

SIA vs SSA

Joe Carlsmith

September 2022

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”).¹ 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 observers. Various philosophers treat this implication (known as the “Presumptuous Philosopher”) as a basically decisive objection to SIA.²

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.³

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

¹Bostrom (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.

²See e.g. Leslie (1996); Ćirković (2001); Bostrom (2002a); and Arntzenius and Dorr (2016).

³Here 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.

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);
- 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").⁴ 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.⁵ Further, let's call all of a subject's evidence at a given time her "epistemic situation."⁶ 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).⁷ I want the anthropic principles I discuss to treat

⁴See Lewis (1979) for a classic discussion.

⁵Elga (2004) and Manley (unpublished) also call this a "predicament." Formally, 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.

⁶Here 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 in 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.

⁷While 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 in a more coarse-grained

uncertainty about whether you are 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).⁸

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.⁹

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.

⁸The 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)).

⁹The 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.¹⁰

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*.¹¹ 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.¹² 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 wounded deer suddenly appearing before you.¹³ I'll sometimes call the

¹⁰See Elga (2004) for a classic discussion, and Weatherson (2005) for some objections.

¹¹I 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.

¹²See e.g. Manley (unpublished), p. 20-22 and Isaacs et al (2021), p. 20, who formulate SIA and SSA in a manner very similar to the way I do.

¹³Thus, 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”—

evidence that informs the prior probability distribution “non-anthropoc 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.¹⁴ 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.¹⁵ 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 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).¹⁶ Indeed, I don’t view this particular story about priors as especially

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.

¹⁴After 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.

¹⁵See Manley (unpublished) and Issacs et al (2021) for more on this sort of story. On such a story, the “non-anthropoc evidence” I discussed above would be just: the *de dicto* evidence.

¹⁶One 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

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.¹⁷ What matters most, in my opinion, is the ultimate *verdicts* of the theories in question in the types of cases I'll discuss (for example,

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.

¹⁷Thus, 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 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

GOD’S EXTREME COIN TOSS)—verdicts produced by the underlying way in which SIA and SSA favor different types of objective worlds.¹⁸ 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.¹⁹ 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.)

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:

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

And SSA says that:

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

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.

¹⁸Specifically, 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.

¹⁹I 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.

(In what follows, I'll mostly present calculations of this form in terms of odds-ratios, which I find easier to think about.²⁰)

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.²¹

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.²²

²⁰The 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 $p_r(O_x)n(O_x) : p(O_y)n(O_y)$, and SSA says that it's $p_r(O_x)\frac{n(O_x)}{r(O_x)} : p(O_y)\frac{n(O_y)}{r(O_y)}$.

²¹For 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.

²²We'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:

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 $1/2$?" A key answer is: if you don't update the prior, and instead skip straight to apportioning

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 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, $1/3$ on heads); whereas SSA updates *against* the Tails-AB and Tails-BA worlds (since A-havers are only $1/2$ of the reference class in those worlds), for a final odds ratio of 2:1: $\frac{1}{2}$: $\frac{1}{2}$ (and so, $1/2$ on heads). Thus, you get the classic "thirder" and "halfer" 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,

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.

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.²³

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.²⁴

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 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

e.g., even if the universe’s real conditions are very different, *some* observer will make the observations you are making).

²³Now, 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.

²⁴See also Bostrom’s (2002a, p. 122) “heavenly messenger” analogy, and Manley’s (unpublished, section 5) analogy with marbles.

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.²⁵ However, I think this story may run into instabilities (see footnote), so I tend to stick with the story above.²⁶

²⁵See e.g. Olson and Ord (2021) for an example of this sort of framing.

²⁶On 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: “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

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.

(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

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.

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%.²⁷ Thus, with a red jacket in GOD’S COIN TOSS WITH EQUAL NUMBERS above, it

²⁷Indeed, 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.

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 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.²⁸ 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

²⁸That is, on R-SIA + SSA, the posterior odds ratio $p(O_x) : p(O_y)$ is $p_r(O_x)r(O_x) \frac{n(O_x)}{r(O_x)} : p_r(O_y)r(O_y) \frac{n(O_y)}{r(O_y)}$, which reduces to $p_r(O_x)n(O_x) : p(O_y)n(O_y)$.

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. 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.²⁹ 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?

(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.

²⁹I discuss some other candidate views in the footnotes of section XVI below.

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.³⁰ 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-anthropocentric) 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.

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.³¹

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 T₂.³² 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).

³⁰For 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.

³¹Of 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.

³²Though note that betting in anthropics implicates a number of additional issues—see section XIII below.

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 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.³³ 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-anthropropic 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.³⁴

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-anthropropic 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-anthropropic empirical evidence favors T₁ by a factor of a thousand seems like it might be positing something very weird 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

³³See Carl Shulman's comment on Grace (2011).

³⁴That 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.

certainly wouldn't happen.³⁵

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-anthropocentric 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.³⁶ 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.³⁷

Indeed: SSA, famously, can lead to the "Doomsday Argument," which is structurally analogous to the case just given.³⁸ 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 and flourish long enough for ~200 trillion humans to live instead. On the basis of the available non-anthropocentric 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.³⁹ 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).⁴⁰ But the bare possibility of making such an argument, on SSA, suggests that un-presumptuousness isn't exactly SSA's strong suit, either.

³⁵That 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. . .

³⁶Or at least, it does if we use a reference class that includes the non-earth observers—more on trying to avoid that below.

³⁷See Grace (2011) for more on the parallels here.

³⁸See e.g. Leslie (1996) for classic discussion, and Bostrom (2002a) for in-depth analysis.

³⁹The 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.

⁴⁰Thus, 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.

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.

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.⁴¹

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.⁴² And once can imagine arguing that SIA, too, says weird things about fair coins. After all, SIA is 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

⁴¹ 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.

⁴² See, for example, Lewis (1980) and Thomas (2021a) for discussion.

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?⁴³

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 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.⁴⁴ 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,

⁴³See Bostrom's (2001) *UN++* and *Lazy Adam* for related examples.

⁴⁴Below I discuss SIA's verdicts once you're already in the white room, but I want to start with this version.

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 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 $1/3$ rd. 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.⁴⁵

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.

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

⁴⁵This, in my opinion, is also the thing to say about Yudkowsky's (2009) "Anthropic Trilemma."

the puppy. Indeed, on evidential decision theory, they will pay to get access to this sort of button.⁴⁶ And in 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 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

⁴⁶Evidential 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—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.

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*,⁴⁷ five bug-eyed aliens 10¹⁰⁰ 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.

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.⁴⁸ 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,

⁴⁷See Chanko (1995). “‘There was,’ Miller admits reluctantly, ‘one animatronic pig’”.

⁴⁸See Bostrom (2022a), p. 171.

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 appeal to some notion of metaphysical essence or modal profile.⁴⁹ 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” in 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?⁵⁰ 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 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

⁴⁹This 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.

⁵⁰I 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.

whether you count chimps in the jungle as in the reference class or not.⁵¹

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

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?

⁵¹Indeed, 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 between 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").

- **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.⁵²

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

⁵²Bostrom 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 t_1 ," 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.

in it otherwise.⁵³ Perhaps some philosophers won't be bothered by this. After all, fitting the cases well is an important 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 off 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.⁵⁴

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

⁵³Indeed, 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.

⁵⁴Indeed, 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).

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.⁵⁵

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 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?⁵⁶

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.⁵⁷ 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

⁵⁵Indeed, 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.

⁵⁶See also a closely-related version in Dorr (2002), and a related series of cases in Arntzenius (2002)).

⁵⁷Here I'm returning to a version of SSA that keeps the reference class constant.

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,⁵⁸ and I'm inclined to agree.

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 Dorrr, 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 halving. 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."⁵⁹ 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 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

⁵⁸At least, circa 2009; he's since changed his view (see Armstrong (2011a), for reasons to do with decision theory.

⁵⁹See discussion in Bostrom (2002b), p. 196.

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),⁶⁰ 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.⁶¹ 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.⁶² You might be initially tempted by the following argument for thirding, rather than halving, 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 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).⁶³ Is that surprising? It might initially seem that way.

⁶⁰As a reminder, EDT says to choose the action such that your choosing it is the best *news*; CDT says to choose the action with the best causal effects. See Weirich (2020) for more.

⁶¹See Meacham (2010) for an approach to decision-theory in this broad vicinity; and see Carlsmith (2021a, section IX) for more.

⁶²See e.g. Hitchcock (2004), Arntzenius (2002), Briggs (2010), and Yamada (2019) for more discussion of betting in the context of anthropics.

⁶³See Yamada (2019), p. 1249-1251, for more on the relationship between credences and acceptable betting odds.

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 tails*. 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.⁶⁴

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).⁶⁵ 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).⁶⁶ That is, every altruistic EDT-er bets, sometimes, in a way that would naively (but wrongly) suggest a "*fifth-ing*" pattern of credence.⁶⁷

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

⁶⁴See Arntzenius (2002) for more on this.

⁶⁵See Yamada (2019), p. 1254, for discussion of this.

⁶⁶Thanks to Paul Christiano and Katja Grace for discussion. See Christiano (2021) for more on "*fifth-ing*."

⁶⁷At 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).

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.⁶⁸

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?⁶⁹ Does the fact that we exist mean that there are probably lots of simulations of us?⁷⁰ Can we use earth's evolutionary history as evidence for the frequency of intelligent life?⁷¹ 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 number of people-in-our-epistemic-situation they posit or imply. Thus, and especially

⁶⁸Here'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 to 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).

⁶⁹See Leslie (1996).

⁷⁰See e.g. Shulman and Bostrom (2012) and Xu and Shulman (2021).

⁷¹See e.g. Synder-Beattie et al (2021).

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.⁷²

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 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

⁷²I 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.

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.⁷³

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 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.⁷⁴ 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.⁷⁵ Indeed, this sort of certainty seems like a natural extension of the logic of the presumptuous philosopher.⁷⁶ And this does seem overconfident.

⁷³See Olum (2000, p. 15).

⁷⁴See Beckstead and Thomas (2021) and Wilkinson (2021) for discussion.

⁷⁵Other options include: becoming undefined on infinite cases, or trying to carve out ad hoc exceptions for infinite cases. See Manley (unpublished) for discussion.

⁷⁶Just as obsession with infinitely high-stakes outcomes seems like a natural extension

One might think: surely the universe *could be* finite. Finitude, after all, is a live scientific hypothesis.⁷⁷ 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.⁷⁸ 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 drawn is zero.⁷⁹ 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.⁸⁰ 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.⁸¹ Sup-

of vulnerability to a finite pascal's mugging. See Thomas and Beckstead (2021) for more on the close connection here.

⁷⁷See 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."

⁷⁸See 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.

⁷⁹Assuming it's defined at all.

⁸⁰In 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.

⁸¹In conversation, Paul Christiano characterized this to me as a kind of "double

pose, 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 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 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 of people in your epistemic situation. Moves like these, though, bring in substantial additional complications and objections.⁸²

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.⁸³ 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.⁸⁴

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

whammy.” First SIA becomes certain that it's in an infinite world—and then it immediately breaks.

⁸²For 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.

⁸³See Weatherson (2005) and Manley (unpublished).

⁸⁴See e.g. the discussion in Wolchover and Byrne (2014): “Few consider the problem to be solved.”

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.

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 $1/2$ 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 $1/2$ 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.⁸⁵ 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-

⁸⁵That is, on Sunday, a policy of accepting a bet of “\$20 conditional on tails, \$10 conditional on heads” is neutral in expectation.

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.⁸⁶

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.⁸⁷ 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 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.⁸⁸ So one might reasonably wonder: given the

⁸⁶See Yudkowsky and Soares (2009, p. 8) for discussion.

⁸⁷See, for example, Levinstein and Soares (2020) and Meacham (2010).

⁸⁸As 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-halving" behavior I discussed in the text, which I see as incompatible with very basic

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.⁸⁹ 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.

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-halving 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.

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).

⁸⁹Thanks 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.

Consider, for example, SLEEPING BEAUTY. What are you going to say in SLEEPING BEAUTY, if not $1/2$, or $1/3$?⁹⁰ 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 Dorr/Arntzenius argument that "you should be $1/4$ th 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 third." 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).⁹¹

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.⁹² 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 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

⁹⁰Let's leave aside "fifthing" for now, along with moves that attempt to incorporate your uncertainty about which theory of anthropics is correct.

⁹¹This is the route taken by "UDASSA"—see Carlsmith (2021b) for discussion.

⁹²See 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.

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.⁹³

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.⁹⁴ 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

⁹³See Grace (2010) and Olson and Ord (2021).

⁹⁴See also Shulman and Bostrom (2012).

“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).

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.⁹⁵ 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 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.⁹⁶

⁹⁵This 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.

⁹⁶This essay owes an especially large amount to discussion with Katja Grace, and to her

work on anthropics. Thanks, as well, to Amanda Askill, Nick Beckstead, Paul Christiano, Tom Davidson, Carl Shulman, Bastian Stern, and Ben Weinstein-Raun for discussion.