a). "Doomsday and objective chance" *Global Priorities Institute* (GPI Working Paper No. 8) https://globalprioritiesinstitute.org/wp-content/uploads/Thomas-Doomsday-and-Objective-Chance-Version-2.pdf

Thomas, Teruji (2021b). "Simulation expectation." *Global Priorities Institute* (GPI Working Paper No. 16) https://globalprioritiesinstitute.org/wp-content/uploads/Teruji-Thomas-Simulation-expectation.pdf

Thornley, Elliott (2021). "The impossibility of a satisfactory population prospect axiology (independently of Finite Fine-Grainedness)." *Philosophical Studies* 178, no. 11: 3671-3695

Thorstad, David (2022). "Existential risk pessimism and the time of perils." *Global Priorities Institute* (GPI Working Paper No. 1) https://globalprioritiesinstitute.org/wp-content/uploads/David-Thorstad-Existential-risk-pessimism-.pdf

Todd, Benjamin (2021). "Is effective altruism growing? An update on the stock of funding vs. people." *80,000 Hours* https://80000hours.org/2021/07/effective-altruism-growing/

Tomasik, Brian (2016). "How the Simulation Argument Dampens Future Fanaticism." *Center on Long-Term Risk* https://longtermrisk.org/files/how-the-simulation-argument-dampens-future-fanaticism.pdf

Torres, Émile P (2021). "Against longtermism." *Aeon*, ed. Sam Dresser https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo

Vallentyne, Peter and Kagan, Shelly (1997). "Infinite Value and Finitely Additive Value Theory." *The Journal of Philosophy* 94, no. 1: 5-26

Van Liedekerke, Luc (1995). "Should Utilitarians Be Cautious About An Infinite Future?" *Australasian Journal of Philosophy* 73, no. 3: 405-407

Weatherson, Brian (2003). "Are You a Sim?" *The Philosophical Quarterly* 53, no. 212: 425-431

Weatherson, Brian (2005). "Should We Respond to Evil with Indifference?" *Philosophy of Phenomenological Research* 70, no. 3: 613-635

Weirich, Paul (2020). "Causal Decision Theory." *Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/entries/decision-causal/#ExpeUtil

Wiblin, Robert and Harris, Keiran (2018). "Tackling the ethics of infinity, being clueless about the effects of our actions, and having moral empathy for intellectual adversaries, with philosopher Dr Amanda Askell" [Podcast] *80,000 Hours* https://80000hours.org/podcast/episodes/amanda-askell-moral-empathy/

Wiblin, Robert and Harris, Keiran (2021). "Ajeya Cotra on worldview diversification and how big the future could be." [Podcast] *80,000 Hours* https://80000hours.org/podcast/episodes/ajeya-cotra-worldview-diversification/#top.

Wilkinson, Hayden (2021a), "Infinite aggregation: expanded addition." *Philosophical Studies* 178, no. 6: 1917-1949

Wilkinson, Hayden (2021b). "Chaos, add infinitum." *PhilPapers* https://philpapers.org/archive/WILCAI-11.pdf

Wilkinson, Hayden (2022a). "In Defence of Fanaticism." *Ethics* 132, no. 2: 445-477

Wilkinson, Hayden (2022b), "Infinite Aggregation and Risk." *Australasian Journal of Philosophy,* DOI*:* 10.1080/00048402.2021.2013265

Wolchover, Natalie and Byrne, Peter (2014). "In a Multiverse, What Are the Odds?" *Quanta Magazine* https://www.quantamagazine.org/the-multiverses-measure-problem-20141103/

Xu, Mark and Shulman, Carl (2021). "The Simulation Hypothesis Undercuts the SIA/Great Filter Doomsday Argument." *LessWrong* https://www.lesswrong.com/posts/HF2vpnmgqmHyLGRrA/the-simulation-hypothesis-undercuts-the-sia-great-filter

Xu, Mark (2021). "Strong Evidence is Common." *Artificially Intelligent* https://markxu.com/strong-evidence

Yamada, Masahiro (2019). "Beauty, odds, and credence." *Philosophical Studies* 176: 1247-1261

Yudkowsky, Eliezer (2009). "The Anthropic Trilemma." *LessWrong* https://www.lesswrong.com/posts/y7jZ9BLEeuNTzgAE5/the-anthropic-trilemma

Yudkowsky, Eliezer and Soares, Nate (2018). "Functional Decision Theory: A New Theory of Instrumental Rationality." *arXiv* https://arxiv.org/pdf/1710.05060.pdf

Zame, William R. (2007), "Can intergenerational equity be operationalized?" *Theoretical Economics* 2, no. 2: 187-202

Zuber et al. (2021). "What Should We Agree on about the Repugnant Conclusion?" *Utilitas* 33, no. 4: 379-383