SIA > SSA

Joe Carlsmith

September 30, 2021

This post is the first in a four-part sequence explaining why I think that one prominent approach to anthropic reasoning is better than another. 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." What should your credence be that the coin landed heads?

The approach I like better—the "Self Indication Assumption" (SIA)—says: ~one in a million. SIA thinks you're more likely to exist in worlds with more people in your epistemic situation. Here, this is the tails-world by far.

The approach I like worse—the "Self-Sampling Assumption" (SSA)—says: one half. SSA thinks you're more likely to exist in worlds where the people in your epistemic situation are a larger fraction of the people in your "reference class." Don't ask me what a reference class is, but in this case, let's assume that the people in your epistemic situation are 100% of it either way. So SSA sticks with the one half prior.

I open with this case because it's one of the worst for SIA, the approach I favor. In particular, we can construct scientific analogs, in which SIA becomes ludicrously confident in a given cosmology, simply in virtue of that cosmology positing more people in our epistemic situation. For many, this implication (known as the "Presumptuous Philosopher") is a ~decisive objection to SIA.

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.

I proceed as follows. The first part of the sequence ("Learning from the fact that you exist") describes SIA and SSA. In particular, I emphasize that pace some presentations in the literature, SIA should not be seen as an *additional assumption you add to SSA*—one that "cancels out" SSA's bad implications, but accepts SSA's worldview. Rather, SIA is a different (and more attractive) picture altogether.

The second part ("Telekinesis, reference classes, and other scandals") lays out the bulk of my case against SSA. In particular, SSA implies:

- comparable scientific presumptuousness;
- ~certainty that fair coins that haven't yet been flipped will come up heads;
- expecting to be able to move boulders with your mind;
- a made-up and indeterminate ontology of reference classes that really seems like it shouldn't be a thing, and which also shouldn't be contorted willy-nilly to get whatever results you want (and also, this doesn't work anyway);
- sensitivity to differences that don't matter (like whether a definitely-not-you observer is killed and/or given evidence you don't get, vs. never created; or whether there are ten vs. twenty chimps in the jungle surrounding your thought experiment);

- a "special snowflake" metaphysics on which, because you actually exist, you had to exist in any world still compatible with your evidence (though I think that SSA has some replies here);
- intense updates towards solipsism.

The third part ("An aside on betting in anthropics") briefly discusses betting in anthropics. In particular: why it's so gnarly, why I'm not focusing on it, and why I don't think it's the only desiderata.

The fourth part ("In defense of the presumptuous philosopher") discusses prominent objections to SIA in more detail. In particular:

- I suggest that the Presumptuous Philosopher is a strong candidate for what I call a "good bullet": that is, a counterintuitive result, acceptance of which resolves a lot of gnarly issues into a simple and pretty satisfying theory, and rejection of which invites endless complication and counterexample (fans of the repugnant conclusion, take note). This doesn't mean we should actually bite. But we shouldn't die on the hill of non-biting.
- Pascal's muggings and infinities are problems for SIA, but this puts SIA in pretty respectable company (expected utility theory, population ethics)—and in particular, company that still seems to make itself useful.
- SIA has problems with "counting observers." I haven't thought that much about this one, but I have some feeling like: don't we all?
- Given some values and decision theories, SIA (like SSA) suffers from inconsistencies between "the policy you'd want to commit to, from some 'prior' epistemic perspective" and "how you behave ex post." But these inconsistencies are common in many other contexts, too; and if you're worried about them, use an "updateless" decision theory. But I suggest not throwing out the concept of epistemology along the way.
- We can maybe do a bit to make SIA more intuitive (though not as much as I'd like).

That said, even if SSA is worse than SIA, it's not like SIA is sitting pretty (I especially don't like how it breaks in infinite cases, and there are presumably many other objections I'm not considering). I briefly discuss whether we should expect to find a better alternative—the "Anthropic Theory X" above. My current answer is: maybe (and maybe it's already out there), 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 close by quickly mentioning some of SIA's possible implications in the real world (for example, re: doomsday arguments). I think we should tread carefully, here, but stay curious.

Acknowledgments: This sequence owes a huge amount to discussion with Katja Grace, and to her work on anthropics (see summary here, her honors thesis here, and the many links throughout the sequence). My thanks, as well, to Amanda Askell, Nick Beckstead, Paul Christiano, Tom Davidson, Carl Shulman, Bastian Stern, and Ben Weinstein-Raun for discussion.

Part I

Learning from the fact that you exist

1 Surprised I Am and ASS-backwards

Cases like God's extreme coin toss involve reasoning about hypotheses that specify both an objective world (e.g., a heads world with one person, or a tails world with a million), and a "location" of the "self" within that world (e.g., in the tails world, the "self" could be the person in the first room, the second room, etc). Call hypotheses of this form "centered worlds." The question is how to assign probabilities both to objective worlds and centered worlds, granted (a) some prior over objective worlds, (b) knowledge that you exist, and (c) your other knowledge about your situation. I'll call this broad topic "anthropics," though others might define the term differently.

A classic reference here is Bostrom (2002), which I'll be focusing on a lot—it's where I've spent most of my time. I'm going to be disagreeing with Bostrom quite a bit in this sequence, but I want to say up front that I think his book is great, and that it clarifies a lot of stuff. In fact, this whole sequence is very much "living in the world that Bostrom built," and a lot of the points I'm going to make are made by Bostrom himself—it's just that I'm making them with much more of a "this is why Bostrom's view is untenable" flavor.

SIA and SSA are two prominent approaches to anthropic reasoning (Bostrom favors a version of SSA, and dismisses SIA in a few short pages). Unfortunately, the names and standard glosses of these principles seem almost optimized for obscurity, and for many years, casual exposure left me unable to consistently remember which was which, or what they really meant. Katja Grace once suggested to me that partisans of SIA remember it as "Surprised I Am" (e.g., the view that updates on your own existence) and SSA as "ASS-backward" (e.g., the bad view). Another option would be to rename them entirely, but I won't attempt that here. For those familiar with the Sleeping Beauty problem, though, you can think of SIA as "thirding," and SSA as "halfing"—at least to a first approximation.

(Note: Bostrom presents SIA as an assumption you can add to SSA, yielding "SSA + SIA." This formulation ends up equivalent to my own, but I think it's worse, and I explain why in section 4. For now, I'll treat them as distinct and competing theories.)

How do SIA and SSA approach cases like God's extreme coin toss? Quantitatively: SIA updates the "prior" in proportion to the number of people in your epistemic situation in each objective world. SSA updates it in proportion to the fraction of the people-in-your-epistemic situation who are in the reference class, in that world. Then they both apportion their new credence on each objective world equally amongst the centered worlds (e.g., the hypotheses about "who you are") compatible with that objective world (e.g., among the people in that world you might be).

To see how this works, consider the following case:

> God's coin toss with equal numbers: God tosses a fair coin, and 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 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 (note from Joe: this is the type of "assuming X about the reference class" that you have to say *all the time* if you're SSA). 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.

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: "Wait, why are we updating the prior at all? Why play this game to begin with?" A key answer is: if you don't update the prior, and instead skip straight to apportioning your prior credence amongst the red-jacketed people in each world, you say silly things about this case. 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." But notice: 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.

(Bostrom actually ends up endorsing a version of SSA that fails to make this update—but that's not to its credit. I discuss this in part 2, section 9.)

2 Storytelling

SIA and SSA both get this "square one" right; but they differ in their verdicts in other cases. Before getting to those cases, though, can we say anything about what SIA and SSA are doing on a qualitative level? What is the "story" or "conception of the world" motivating these theories, and their differences?

It's actually pretty unclear in both cases. But, here's a shot at story-telling, which will hopefully illustrate how I, at least, tend to think about these views.

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 people-in-your-epistemic-situation from the "hat" of the platonic realm. That is, and put fancifully: before you were created with a red jacket in a white room, God said to himself "I need to create X number of people with red jackets in white rooms." He then reached into the platonic realm and groped around blindly 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, indeed, that God can draw as many people as he wants out, without altering the probability that he draws you). But yet, by the most (infinitely?) ridiculous luck, God's great fingers wrapped around your ghostly non-body. You got pulled, as the other red-jacketed souls looked on in awe and horror and jealousy and relief. 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. Or at least, that's the idea. (Notably, if the space of red-jacketed-white-roomed-people is infinite, then the probability that you get pulled by a finite world is zero, however finitely-many the pulls. And yes, SIA does imply certainty that you live in an infinite world. And yes, this is indeed a problem. See discussion in part 4, section 14.

To be clear: I don't especially like this story. And we can look for others, perhaps less exotic. 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. I discuss this conception more in part 4, section 15. However, I think it may run into instabilities, so I tend to stick with the story above.

Let's turn to SSA's story. Or at least, SSA's story as I tend to tell it. It's not a neutral rendition.

Like SIA, SSA learns something from the fact that you exist. In particular, SSA learns that you would've *necessarily* existed in any world that you can't currently rule out—e.g., any world with *anyone* in your epistemic situation. That is, granted that you *do* exist, SSA assumes that if God were going to create any world compatible with your current evidence, then He would have "gone looking for you" in the hat of possible people, then "inserted you" into that world—regardless of how many people it contains. He was, apparently, hell-bent on creating you, come what may, in all of the worlds you haven't yet figured out don't contain you—after all, you exist. It's a strange sort of relationship you have, you and God.

(Here I think the SSA-er says: "no, it's not like that. Rather, it's that given that I exist, then *if* any of those other worlds are real, *then* it's the case that I exist in those worlds. So I am licensed, in reasoning about which possible worlds are actual, in assuming that I get created in all of them." I discuss the dialectic here in a bit more detail in part 2, section 11.)

Importantly, though, on SSA, 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? No one knows. It's entirely made up. (I'll return to this problem later.) Still, on SSA, that's how God operates: he picks some set of people who in some sense "you could have been"—even though for some of them, you often know you *aren't*—and then makes one of them, at random, you.

Bostrom 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 I don't remember Bostrom offering an account. We can ask similar question about SIA, though, and the "hat of possible people" story I offered above isn't exactly an "oh of course no problems with that one."

To see where the reference class bit 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 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, to its great discredit, gives different answers depending on whether you count chimpanzees in the jungle as in your reference class or not. Thus, and using the language of the story above, it reasons: "Well, I know I exist, and I can't yet rule out heads or tails. So, regardless of whether the coin landed heads vs. tails, I was going to exist. (This is where SIA says: what? That's wrong.) 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 bacteria, an ant, a genetically engineered post-human, a brain emulation, a nano-bot, a paperclipping AI, a grabby alien, and so on. Indeed, as I'll discuss below in the context of the "Doomsday Argument," on SSA, the very future of humanity plausibly hinges 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 you "couldn't have 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.)

This chimp case may be the earliest and simplest result where I basically just get off the boat with SSA. I take one look at those chimps, and the question of whether they're in the reference class, and I feel like: "I'm out." But I don't necessarily expect others to feel the same way, and there's much more to say on either side regardless.

3 Can't we just use the minimal reference class?

Perhaps you're wondering, for example: can't 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 on the number of people in your epistemic situation at all, since the percentage of observers in your reference class who are in your epistemic situation will always be 100%. Thus, with a red jacket in God's coin toss with equal numbers above, it ends up at 50% on heads, and 50% on tails—even though on heads, only one person out of ten had a red jacket, but on tails, everyone did. In this sense, it starts reasoning like the "heads is always 50% no matter what I've learned about my jacket color" person—and it falls afoul of basic Bayesianism in the same way.

Indeed, a central problem motivating Bostrom is that he thinks that if you can't make updates like favoring tails in cases like God's coin toss with equal numbers, then you can't do science given the possibility of "big worlds"—that is, worlds where, for any given observation, there is some observer (for example, a Boltzmann brain) 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 make the observation in question—but the minimal reference class makes this impossible. That said, I haven't thought very much about Bostrom's "science in big worlds" considerations, and I don't think the argument against SSA-with-the-minimal-reference-class hinges on them. Regardless of the situation with Boltzmann brains, we should have the resources to favor tails in the "square one" case.

Note how elegantly SIA gets around this problem. SIA 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 some "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.

SSA, though, seems stuck with some sort of non-minimal reference class. Exactly how non-minimal is a further question—one that I'll return to in part 2, section 8.

4 Better and worse ways to understand SIA (or: how to actually stop using reference classes)

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

A bit of notation will be helpful. Let's call n the number of people in your epistemic situation, in a given objective world. And let's call r the number of people in your reference class, in that world. As I presented it, SIA updates the prior over objective worlds in proportion to n. SSA updates it proportion to n/r.

Now consider a different theory, which I'll call "Reference-class-SIA" (or R-SIA) and which corresponds more closely to one type of presentation in the literature. 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 n, R-SIA scales the prior in proportion to r.

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. Thus, tacking R-SIA onto SSA is sometimes said to "eliminate" the problematic dependence on the reference class that SSA otherwise implies: whatever reference class you choose, you get the same answer. And it is also said to "exactly cancel" some of 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. And after doing this extravagant dance, you're left with good ol' n (SIA).

But I think this framing undersells SIA's appeal. The appeal of SIA with respect to reference classes isn't that "you can pick whatever reference class you want." *It's that you don't have to think in terms of made-up reference classes at all.* Rather, you can just think entirely in terms of "people in your epistemic situation"—that is, in terms of n. Somehow, R-SIA + SSA feels to me like its ceding too much ground to SSA's narrative. It's living too much in SSA's weird, reference-classes-are-somehow-a-fundamental-thing-even-though-no-one-has-any-account-of-them world. It's trying to patch SSA with some extra band-aid, rather than rejecting it entirely.

Similarly, the appeal of SIA with respect to SSA's counterintuitive implications isn't that it adds just the right additional ridiculous update to counteract SSA's other ridiculous update. It's not that SIA lunges a million miles left, to cleverly (luckily? trickily?) 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. It just stays put, at home, where you always wanted to be. In this sense, SIA as I presented it above feels to me "simpler" than R-SIA + SSA—and I think the simple version captures better what reasoning using SIA actually feels like.

What's more, thinking in terms of R-SIA leads people to attach slogans to SIA that it doesn't strictly imply in practice. In particular, my sense is that people think of SIA as the view that favors worlds with more "observers"—and if you're using R-SIA with the reference class "observers," this is indeed a natural gloss. But if SIA as I presented it above doesn't actually care about observers per se (and neither does R-SIA, once you tack on SSA as well). Rather, it only cares about observers *in your epistemic situation*. You can try to sell me on a hypothesis that contains a zillion extra observers wearing blue jackets; but if I am wearing a red jacket, then on SIA, this feature of the hypothesis leaves me cold (though if it implies something about the number of red-jacketed people as well, or the number of people who could, for all I know, have been given red jackets, that's a different story). The same holds for bug-eyed aliens, chimps in the jungle, paper-clipping superintelligences, civilizations like our own on planets we can tell that we're not on, and all the rest of the cosmic zoo. SIA doesn't like observers; it likes uncertainty about "who/where I am." And we already know lots of stuff about ourselves.

That said, this consideration only goes so far. In particular, *if you don't know anything about yourself except that you're an observer*, then SIA does indeed like observers per se; and if you "forget" everything about yourself, then on SIA your credence in lots-of-observers-per-se worlds does indeed inflate. And more generally, the number of observers-per-se may correlate strongly with the number of observers in your epistemic situation, and/or the ones that could, for all you know, be in your epistemic situation, and hence be you (added 10/2: I say a bit more about the distinction between "people in your epistemic situation" and "people who, for all you know about about a given objective world, *might be* in your epistemic situation" here). Ultimately, though, it's the people-you-could-actually-be that SIA is really after.

Leaving R-SIA + SSA to the side, then, I'll focus on comparing SIA and SSA. Which theory is better?

Note that I say "better," not "true" or "best." These aren't the only approaches to anthropics, and given the various weird implications and uncertainties I'm about to discuss, it seems plausible that the true/best theory (is there are "true theory" of what your credence "should be"?) lies elsewhere (see discussion in part 4, section 16). Indeed, there's a whole literature on anthropics out there, which I haven't attempted to survey. Rather, I'm sticking to a comparison between two basic, prominent, first-pass views.

Indeed, really I'd prefer to ask a narrower question about these views. Not "which is better?", but "which is better *mostly in light of the considerations discussed in Bostrom (2002), plus a few other considerations that Joe encountered while writing this blog post?*". That is, I'm not, here, really attempting to exhaustively canvass all the relevant arguments and counterarguments (though I'm interested, readers, to hear which of the arguments I don't include you find most persuasive). Rather, I'm trying to report my (admittedly sometimes strong) inclinations after looking into the topic a bit and thinking about it.

All that said: SIA currently seems better to me. Part 2 and Part 4 of this sequence explain why. (Part 3 is a bit of an interlude.)

Part II

Telekinesis, reference classes, and other scandals

This post is the second in a four-part series, explaining why I think that one prominent approach to anthropic reasoning (the "Self-Indication Assumption" or "SIA") is better than another (the "Self-Sampling Assumption" or "SSA"). This part focuses on objections to SSA. In particular, SSA implies:

- scientific presumptuousness comparable to SIA's;
- ~certainty that fair coins that haven't yet been flipped will come up heads;
- expecting to be able to move boulders with your mind;
- a made-up and indeterminate ontology of reference classes that really seems like it shouldn't be a thing, and which also shouldn't be contorted willy-nilly to get whatever results you want (and also, this doesn't work anyway);
- sensitivity to differences that don't matter (like whether a definitely-not-you observer is killed and/or given evidence you don't get, vs. never created; or whether there are ten vs. twenty chimps in the jungle surrounding your thought experiment);
- a "special snowflake" metaphysics on which, because you actually exist, you had to exist in any world still compatible with your evidence (though I think that SSA has some replies here);
- intense updates towards solipsism.

5 The inevitability of presumptuousness

To get an initial flavor of some basic 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. 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. And at least in Bostrom's dialectic, the most prominent problem cases for each view basically amount to a restatement of this fact. (In particular, for those familiar with Bostrom, 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.) 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 trillion observers. According to T2, there are a trillion trillion trillion observers. The (non-anthropic) empirical evidence is indifferent between these theories, and the scientists are preparing to run a cheap experiment that will settle the question. However, a philosopher who accepts SIA argues that this experiment is not necessary, since T2 is a trillion times more likely to be correct.

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 an *empirical* science? What's this philosopher doing, weighing in on an empirical dispute with such confidence, on the basis of no evidence whatsoever save that she exists? Go back to your armchair, philosopher! Leave the science to the scientists!

Indeed, we can make the presumptuous philosopher look even more foolish. We can imagine, for example, 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. 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 writes, "is making a fool of [her]self" (p. 9).

For many people, this is close to the end of the train for SIA: presumptuousness of this kind (not to mention humiliation—and in front of the *scientists*, no less) is just too much to handle. And to be clear: 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) just as bad (and presumptuous, unscientific, humiliating, etc), and (b) worse.

Let's start with the nit-picks. First, the example as usually presented is phrased in terms of *observers* (I attribute this to the literature's focus on R-SIA discussed above): but as I discussed in part 1, to be relevant to SIA's ultimate verdicts, it needs to be phrased in terms of *people in your epistemic situation*. That is, it needs to be the case that on T2, there are trillion times more *candidates for people who might be you*. Suppose, for example, that the cosmologies in question work like this. In both cases, earth sits at the exact center of a giant, finite sphere of space. On T1, the sphere is smaller, and so has more non-earth 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 it's the same number of people-who-could-be-me either way. Thus, SIA follows the science: just do the experiment. For the case to work, then, the cosmologies in question need to be such that the extra observers *could be us*. Of course, lots of cosmologies are like this (just because we know we're on earth doesn't mean we know "where earth is" in an objective world), so you can in fact make versions of the case that work: hence the label "nit-pick." But it's a nit-pick that will become relevant below.

My second nit-pick is that pretty clearly, you shouldn't be 100% on a given theory of anthropics (see Carl Shulman's comment here). So while it's true that these sorts of credences are implied by SIA, they're not implied by a reasonable-person's epistemic relationship to SIA.

My third nit-pick is that I think it's at least a bit unfair, in a debate about the right credences to have in this scenario, to imagine the philosopher losing all these bets. That is, if SIA is right, then it's not the case that the non-anthropic empirical evidence is the sole relevant guide as to what will result from the experiment—the fact that you exist at all, in your epistemic situation, is also itself a massive update. Indeed, if we take this update seriously, then to even end up in a situation in which the non-anthropic empirical evidence favors T1 by a factor of a thousand seems like it might be positing something very weird having happened—something we might expect to induce the type of model-uncertainty I just mentioned. And more broadly, to SIA, imagining the philosopher losing these bets is similar to imagining someone betting hard against Bob winning the lottery, and losing twice in a row: by hypothesis, it almost certainly wouldn't happen. (That said, after the first loss, the model uncertainty thing comes into the lottery case as well: e.g., something fishy is going on with Bob...)

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

To see this, return to the version of the case just discussed, in which on both theories, earth is at the center of a giant sphere of space, but on T2, and the sphere and observer count are bigger. Let's say the non-anthropic empirical evidence, here, is 50-50, or a thousand to one in favor of T2, or whatever. As mentioned above, now SIA just follows the science. SSA, though, suddenly jumps into the role of presumptuous philosopher (or at least, it does if we use a reference class that includes the non-earth observers—more on epicycles that try to avoid this below). 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 (see Grace's "The Unpresumptuous Philosopher" for more on the parallels here).

Indeed: SSA, famously, can lead to the Doomsday Argument, which is structurally analogous to the case just given. Thus, suppose that you are considering 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-anthropic empirical evidence (for example, re: 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. (The usual doomsday argument appeals to your "birth rank," but I don't actually think this is necessary: it's just about the size of the reference classes). Whether this argument actually goes through in the real world, even conditional on SSA, is a much further question (it depends, in particular, on what reference class we use, and what other hypotheses are in play). But the bare possibility of making such an argument, on SSA, suggests that un-presumptuousness isn't exactly SSA's strong suit, either.

Is SIA's version of presumptuousness somehow worse? I don't see much reason to think so in principle. The core counter-intuitive thing, after all, was imagining philosophers (those sorry bastards—is it even a real field?) making extravagant bets with the sober cosmologists, on the basis of whatever the heck anthropics is supposed to be about. Surely, the intuition apparently goes, anthropics can't, as it were, be a thing. It can't, as it were, actually tell you stuff. Or at least not, you know, extreme stuff. Sure, maybe it can push you to be *a third* on heads, if God's coin toss with one vs. two people; or, like, *two-thirds*, if you're SSA and you know stuff about your jacket. But surely you can't just apply the same reasoning to a more extreme version of the case. That would mean you could make your credence on heads be *whatever*, just by adding more people/changing the jackets. That would be unfair. That could cause extreme credences on things, and endorsing epistemic principles that imply extreme credences in some cases is not, apparently, the philosopher's role—especially not in cases that are science-flavored, as opposed to people-in-boxes-with-jackets-flavored.

I'm writing somewhat in jest, here (though I do think that Bostrom's treatment of Sleeping Beauty suggests this kind of aversion towards extreme credences). But clearly, there is indeed something important to the intuition that the way to do cosmology, in either case, is to bet on the cosmologists, not against them. And to the extent that SIA or SSA would lead to actively bad empirical predictions, this seems like weighty (~decisive?) counter-evidence, whatever the theoretical virtues at stake.

For now, though, I'm left feeling like SIA and SSA are both presumptuous, including in cases that the scientists have opinions about. So presumptuousness, on its own, doesn't seem like a good desiderata.

6 Can you move boulders with your mind?

However: I also think that SSA's brand of presumptuousness is worse. In particular, it involves (a) betting not just against the scientists, but against the objective chances, and (b) it implies that a kind of telekinesis is possible. (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.)

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

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%—what with the fair coin thing, the not-having-been-tossed thing, and so on (if we want, for good measure, we can also make it a quantum coin—see Quantum Joe). 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 presumably (though betting in anthropics-ish cases can get complicated—see part 3), it bets, mortgages the house, etc accordingly.

The philosophy literature has a lot to say about when, exactly, ones credences should align with the objective chances, and I'm not going to try to unravel the issue here. What's more, one can imagine arguing that SIA, too, is weird 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). For now, though, I'll just note that SSA's verdict here seems like a really bad result to me—and in particular, a *worse* result than the Presumptuous Philosopher above. The Presumptuous Philosopher, to me, reads as a "can anthropics actually provide strong evidence about stuff?" type of case. Whereas the red-jacketed high-roller reads as more of a "can anthropics tell you that a fair coin, not yet flipped, is almost certain to land heads?" type of case. 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 front of you is a puppy. Next to you is a button that says "create a trillion more people." No else exists. A giant boulder is rolling inexorably towards the puppy. It's almost certainly going to crush the puppy. You have to save the puppy. But you can't reach it. How can you save it? You remember: you're SSA. You hold in your hands the awesome power of reference classes. 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?

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

One might think this a *very strange* expectation. Or more specifically: one might just think, point blank, that this type of move *won't work*. You can't move that boulder with your mind. That puppy is dead meat. But SSA expects the puppy to live. After all, if the puppy dies, then there will be a trillion extra people, and you would've been a tiny fraction of the reference class.

(Note that if we include puppies in the reference class, SSA also updates, upon waking up in this world, towards the puppy being an illusion—since if the puppy was real, then it was only 50% that you were you, instead of the puppy. And if we include the *boulder* in the reference class...)

We can also imagine versions where the boulder has either already crushed the puppy, or not, but you don't know which, and you make a commitment to press the button if you learn that the puppy is dead. This version combines the "telekinesis" vibe with a "backwards causation" vibe. That said, this blog isn't known for its strong stance against backwards causation vibes, so I'll focus on the forwards version.

Even if the forwards version, though, I can imagine protest: "Joe, I thought you were into zany acausal control stuff. Isn't this the same?" I don't think it's the same. In particular, I'm into acausal control *when it works*. If a roughly infallible predictor put a force-field around the puppy if and only if they predicted you were going to press the button, then by all means, press. My objection here is that I don't think SSA's move in this case is the type of thing that works. Or at least, I think that positing that it works is *substantially* more presumptuous than positing that anthropics can provide strong evidence about cosmology in general.

7 Does SIA imply telekinesis, too?

Now we might start to wonder, though: does SIA imply telekinesis, too? After all, SIA likes worlds with lots of people in your epistemic situation. Can't we use a button that makes lots of those people in particular to manipulate the world, telekinesis-style?

Sort of, but SIA's version of this, in my opinion, is less bad than SSA's version. Consider:

> 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 if and only if the boulder swerves. Having set up the machine, you prepare to enter a sealed white room. Should you expect the boulder to swerve, and the puppy to live?

Here, SIA still answers no. To see why, though, it helps to make a move that I expect basically all good anthropic theories will need to make—namely, to treat "you," for epistemic purposes, as a *person-moment*, rather than a *person-over-time*. After all, anthropics is about reasoning about the probabilistic relationship between objective worlds and centered-worlds, and centered-worlds can pick out *both* an agent *and* a time (hence, a person-moment) within an objective world. So a fully general theory needs to handle person-moments, too.

Thus, for example, the classic case of Sleeping Beauty 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, SIA answers 1/3, using reasoning that treats you as a person-moment instead of a person-over-time: after all, there are twice as many person-moments-in-your-epistemic-situation given tails than heads. And Bostrom, too, offers and accepts a reformulation of SSA that appeals to person-moments instead of people-over-time (he calls this reformulation the "Strong Self-Sampling Assumption," or "SSSA"). I'll generally ignore the differences between person-moments and persons-over-time versions of SIA and SSA, but my background assumption is that person-moments are more fundamental; and bringing this out explicitly can be helpful when thinking about cases like Save the puppy as SIA.

In particular: 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 "the boulder swerved" worlds 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. 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 she 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 a core SIA thing.

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 merge them into one person; 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 sunlight of this awful thought experiment should expect to see the puppy dead. (This, in my opinion, is also the thing to say about Yudkowsky's "Anthropic Trilemma.")

That said, it's true that, if you don't do any killing/merging etc, and instead let *ev-*

eryone 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 guy *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 *he* (or more specifically, his next person-moment) emerges into the daylight, he's 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-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, one sorry sucker 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: "dude, don't believe this SIA bullsh**: the puppy is 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 the original Joe—e.g., the original series of person-moments—rather than the copies, then it's true that you want to force a "puppy is dead" belief, because the original Joe ends up almost exclusively in "puppy is dead worlds." But this move has a side effect: it makes a trillion copies (or whatever) 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 willy-nilly. You've got broader epistemic responsibilities. Indeed, most of your "epistemic influence," if we weight by both probability *and* number of minds-influenced, is funneled towards the "puppy is alive" worlds. That said, once we're bringing questions about who you care about, and what sorts of pre-commitments (epistemic and otherwise) you want to make, we're getting into pretty gnarly territory, which I won't try to disentangle here (see part 3 for a bit more discussion).

For now, I'm happy to acknowledge that SIA isn't sitting entirely pretty with this sort of case. But I think SSA is sitting uglier. In particular, SSA *actively expects* this sort of "use the button to the save the puppy" thing to work. It will pay to get access to this sort of button; it will start calling in the "puppy saved!" parade even before it enters any kind of sealed-white-room. From SIA's perspective, by contrast, this sort of button-maneuver, and these sorts of sealed-white-rooms, are much less appealing. Exactly what type of not-appealing depends on factors like whether SIA cares about Joe-copies, but in general, even if in some cases SIA *ends up* expecting telekinesis to have worked, it will generally avoid, or at least not seek out, cases where it ends up with this belief. SSA, by contrast, believes in telekinesis ahead of time, and 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 wackier 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.

8 Against reference classes

Let's talk about reference classes. In particular, why they're bad (this section), and why using them to try to get out of the cases above is an un-Bayesian epicycle that doesn't work anyway (next section).

Why are reference classes bad? Well, for one thing, what even are they? What is the story about reference classes, such that they are a thing—and not just any old thing, but one sufficiently important as to warrant massive updates as to what sorts of world you're likely living in? SSA's toy story, as I've told it, is that the reference class is the set of beings in a given world such that God, dead set on creating you somehow (according to SSA), randomly "makes you one of them." But then, of course, SSA doesn't actually believe this in any literal sense. But what does SSA actually believe? What "way" does the world have to be, in order for SSA's reference class reasoning to make sense? What could even make it the case that the "true" reference class is one thing vs. another?

I have yet to hear such questions answered. To the contrary, as far as I can tell, 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 *contort* the reference class—sometimes, in my opinion, unsuccessfully—in pursuit of those answers. But we are left with very little sense of what constraints—if any—such contortions need, in principle, to obey.

In the absence of any such underlying metaphysical picture—or indeed, any non-mysterious characterization of reference classes more broadly—one could be forgiven for wondering whether the reference class could, as it were, 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* ("'There was,' Miller admits reluctantly, 'one animatronic pig'"), five bug-eyed aliens 10^{100} light-years away, and a King of France who never existed. When God created this world, he made "me" one of these creatures at random (the relevant King of France happened to not be present in this world). Probably, I was going to be a pig. (In fact, given that I'm Joe, this is evidence that Babe actually involved fewer than 47 pigs...).

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 "God's coin-toss with equal numbers" type cases, and it violates conditionalization to boot.

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. I'll say more about why this doesn't work below. Indeed, my current take is that if you're *going* to go for SSA, you should go for 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, especially if you're not a deep realist about "observers" (though SIA has somewhat related problems about "counting people-in-your-epistemic-situation"). Beyond this, though, if we're really trying to be maximal, we might wonder: why stop with observer-like things? Why not, for example, throw in some unconscious/inanimate things too? Sure, I know that I'm an observer-like thing. But the whole point of reference classes is to include things I know I'm not. So why not include rocks, galaxies, electrons? Why not the composite object consisting of the moon and my nose? Why not, for that matter, abstract objects, like the natural numbers? Viewed in this light, "things" seems a more maximal reference class than "observer moments" (and perhaps "things" is itself less-than-fully maximal; do the things have to "exist"? Can merely possible things count? What about impossible things?). And if "observer-moments" turns out to be less-than-fully maximal, it loses some of its non-arbitrariness appeal (though perhaps there's some way of salvaging this appeal—I do think that "observer-moments" is intuitively a more natural reference class than "things." Maybe we say something about: the "things" you don't rule out once you realize that you exist and are asking questions? But why that?).

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, with some philosophers, to some notion of metaphysical "essence." Thus, we might say, you *couldn't* have been a pig, or an alien, or an electron; perhaps, even, you couldn't have been someone with different genes. And if you *couldn't* have been something, then perhaps God couldn't have randomly made you that type of thing, either. Indeed, my sense is that sometimes, the notion of "reference classes" is construed in some vaguely-reminiscent-of-metaphysical-essences kind of way (e.g., "but you *couldn't* have been an electron; you're an observer!"), even absent any kind of explicit account of the concept at stake.

But do we *really* want to bring in stuff about metaphysical essences, here? *Really*, SSA? Bostrom, at least, seems keen to distance himself from this sort of discourse; and I am inclined to heartily agree. And once we start making cosmological predictions on the basis of whether Saul Kripke would grant that I "could've" been a brain emulation, one starts to wonder even more about presumptuousness.

Are there other non-arbitrary reference options, between minimal and maximal? Maybe: humans? But...why? Do they need to be biological? Can they be enhanced? How much? Why or why not? Why not say: creatures in the genus homo? Why not: primates? Why not: intelligences-at-roughly-human-levels? Why not: people-with-roughly-Joe's-values? Why not: people-with-Joe's-DNA-in-particular? I've yet to hear any answers, here. Indeed, as far as I can tell, we're basically in the land of just entirely making up whatever we want, subject to whatever constraints on e.g. simplicity, vaguely-intuitiveness, etc that we have the philosophical decency to impose on ourselves. The discourse, that is, is totally untethered. And no surprise: it never had a tether. We never knew what we were trying to talk about in the first place.

What's more, this untethered quality has real effects on our ability to actually 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 whether you count chimps in the jungle as in the reference class or not. Indeed, reading over Katja Grace's overview of her attempt apply SIA and SSA to reasoning about the Great Filter, I was struck by the contrast been SIA's comparatively crisp verdicts ("SIA increases expectations of larger future filter steps because it favours smaller past filter steps"), vs. the SSA's greater muddle ("SSA can give a variety of results according to reference class choice. Generally it directly increases expectations of both larger future filter steps and smaller past filter steps, but only for those steps between stages of development that are at least partially included in the reference class.").

One of Bostrom's main responses to all of 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 Bayesian 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: e.g., 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. In particular, I think that seeing reference classes as a mysterious subjective object like the "prior" does put them in somewhat more respectable company. And indeed, some implications of the subjectivity at stake are similar: for example, just as agents with different priors can continue to disagree after sharing all their information, so too can agents with different reference classes, but the same priors. (Which isn't to say this is a good result; it's not. But it establishes more kinship with the prior.) Still, though, I think we should view introducing yet another mysterious subjective object of this kind as a disadvantage to a theory—especially when we can't really give an account of what it's supposed to represent.

At heart, I think my true rejection of reference classes might just be that they feel janky and made-up. When I look at the GOD'S COIN TOSS WITH CHIMPANZEES; when I find myself having to say "of course, if there are ten other people *watching* Sleeping Beauty's experiment, then depending on whether they're in the reference class, and how many person-moments they've had, Beauty's credence should actually be X; but let's bracket that for now..."; when I find myself without any sense of what I'm actually trying to talk about; I have some feeling like: Bleh. Yuck. This is silliness. Someone I know once said, of SSA, something like: "this is repugnant to good philosophical taste." I've found that this characterization has stuck with me, especially with respect to reference classes in between minimal and maximal. When forced to talk about such reference classes, I feel some visceral sense of: ugh, this is terrible, let's get out of here. SIA is sweet relief.

9 Against redraw-the-reference-class epicycles that don't work anyway

There's a particular 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 you credence be on heads?
>>
>> Light+Red: God keeps the lights in all the rooms on. You wake up and see that you have a red jacket. What should your credence be on heads?

In Darkness and Light + Red, SIA and SSA (respectively) each give extreme verdicts about the toss of a fair coin. These examples served as the templates for other putatively problematic implications of SIA (the Presumptuous Philosopher) and SSA (e.g., the Doomsday Argument, Red-Jacketed High-Roller, Save the Puppy). Bostrom hopes to avoid them both. That is, he hopes to thread some sort of weird needle, which will allow him to be 50% on heads in Darkness, and 50% on head 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. Heads will always be 50%; the scientists will always be right; the puppy will always die; and the world will be safe from anthropics—at least, for now.

How can we reach such a happy state? As far as I can tell, the idea is: define the reference classes such that you get this result. (See Bostrom (2002), p. 167, and p. 171-2 for fairly explicit comments about this intention.) In particular: claim 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 standard Bayesianism would imply, you "start over" with the whole SSA game, but this time, with a new and improved reference class—a reference class that allows you not think it was unlikely, conditional on tails, that you ended up with a red jacket. After all, on this new reference you 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? Not in my book. 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, you do this thing where your credence on "I'm in a tails world" ends up changing. Bostrom does a dance, here, about how no, really, his model is (or at least, can be, if you want it to be) kosher Bayes after all, because 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 and I know it"). I haven't tried to engage with this dance in detail, but my current take is: I bet it doesn't work. In particular, my suspicion is that Bostrom's treatment is going to throw the doors wide open for person-moments to reason in very unconstrained ways even in non-anthropics-y cases (see e.g. Grace's discussion here); and that more generally, Bostrom is going to end up treating the type of Bayesian reasoning that you should be doing in this sort of case as more different from normal reasoning than it should be.

My higher-level objection, though, is that it seems pretty clear that Bostrom is making this move specifically in order to give a certain set of answers in a certain set of otherwise problematic cases, and that he would have little interest in it otherwise. Indeed, he frames this move as in some sense "optional"—something you can, as it were, 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.

Of course, contorting your fundamental principles to curve-fit a specific (and often artificially-limited) batch of cases, with little regard for other theoretical desiderata, is the bread and butter of a certain type of philosophical methodology. But that's not to the field's credit. Indeed, plausibly such a methodology, for all its charms, often sends the philosophers astray—and I expect that trying to use it to say 50% in both Darkness and Light + Red will lead us astray here. At the very least, Bostrom's version sets off a lot of alarm bells, for me, about over-fitting, epicycles, and the like. And it makes me wonder, as well, about what sorts of limits—if any—are meant to apply to 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 problems similar to those that the minimal reference class does (my thanks to Bastian Stern suggesting this point in conversation). To see this, consider a version of God's coin toss with equal numbers:

> God's coin toss with equal 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 really 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. Indeed, Bostrom thinks that if you can't say things like that, you can't do science in big worlds.

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

Indeed, I feel confused by Bostrom's treatment of this issue. After introducing his treatment of the original Light + Red on p. 165 of the book, 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 surely Bostrom 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 are basically the same! The only difference is the extra red-jacketed people in heads! Using different reference classes in the two cases would be just...too much. At that point, we should just throw in the towel. We should just say: "the reference class is whatever the heck I need to say it is in order to have the credence I want, which in this particular case is, let me check...50%."

To be clear, I don't actually think that Bostrom would endorse using different reference classes in these two cases. But as far as I can tell, his discussion in the book implies this, and makes it necessary. Maybe I've misunderstood something, or missed some other discussion of the issue elsewhere?

Moving beyond Bostrom in particular: my suspicion is that something in the vicinity of these objections is going to apply, in general, to attempts to contort the reference classes to avoid SSA's problematic implications in cases like Save the Puppy (especially to avoid them *in principle*, as opposed to in some particularly putatively real-world application). Thus, to avoid telekinesis in Save the Puppy, my sense is that you'll have to do something un-Bayesian (e.g., not update when you learn that you are the single, pre-boulder squishing/swerving person, rather than one of the possible people created by the button in the no-swerve worlds), epicyclic (it's going to seem like: what? why?), and in tension with what one would want to say in a nearby, equal-numbers version (though maybe it's harder to find equal-numbers versions of Save the Puppy? I haven't thought about it much.)

10 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:

> 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 all the red jacketed people. If tails, he kills all the blue jacketed people. 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?

(This is a condensed version of an argument from Stuart Armstrong; see also a closely-related version in Dorr (2002), and a related series of cases in Arntzenius (2003)).

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 (or at least, it does if you don't try any of Bostrom's reference-class shenanigans above). 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 goes all SIA-ish. In particular, when it first wakes up, but it hasn't yet heard or not heard the bell, it updates against having a red jacket, to 10%: after all, it's an equal-numbers case, and most people have blue jackets. Then, because the chance of death is 50% conditional on either having a blue jacket, or a red jacket, it stays at 10% after hearing the bell: survival is no update.

But are these cases actually importantly different? Armstrong (at least, circa 2009; he's since changed his view, for decision-theory reasons) doesn't think so, and I'm inclined to agree. And note that we can construct a kind of "spectrum" of cases leading from the first case to the second, where it seems quite unclear what would constitute an epistemically-relevant dividing line (see Armstrong's post for more).

Dorr makes a similar argument in *Sleeping Beauty*. Consider a version where Beauty is woken up on both Monday and Tuesday conditional on both heads and tails, but then, if it's heads and Tuesday, she hears a bell after an hour or so. Surely, argues Dorr, Beauty ought to be 50-50 on heads vs. tails prior to hearing-the-bell-or-not, and 25% on each of

Heads-Monday, Heads-Tuesday, Tails-Monday, and Tails-Tuesday. Then, after she *doesn't* hear the bell, surely she should cross off "Heads-and-Tuesday," re-normalize, and end up at 1/3rd on heads like a reasonable SIA-er. And indeed, this is what SSA *does* do (unless, of course, we futz with the reference classes), *if* Beauty is also woken up in "Heads-and-Tuesday" and can hear this type of bell. But if Beauty *isn't* woken up in Heads-and-Tuesday at all, then suddenly SSA is back to halfing. Does this difference matter? It really seems like: no.

What we're seeing in these cases is basically SSA's "sensitivity of outsiders," made especially vivid and counter-intuitive. 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 funky 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. That said, I've griped about reference classes quite a bit already, and I'm not actually sure that the "what's up with the relevance of these outsiders?" objection actually adds much to the "what's up with reference classes in general?" objection (though it definitely prompts in me some sense of: "*man* this janky").

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 actually get pretty close to 1/3-ing: if heads, you had a 1/~1M chance of being in a white room rather than the jungle, and if tails, you had a 2/~1M chance of this, so finding yourself existing in a white room is actually a ~2:1 update in favor of tails. SSA-ers might try to use similar "appeals to outsiders" to try to avoid saying bad things about the doomsday argument. Thus, if there are (finite) tons of observers and they're all in the reference class, the difference between DOOM SOON and DOOM LATER does less to the fraction of people-in-your-reference-class you are.

I think moves like this might well help to alleviate some of SSA's bad results in real-world cases (though we'd have to actually work out the details, and no surprise if they get gnarly). 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 suddenly 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 futzing with the number of outsiders (or hoping for the right number) feel pretty hacky to me, and not in the original spirit of the view. And regardless, SSA's bad results in cleaner, more thought-experimental cases (e.g., Save the Puppy) will persist.

11 SSA's metaphysical mistake

I've given a lot of specific counter-examples and counter-arguments to SSA. But I also want to talk about where it feels, at least from SIA's perspective, like SSA goes wrong from the get-go: namely, it assumes that you exist no matter what, in any world epistemically-compatible with your existence. This is a core shtick for SSA. It's what allows SSA to not update on the fact that you exist. But at least when viewed in certain light, it doesn't really make sense (perhaps other light is more flattering).

To see what I mean, return to a basic version of God's coin toss, where God creates one person if heads, and a million if tails, all in white rooms. Suppose that the coin has in fact landed tails. You are Bob, one of the million people God has created, and you're wondering whether the coin landed heads, or tails. As a good SSA-er, you basically reason: "well, I exist. So if it landed heads, I'd exist; and if it landed tails, I'd exist. So: 50-50." But now consider Alice, in the next room over. She, too, is an SSA-er. So she, too, reasons the same.

But notice: *Bob and Alice can't both be right*. In particular, Bob is treating the heads world like it would necessarily create *him*; and Alice is doing the same; but there ain't room enough in the heads world for both (thanks to Katja Grace for suggesting this framing). And indeed, we can specify that, had the coin landed heads, the person who would've been created is in fact not Bob *or* Alice but rather *Cheryl* of all people. And is that so surprising? What mechanism was supposed to guarantee that it would be Bob, or Alice, or any other particular anthropic-reasoner? The mere fact that Bob and Alice found themselves existing in the actual world, and thus were able to wonder about the question? But why would that matter?

This isn't necessarily a tight argument (indeed, I discuss some possible replies below). But I'm trying to point at some kind of "why reason like that?" energy I can get into in relation to SSA. Maybe this example can make it vivid.

My sense is that for Bostrom, at least, the story here is supposed to be hinge centrally on the fact that if you hadn't existed, you wouldn't be around to observe your non-existence (see, e.g., his discussion on p. 125). But why, exactly, would this fact license assuming, granted that you *do* exist, that you would've existed no matter what (at least in worlds you can't currently rule out)? Here I think of classic examples reminiscent of Armstrong's case above. Suppose you're one of a hundred people in white rooms. God is going to kill ninety-nine of you, if heads, or one, if tails, then ring his bell either way. His bell rings. If you had been killed, you wouldn't have been around to hear it. Does this mean you were guaranteed to survive? Does this mean you shouldn't update towards tails? No. So what's the story? Why is never-having-been created different from getting killed?

In general, it can feel to me like, because you happen to exist, SSA treats you like some sort of special snow-flake person—some sort of privileged ball that God must've gone "fishing for" in the urn, since after all, it got drawn. Or perhaps, on a different framing, SSA treats you like a kind of "ghostly observer," who has learned, from the fact that it exists and is making observations, that it "would've been someone" in any world, and the only question is who. On this framing, it's not that as Bob, you should assume that God would've created *Bob* in the heads world. Rather, God could've created anybody he wanted: but that person would've been *you* regardless—e.g., the ghostly observer would inhabit a different body. That is, in this case, had the coin landed heads, God may well have created Cheryl; but "you," in that case, would've been Cheryl. (And presumably, the same would be true of Alice? You'd *both* have been Cheryl? Or something?)

Indeed, maybe the most SSA-ish story is something like: look, there's a world spirit. You're it. We're all it. The world spirit, um, experiences a random sample of all the observer-moments in the world, no matter how many of them God creates. Thus, if God created just Cheryl, you'd be experiencing Cheryl. If God created Alice, Bob, etc, you'd be experiencing one of them or other. Thus, you'd be experiencing someone either way, and you shouldn't update from the fact that you're experiencing anything at all. (However, if you learn that you're experiencing Cheryl in a land of chimps, you should update towards the chimps being illusions.) I doubt people will want to put things in these terms, but I think that this picture would in fact make sense of SSA's reasoning.

That said, I think that SSA-ers have other options/replies available here as well. In particular, I think that SSA can say something like: "look, I do in fact exist. Thus, if any of these epistemically-possible worlds are *actual*, then they do in fact contain me. So, it makes sense, in considering what credence to put on these worlds, to condition on my having been created in them—since if they were actual, I would've been." This sort of line does have its own pull, and I think really running to ground some of the differences here might get tricky. In particular it looks like there's some semantic difference re: "would I have existed if e.g. the coin had come up heads." The thing I specified re: Cheryl was that on a counterfactual "if," the answer (when Bob asks) is no. But the SSA-er presumably wants to say that a different "if" is relevant, one more akin to "if the coin came up heads, do I exist?"—and I don't currently have an especially strong opinion about where debate about the "ifs" here will go.

Overall then, I think it's probably best to construe the arguments in this section centrally as "here's a way that someone in an SIA-like mindset can end up looking at SSA and saying: what?" Really pinning down the dialectic would take further work.

12 SSA's solipsism

One last dig at SSA: it loves solipsism. If you were the only thing that exists, it would be so likely that you are you. Like, 100%. Compared to these hypotheses where there are all these other people (8 billion of them? 100 billion throughout history? More to come? Come on. Don't be ridiculous.), and you just *happen* to be you, SSA thinks that solipsism look *great*. 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. And weren't you way more than one-in-a-hundred-billion on solipsism anyway? Don't you remember

Descartes? How did you really know those other people existed in the first place? It would look the same regardless, you know. And don't even get me started on the idea that *animals* are conscious, or that aliens exist. Please.

In fact, while we're at it, what's all this about your memories? That sounds like some "other-person-moments in the reference class" bullsh**to me. How many of them did you say there were? What's that? We never defined any sort of temporal duration for a person-moment because obviously that's going to be a silly discourse, but apparently we're going to use the concept anyway and hope the issue never makes a difference? Hmm. Sounds suspicious to me. And sounds like the type of thing that would make it less likely that you were having these experiences in particular. Best to just do without. That 13th birthday party: never happened. And obviously your future, too, is out the window.

I jest, here, but it's a real dynamic. Just as SIA loves big worlds, if you don't know you are, SSA loves small worlds, if you do. And the solipsist's world is the smallest of all.

Part III

An aside on betting in anthropics

This post is the third in a four-part sequence, explaining why I think that one prominent approach to anthropic reasoning (the "Self-Indication Assumption" or "SIA") is better than another (the "Self-Sampling Assumption" or "SSA"). This part briefly discusses betting in anthropics. In particular: why it's so gnarly, why I'm not focusing on it, and why I don't think it's the only desiderata. If you're not interested in betting-type arguments, feel free to skip to part 4.

I've now covered my main objections to SSA. In part 4, I say more in defense of SIA in particular. Before doing so, though, I want to mention a whole category of argument that I've generally avoided in this post: that is, arguments about what sorts of anthropic theories will lead to the right patterns of *betting behavior*.

I expect to some readers, this will seem a glaring omission. What's the use of talking about credences, if we're not talking about betting? What are credences, even, if not "the things you bet with"? Indeed, for some people, the question of "which anthropic theory will get me the most utility when applied" will seem the only question worth asking in this context, and they will have developed their own views about anthropics centrally with this consideration in mind. Why, then, aren't I putting it front and center?

Basically, because questions about betting in anthropics get gnarly really fast. I'm hoping to write about them at some point, but this series of posts in particular is already super long. That said, I also don't think that questions about betting are the only desiderata. Let me explain.

Why is betting in anthropics gnarly? At a high-level, it's 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're EDT-ish or CDT-ish, your level of altruism towards copies of yourself/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 various "act as you would've pre-committed to acting from some prior epistemic position" type moves (e.g. "updatelessness")—either at the level of choices (whether your own choices, or those of some group), or at the level of epistemology itself. Anthropics-ish cases tend to implicate these issues to an unusual degree, and in combination, they end up as a lot of variables to hold in your head at once. Indeed, there is some temptation to moosh them together, egg-ed on by their intertwined implications. But they are, I think, importantly separable.

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

But this argument doesn't work if Beauty's person-moments are EDT-ish (and 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 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 betting behavior doesn't align with your credences. Is that surprising? Sort of. But in general, if you're going to take a bet different numbers of times conditional on outcome vs. another, the relationship between the 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. (See Arntzenius (2002) for more.)

Indeed, the EDT-ish *thirder*, here, actually ends up betting like a *fifth*-er. That is, if offered a "win twenty if heads, lose ten if tails" bet upon each waking, she 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 (thanks to Paul Christiano and Katja Grace for discussion). That is, every (non-updateless) altruistic EDT-er is a fifth-er sometimes.

(Or at least, this holds if you use the version of EDT I am most naively attempted by. Paul Christiano has recently argued to me that you should instead use a version of EDT where, instead of updating on your observations in the way I would've thought normal, then picking the action with the highest EV, you instead just pick the action that has the highest EV-*conditional-on-being-performed-in-response-to-your-observations*, and leave the traditional notion of "I have a probability distribution that I update as I move through the world" to the side. I haven't dug in on this, though: my sense is that the next steps in the dialectic involve asking questions like "why be a Bayesian at all.")

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. And we'll also need to start asking questions about what sort of pre-commitments it would've make sense to have made, from what sorts of epistemic/cooperative positions, and what the implications of that are or should be. I think questions like this are well worth asking. But I don't really want to get into them here.

What's more, I don't think that they are the *only* questions. In particular, to me it seems pretty possible to separate the question of how to bet from the question of what to believe. Thus, for example, in the EDT-ish halfer case above, it seems reasonable to me to imagine thinking: "I'm 50% on heads, here, but if it's tails, then it's not just me taking this 'win $10 if heads, lose $10 if tails' bet; it's also another copy of me, whose interests I care about. Thus, I will demand $20 if heads instead." You can reason like that, and then step out of your room and continue to expect to see a heads-up coin with the same confidence you normally do after you flip. Maybe this is in some sense the wrong sort of expectation, but I don't think your betting behavior, on its own, establishes this.

(One response here is: "you're mistakenly thinking that you can bet for two, but expect for one. But actually, you're expecting for both, too. And don't you care about the accuracy of your copy's beliefs too? And what is expecting if not a bet? What about the tiny bit of pleasure or pain you'll experience upon calling it for tails as you step out of the room? Don't you want that for both copies?". Maybe, maybe.)

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 behind some veil of ignorance have pre-committed to believing/acting-like-I-believe?", or something like that. Those are (or, might be) important questions too. But sometimes you're just, as it were, curious about the truth. And more generally, in many cases, you can't actually 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? Does the fact that we exist mean that there are probably lots of simulations of us? Can we use earth's evolutionary history as evidence for the frequency of intelligent life? 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.

Indeed, I've been a bit surprised by the extent to which some people writing about anthropics seem interested in adjusting (contorting?) their epistemology per se in order to bet a particular way—instead of just, you know, betting that way. This seems especially salient to me in the context of discussions about dynamical inconsistencies between the policy you'd want to adopt *ex ante*, and your behavior *ex post* (see, e.g., here). As I discussed in my last post, these cases are common outside of anthropics, too, and "believe whatever you have to in order to do the right thing" doesn't seem the most immediately attractive solution. Thus, for example, if you arrive in a Newcomb's case with transparent boxes, and you want to one-box anyway, the thing to do, I suspect, is not to adopt whatever epistemic principles will get you to believe that the one box is opaque. The thing to do is to one-box. (Indeed, part of the attraction of updateless-ish decision theories is that they eliminate the need for epistemic distortion of this kind.) I expect that the thing to say about various "but that anthropic principle results in actions you'd want to pre-commit to not taking" (for example, in cases of "fifth-ing" above) is similar.

Still, I suspect that some will take issue with the idea that we can draw any kind of meaningful line between credences and bets. And some will think it doesn't matter: you can describe the same behavior in multiple ways. Indeed, one possible response to cases like God's coin toss is to kind of *abandon* the notion of "credences" in the context of anthropics (and maybe in general?), and to just act as you would've pre-committed to acting from the perspective of some pre-anthropic-update prior, where the right commitment to make will depend on your values (I associate this broad approach with Armstrong's "Anthropic Decision Theory," though his particular set-up involves more structure, and I haven't dug into it in detail). This is pretty similar in spirit to just saying "use some updateless-type decision theory," but it involves a more explicit punting on/denial of the notion of "probabilities" as even-a-thing-at-all—at least in the context of questions like which person-in-your-epistemic-situation you are. That is, as I understand it, you're not saying the equivalent of "I'm ~100% that both boxes are full, but I'm one-boxing anyway." Rather, you're saying the equivalent of: "all this talk about 'what do you think is in the boxes' is really a way of re-describing whether or not you one-box. Or at least, it's not important/interesting. What's key is what you *do*: and I, for one, one-box, because from some epistemic perspective, I would've committed to doing so."

This approach has merits. In particular, it puts the focus directly on action, and it allows you to reason centrally from the perspective of the pre-anthropic-update prior (even if you eventually end up acting like an SIA-er, a Presumptuous Philosopher, and so on), which can feel like a relief. Personally, though, I currently prefer to keep the distinctions between e.g. credences, values, and decision-theories alive and available—partly to stay alert to implications and subtleties one would miss if you mooshed them together and just talked about e.g. "policies," and partly because, as just discussed, they just seem like different things to me (e.g., "do I believe this dog exists" is not the same as "do I love this dog"). And I also have some worry that the "pre-anthropic update prior" invoked by this approach is going to end up problematic in the same way that the prior becomes problematic for updateless decision theories in general. (E.g., how do you know what pre-commitments to make, if you don't have credences? Which credences should we use? What if the epistemic perspective from which you're making/evaluating your pre-commitment implicates anthropic questions, too?)

At some point, I do want to get more clarity about the betting stuff, here—obviously, it's where rubber ultimately meets road. For now, though, let's move on.

Part IV

In defense of the presumptuous philosopher

This post is the last in a four-part sequence, explaining why I think that one prominent approach to anthropic reasoning (the "Self-Indication Assumption" or "SIA") is better than another (the "Self-Sampling Assumption" or "SSA"). This part discusses some prominent objections to SIA. In particular:

- *I suggest that the "Presumptuous Philosopher," a canonical counter-example to SIA, is a strong candidate for what I call a "good bullet": that is, a counterintuitive result, acceptance of which resolves a lot of gnarly issues into a simple and pretty satisfying theory, and rejection of which invites endless complication and counterexample (fans of the repugnant conclusion, take note). This doesn't mean we should actually bite. But we shouldn't die on the hill of non-biting.*
- *Pascal's muggings and infinities are problems for SIA, but this puts SIA in pretty respectable company (expected utility theory, population ethics)—and in particular, company that still seems to make itself useful.*
- *SIA has problems with "counting observers." I haven't thought that much about this one, but I have some feeling like: don't we all?*
- *Given some values and decision theories, SIA (like SSA) suffers from inconsistencies between "the policy you'd want to commit to, from some 'prior' epistemic perspective" and "how you behave ex post." But these inconsistencies are common in normal life, too; and if you're worried about them, you can use an "updateless" decision theory. But I suggest not throwing out the concept of epistemology along the way.*
- *We can maybe do a bit to make SIA more intuitive (though not as much as I'd like).*

That said, even if SSA is worse than SIA, it's not like SIA is sitting pretty (I especially don't like how it breaks in infinite cases, and there are presumably many other objections I'm not considering). I briefly discuss whether we should expect to find a better alternative ("Anthropic Theory

X"). My current answer is: maybe (and maybe it's already out there), 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 close by quickly mentioning some of SIA's possible implications in the real world, re: doomsday arguments, simulations, and multiverses. I think we should tread carefully, here, but stay curious.

13 Good bullets

A lot of this sequence so far has been about objecting to SSA. So let's return, now, to SIA's problems: and in particular, the Presumptuous Philosopher.

I granted, earlier, that the Presumptuous Philosopher is a bad result (I'll also discuss some more extreme versions of this result—for example, in infinite cases—in the next section). 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. In particular, people in your epistemic situation seem so...cheap. You can just 2x them, a trillion-trillion-x them, with a brief waggle of the tongue. Must our credences swing, wildly, in proportion to such whimsy?

I'd like to think that on the actually-good theory of anthropics, combined with our knowledge of everything else, this doesn't happen so easily in the real world (but that it probably does happen in God's coin toss type cases, as e.g. thirding in Sleeping Beauty seems to require). But I don't have the actually-good theory of anthropics. Rather, what I have, so far, is SIA and SSA: and SSA, as I've discussed, seems to me pretty bad. What, then, should we do? And in particular, if SIA and SSA are our only options (they aren't, but see section 16 for more), how can we live with ourselves?

I find it helpful, in such an unpleasant dialectical situation, to bring to mind the concept of a "good bullet," which I will illustrate with an analogy to population ethics.

Once upon a time, there was a conclusion called the repugnant conclusion. This conclusion said that if you add zillions of only-a-bit-good lives in a population, then with only a brief waggle of the tongue, the goodness of that population can swing wildly—enough, indeed, to outweigh anything else. Such a conclusion was implied by a very natural, simple, intuitive theory of the goodness of populations (e.g., add up the goodness of the lives involved); it was supported by a number of extremely strong abstract arguments, based on seemingly undeniable principles; and the most immediate alternative theories immediately faced much *worse* counterexamples to boot. Indeed, there were even *proofs* that if you were going to deny the repugnant conclusion (or indeed, a weakened version of it), you would necessarily say something else very counterintuitive.

Few philosophers were actively excited about the repugnant conclusion. But some would stop at nothing to avoid it. Increasingly desperate, this latter group started saying crazy things left and right. They started putting people in hell to avoid making extra happy people who aren't happy *enough*. They started denying that betterness is a consistent ordering. They started talking a lot about indeterminacy and incomparability and incommensurability. They started talking about how maybe ethics is impossible, maybe it's no use making choices, maybe caring about stuff in a coherent way is a hopeless endeavor. Some went to their very graves, after decades, still writing papers on the topic, still searching for some way out. They had picked, it seemed, a bullet they would never bite, an immovable rock on which all else must be built, and thus, around which all else must contort—how garishly, whatever the costs.

To others, though, the situation was different. They felt like: "wait, so if I bite this one bullet that maybe isn't even that bad from a certain perspective, and which seems like a natural and direct extension of reasoning I accept in other circumstances, then I get in return a nice simple intuitive theory, rather than one of these much-worse alternatives? This is sounding like a pretty good deal." On this view, it's not that the bullet, when you bite it, tastes *good*: it doesn't. But if you bite it—or at least, if you allow yourself to *consider* the *possibility* of biting it—you're suddenly allowed to see clearly again. Things come back into focus. The craziness is over. You can rest. (At least, for now.) In this sense, even if the bullet's taste is bitter—even if you remain open to it being wrong, and interested in genuinely-more-attractive alternatives—it's a "good bullet" nonetheless.

I'm not going to take a stand, here, about whether this "good bullet" conception of the repugnant conclusion is right. But I want to flag the possibility that we should be telling a similar story about the Presumptuous Philosopher.

One reason to suspect this is that Bostrom seems to treat the Presumptuous Philosopher as basically the end-of-discussion-objection to SIA, and my sense is that others do as well. Thus, the dialectic goes, "Of course, we could solve this ridiculous problem if we just used SIA. But: the Presumptuous Philosopher. Ok, back to SSA." (OK, this is a bit of a caricature. But I do get a "this is the central and canonical counter-example to SIA" vibe from the Bostromian literature.) And you've been dismissing an otherwise-attractive theory centrally in virtue of a single case...

Another reason to suspect that the Presumptuous Philosopher is a good bullet is that when we try to avoid it, at least via SSA (and see section 16 for more pessimism), we get all the crazy contortions, epicycles, boulder-swervings, reference-class indeterminacies, solipsisms, metaphysical mistakes, and so forth discussed above. I do feel some pattern matching to the population ethics literature, here.

A final reason to suspect that the Presumptuous Philosopher is a good bullet that it's basically just a restatement of the verdicts that we (or at least, I) *want* from SIA (e.g., thirding in Sleeping Beauty), and it follows from pretty straightforward and compelling arguments, like the Dorr-Arntzenius argument above (e.g., you should be 1/4 on everything when you're woken up twice regardless; so what happens when you learn it's not Heads-Tuesday?). I discuss this a bit more in section 16 below.

Indeed, to stretch the analogy with the repugnant conclusion further, I think there are actually some close spiritual similarities between totalism in population ethics and SIA in anthropics (thanks to Katja Grace for suggesting this). In particular, both of them—at least seen in a certain light—involve a central focus on possible people. That is, totalism cares not just about the people who actually live, but the people who *don't get live because* *you didn't create them*. Similarly, SIA cares, not just about the fact that you're actually alive, but the fact that you *might not have been*: it treats you as a possible person who happened to become actual, rather than as someone who would've been actual no matter what. And one lesson the totalist takes away from population ethics is that possible people tend to get short shrift in actual people's intuitive consideration. Perhaps something similar is happening with resistance towards SIA.

Even if the Presumptuous Philosopher is a strong candidate for a good bullet, though, this isn't to say that we should just go ahead and bite. Heck, I've only looked at two theories, here, and only a smattering of considerations. But I think we should keep the possibility of biting on the table.

14 Pascal's muggings, infinities, and other objections to SIA

Let's look briefly at a few other objections to SIA, including some more extreme variants on the Presumptuous Philosopher: namely, a "pascal's mugging" variant, and an infinite variant.

The pascal's mugging variant is just: I can posit extra people-in-your-epistemic-situation faster than you can decrease your credence on my hypotheses. Thus, boom, I suggest that the hypothesis that there are a Graham's number of you-copies-having-your-experiences buried deep in your closet; I suggest that this world overlaps with a Graham's number of other hidden realms containing people like us (see Olum 2000, p. 15); I suggest a Graham's number of hypotheses like these; and so on. Isn't this a problem?

Yeah, it does seem like a problem. But on the other hand, it feels like the type of "big number, not-small-enough probability" problem we (or at least, some of us) are kind of used to not having a particularly clear picture of how to solve (see here for some of the terrible trade-offs; also this); but which we don't currently give up basic and plausible commitments in the face of. That is, you can make up ridiculously large numbers of lives-to-be-saved, too; but this doesn't mean I stop viewing lives as worth saving (though I do get curious about what's going on), or that I start thinking that I need to know how many lives have been saved already in order to decide whether to save another. I'm inclined to treat the possibility of making up ridiculously large numbers of observers-like-you in a similar way: that is, to worry about it, but not to freak out just yet.

Ok: but 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-like-us? And isn't this obviously overconfident? Surely the universe, as it were, *could be* finite—what with finitude being an actual on-the-table scientific hypothesis, for example (see Sean Carroll's comment at 13:01 here). And even if the scientists end up leaning hard towards an infinite universe, couldn't it have been the case that it was finite instead? Isn't that just: a way things could've been? And if there would've been observers in such a situation, wouldn't SIA doom them to being infinitely wrong?

(Though note, of course, that you shouldn't actually be certain of SIA, and so shouldn't be certain of its conclusions. And it's not like SSA is sitting pretty with respect to "no-certainty about infinites" type considerations: to the contrary, SSA becomes certain that we're *not* in an infinite world, once it narrows down "who it is in that world" to any finite population (see Grace's discussion here)).

But SIA's infinity problems get worse. In particular, once it has become certain that it's in some infinite world or other, it's not actually particularly sure about how to reason about which. Suppose, for example, that God has a button that will create an infinite number of copies of you. If heads, he presses it once. If tails, he presses it twice. Here you are post coin-flip. OK, SIA, are you a halfer now? I guess so? No anthropic update for any worlds with infinitely many copies of you? But don't we still want to be thirders in various cases like that—for example, in the actual world, if it's infinite? What if we start using different sizes of infinity?

Yeah, look, I don't like this either. In particular, the combination of (a) being certain that you're in an infinite world and (b) not knowing how to reason about infinite worlds seem an especially insulting double-whammy (thanks to Paul Christiano for suggesting this juxtaposition). But here, again, I want to make noises similar to those I made about pascal's mugging-ish examples, namely: yes, but aren't we also a bit used to "uh oh: this otherwise attractive view, developed in the context of finite cases, says weird/unclear things in infinite cases, and maybe becomes obsessed with them?". See, e.g., expected utility theory, population ethics, and so on. Indeed, totalism about population ethics becomes similarly obsessed with infinities (that is, in creating/influencing them, rather than believing in them), and it becomes similarly confused about how to compare them. And in both cases—SIA, and totalism—the obsession-confusion combo seems no accident. Infinites are confusing things, especially for views that wanted to rely on a relatively straightforward, everyday usage of "more" or "bigger." But infinities are also confusingly *big*. So if you're excited about big stuff, it's easy to end up obsessing about something confusing.

Indeed, in general, 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 comfort. They aren't good problems. But in my opinion, it's good (or at least, respectable) company. And more, it's company that seems to keep being useful, at least as a first pass, despite these problems lurking in the background. Perhaps SIA is the same.

I'll say one other word in SIA's defense re: certainty about infinities. If you really get into the vibe of "I am a particular possible person-in-my-epistemic-situation, who didn't have to exist," and you actually 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 effectively seals-the-deal in terms of whether there was an infinite number of draws. The hat of possible people-in-your-epistemic-situation, after all, is presumably extremely infinite. So what, you think you were going to just happen to get drawn after a finite number of draws? Har har. Yes, there are a *few people* in that situation, in cases of finitely-many draws. SIA does indeed make those people infinitely wrong. But, like, that's not going to be you. You're not going to actually get drawn in those worlds. So you don't need to worry about being wrong in them. To a first and infinitely close approximation, that is, you only exist in infinite-draws worlds. So certainty about them is fine.

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 a halfer then): so why should you update upon waking? My take here is basically just: once you're thinking about it in terms of person-moments, and once you get into SIA's basic ontology, this problem goes away. Sure, *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. (My suspicion is that something similar is the right thing to say about Roger White's "Generalized Sleeping Beauty" problem. E.g., if you grok that waking up is actually evidence, the problem resolves. That said, I haven't really worked through it.)

There are other objections as well. Paul Christiano, for example, pressed the objection in conversation that SIA's reliance on some notion of "counting" the number of observers isn't going to play nice with quantum mechanics or brute physical notions of probability; and that it's going to break in e.g. cases where you split a computer running a given mind into two computationally-identical slices to different degrees (Bostrom discusses cases like this here). My reaction here is: yeah, sounds like there are some issues here, but also, aren't we probably going to need some way of counting observers in at least some cases—for example, the number of observers in God's white rooms, or the number of wakings in Sleeping Beauty, or the number of attendees at your weekly bingo session? I do think things will get gnarly here: but without having dug into it much at all, I'm inclined to think it's the type of gnarly that lots of views (including SSA) will have to deal with.

Does SIA make bad empirical predictions? That would be pretty damning if so. For example, if it turns out that the universe actually *is* definitely finite in a fundamental sense, there's a strong temptation to throw SIA out the window, unless we've somehow revised it to get rid of its certainty about infinities. But we might wonder about whether we can get started with this whole out-the-window process earlier. For example, does SIA predict that there actually *should be* a Graham's number of observers-like-us hiding in my closet? Like, why isn't our world chock full of observers-like-us in every possible nook and cranny? How come I can move without bumping into a copy of myself?

This feels like the type of objection that might move me a lot if we worked it out, but it also seems a bit tricky. In particular, SIA is only excited about *observers-in-our-epistemic-position*. And our epistemic position is: huh, doesn't seem like there are that many observers-like-me hiding in my closet. So it's not actually clear to me that SIA does predict that I'd see such observers after all. That said, I haven't thought very much about this, and I wouldn't be surprised if there's a good objection in this vicinity.

Does SIA imply dynamic inconsistencies? Yeah, I think it does in some cases. For example, what was up with that fifth-ing thing? That sounds like the type of thing you'd want to self-modify right out (and in particular, it sounded like a kind of "double counting," at least from the perspective of the prior—thanks to Paul Christiano and Carl Shulman for discussion). But as I mentioned in part 3, I think lots of views involve dynamic inconsistencies, and my generic, first-pass response is just: if you're worried about dynamic inconsistencies, go updateless. That's what it's *for*. (Or at least, that's how I tend to think about it. Indeed, in my world, one not-especially-charitable gloss on updateless-type decision theories is "whatever the heck I have to say about decision theory in order to have a consistent policy.") That said, I could imagine getting more fine-grained about when exactly different sorts of dynamic inconsistencies are more or less damning, so there's more to be said.

Even if we accept "just go updateless" as a response, though, I can imagine protest: Joe, if you're ultimately going to just go updateless and start acting like e.g. an altruistic EDT-ish totalist SSA-er, or some such, then this whole thing about how SIA > SSA is a big sham (Carl Shulman suggests the analogy: "I'm a Democrat but I oppose every single policy position of Democrats"). And maybe something like this critique is ultimately going to apply. However, my current view on this is: first, I don't want to get ahead of myself on policy positions. As I mentioned in the section on betting, I have yet to go through and really tease apart the relationships and interactions between what seem to me an importantly diverse array of variables, here (e.g., EDT vs. CDT, updateless vs. not, altruism vs. not, averages vs. totals, bounded vs. unbounded, and so on). And I don't think "just moosh the variables together already" is a heuristic that works in service of clarity. I think that others may well have achieved genuine clarity on this front; but I haven't, yet, and I don't want to jump the gun.

Beyond this, though, even if I ultimately want to go updateless, EDT-ish, altruistic, totalist, or whatever, some of my *readers* might not—and such topics seem, naively, like importantly additional discussions, implicating a lot of additional desiderata. And regardless of the outcome of such a discussion, updateful, CDT-ish, selfish, average-ish, etc people can still wonder whether they should believe the doomsday argument, whether they're highly likely to live in an infinite world, whether telekinesis works, or whatever. That is, naively construed, anthropics presents itself as a question about what *we can learn about our situation* from a certain type of evidence. And I don't want to obscure this question by taking on board a bunch of additional ethical and decision-theoretic assumptions, and then saying that granted those assumptions, the question doesn't matter—especially when other people, who don't accept those assumptions, are still going to ask it.

15 Can we make SIA more intuitive?

I'll mention one other objection to SIA: namely, that it doesn't feel especially intuitive as a "conception of yourself and the world." That is, it would be nice if our theory of anthropics *made sense* as a picture of what we are and what the process for deciding whether or not we get created looks like. We could then *return* to this picture when we get confused, and thereby keep a closer grip on why, exactly, we're reasoning this way in the first place. Indeed, absent such a grip, it can feel like we're just scrambling to draw the right sort of line through the curve of the cases. We're making things up as we go along, with no tether to a view about the way the world is.

In part 1, I've presented one SIA-ish "story" about this: namely, that you are a "possible person-in-your-epistemic-situation," who gets pulled from the "hat" of such people, with a likelihood proportional to the number of "pulls" a world implies. But I think it's pretty reasonable to look at this picture and say: what? It's better, I think, than SSA's "God goes hunting for you specifically in the hat, then throws you randomly into the reference class" picture, but it's not exactly clean and pretty. What exactly are these hats? What is the space of possible people? How does the drawing process work? Does "who I am" change as my epistemic-situation changes? What's all this really about?

(I can also imagine a different sort of objection to this sort of story—one that attempts to get some sort of likelihood ratio on your existing, *conditional on SIA's metaphysics, vs. SSA's metaphysics*. That is, SIA essentially imagines that you've won some ridiculous "possible person who got pulled from the platonic realm" lottery. Whereas SSA imagines that you're a special snowflake and that God was dead set on creating you—or, alternatively, that you're the world spirit who was going to experience whoever God created. And which metaphysics posits that a more unlikely event has occurred?)

I don't have especially good answers re: better SIA stories, but here's at least one alternative, which might be more intuitive for some. On this 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, SIA notices that it exists in its epistemic situation, then says: "Ok, who am I?" It 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 Sleeping Beauty, SIA 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, SIA would adjust accordingly, to make sure that it stays 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."

That said, 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 SIA reason as follows? "Ah, actually, the tails people are each twice as likely to exist as the heads people. So, instead of 1/3rd on each, I'll be 2/5ths on each of the tails people, and 1/5th on the heads person. But now, actually, it looks like tails is 4/5ths and heads is 1/5th. So actually, instead of 1/5th on heads person, I'll be 1/9th..." and so on, until it becomes 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/converge towards false certainty. Partly for that reason, I currently don't lean heavily on it.

I'll mention one other possible reframing, which I haven't worked out in detail, but which feels spiritually similar to me. Consider a case where God flips two coins. The first coin decides how many people he creates: if heads, one; if tails, two. The second coin to decide *who* he creates, if there's only one person. if Heads-Heads, Bob; if Heads-Tails, Alice. If Tails-Heads or Tails-Tails, though, God creates both Alice and Bob.

Suppose you wake up and find that you are Bob. Bob exists in three out of these four cases, all of which are equally likely, so it seems very natural, here, to be 1/3rd on each, and hence 1/3rd on the first coin having landed heads. We don't have to do any special sort of anthropic updating in favor of worlds with multiple Bobs—there aren't any. Rather, here, it's SSA that does weird anthropic stuff: in particular, it downweights the Coin-1-Tails worlds, because in those worlds, it's 50% that you would've been Alice.

SIA's very natural, non-anthropic type of reasoning here feels pretty similar, though, to the type of thing that SIA is trying to do all the time, *but in the context of uncertainty about who it is*. Thus, suppose that in the case above, you wake up and don't yet know whether you're Alice or Bob. However, you know what you *would think* if you knew you were Alice, and you know what you *would think* if you knew you were Bob: in both cases, that is, you'd be 1/3rd on Coin-1-Heads. And having no special reason to think that you're Alice instead of Bob, you're 50-50 on who you are. Thus, given 50% that you should say 1/3rd, and 50% you should say 1/3rd, you say 1/3rd. Maybe thinking about this sort of case can help shed light on SIA's basic shtick?

Still, I don't feel like I've nailed it in terms of "this is the way the world actually is, such that SIA makes sense." And I think a lack of a fully intuitive picture here is a significant barrier to really "believing" in SIA, even if it looks good on more theoretical grounds.

16 Hold out for Anthropic Theory X?

There are, presumably, lots of other objections I'm not discussing. Indeed, I feel some worry that in the process of writing this post, I've given too much attention to the problems with SSA that feel like they shout from the pages of Bostrom's book in particular, and that I haven't gone hunting hard enough for the strongest possible case against SIA (or for SSA's strongest possible defense). But, this sequence is long enough already. For now I'll just say: I expect to keep learning more.

Indeed, I can well imagine a version of this post that focuses less on comparing SSA and SIA, and more centrally on a message in the vicinity of "SIA and SSA are *both* terrible, dear God help us." That, indeed, is sometimes the vibe I get from Bostrom (e.g. here), though his book devotes more attention and sympathy to SSA in particular. And regardless of the comparative merits of SSA and SIA, I wouldn't be surprised if really taking infinite cases in particular seriously forces a kind of start-over-and-build-it-all-back-up-again dynamic, rather than some sort of “patch” to an existing view.

Perhaps, then, faced with such unappetizing options, we should refuse to eat. That is, we should recognize that (in this blog sequence at least), we have yet to find a remotely plausible or satisfying theory of anthropic reasoning, and we should “hold out” until we find an “Anthropic Theory X”—one that gets us everything we want (or at least, much more than SIA and SSA do). (Thanks to Nick Beckstead for suggesting this type of response, and for the name “Anthropic Theory X”—a reference to Derek Parfit’s name for the elusive, fully-satisfying theory of population ethics he was searching for).

And to be clear, such a Theory X may already exist. As I’ve tried to emphasize, I’ve only been discussing two basic and prominent views: I haven’t tried to survey the literature as a whole.

And even if no one has invented such a theory yet, it may still be out there, in theory space, waiting for us to find it. In particular, we do not to my knowledge yet have “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 (thanks to Nick Beckstead, again, for suggesting this consideration).

That said, I’m not sure we’re so far away from proofs in this vein. 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 (at least apparently) structurally identical contexts, that lead to some of their worst results.

Consider, for example, Sleeping Beauty. What, actually, are you going to say in Sleeping Beauty, if not 1/2, or 1/3? (Let’s leave aside “fifthing” for now, along with “incorporate your uncertainty about anthropics” type moves.) 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 “you should be 1/4th if you wake up on both days no matter what, and then if you learn that you’re not Heads-Tuesday you should clearly end up a thirder.” Perhaps, indeed, you’d be inclined to put some of the premises of those arguments into your “impossibility proofs” as one of the Y constraints. (Are “impossibility proofs” and valid arguments especially different? The former has an aura of technical finality and “having really created knowledge.” But we’ve had impossibility proofs for ages that there is no Theory of X of Socratic Immortality, such that (1) Socrates is a man, (2) All men are mortal, and (3) Socrates is not mortal.)

Ok, so say 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, “no magic about the number a zillion” seems like it could also be one of those Y constraints that we could put into our impossibility proofs. But now Theory X is getting pretty darn confident about tails. Presumptuously confident, you might say.

Indeed, as I tried to emphasize above, the Presumptuous Philosopher and its variants are just science-ified versions of a zillion-wakings Sleeping Beauty. We can futz about whether it matters that e.g. the prior is some kind of objective frequency in Sleeping Beauty vs.

some build up of empirical evidence about fundamental reality in a more scientific case. And maybe there's stuff to say here, and other differences to bring out. 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. Uh oh: sure seems like this should be an update towards tails, relative to a merely zillion-waking world. After all, we can count at least a zillion wakings in the infinite world pretty easily (just label them starting from 1). Indeed, sounds like we might be in for it re: for any finite number of wakings, being more confident than that. But that sounds like certainty.

Obviously, there's more to say here, especially about infinite cases. The thing I want to point at, though, is that the distance between "thing we really want to say" and "thing we really 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 case with a different vibe, or a greater feeling of "extremity," or a more serious sense of real-world implication. This dynamic currently inclines me towards pessimism about finding an especially satisfying theory X, at least of a standard kind, that avoids both SIA and SSA's problematic implications. I think SIA and SSA, and their closely-similar variants, might be covering more of the conventionally-plausible ground than it initially appears. Perhaps, indeed, those who seek theory X do better to try to reconceptualize the whole terrain, to "dissolve the problem entirely," rather than to approach it on its own terms (Armstrong's Anthropic Decision Theory has a bit of this flavor). But even then, for whatever reconceptualized equivalent of the question, one wonders: 1/2, or 1/3?

17 Implications

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

People have said various things about this. A classic thing is that we stop believing in the Doomsday Argument, which sounds pretty good. Indeed, one update I've made in the course of writing this post is towards not-buying SSA-ish arguments for doom soon.

But actually, maybe SIA suggests it's own Doomsday Argument: namely, that probably the reason we don't see much intelligent life out there is because ~all life kill itself *after* reaching our stage, rather than ~never reaching our stage at all, because this story makes it more likely that there's lots of life out there at our stage and hence like us (see Grace here; and some related discussion here). But actually, maybe SIA doesn't say this, and instead says we should update towards overwhelming probability of being in a simulation run by a maximally powerful civilization devoting ~all of its resources to simulations of us-in-particular (thanks to Carl Shulman for discussion; see also section 4 here for more on SIA and simulations). But actually, maybe SIA doesn't say any of this, because it has a seizure immediately after becoming certain that we're in an infinite universe (and presumably, the biggest possible infinite universe? Something something modal realism Tegmark-Level-4 Ultimate Multiverse? Or something bigger, with impossible worlds too?).

Overall, I feel pretty far away from any kind of clear picture of what SIA would actually imply, especially given the seizure thing. And indeed, as with the acausal wackiness I discussed in my last post, it feels to me like we're at a sufficiently early stage in reasoning about this stuff that we ought to tread very carefully, and avoid making updates that seem pretty conventionally silly or extreme. What's more, I've been explicitly setting aside stuff about decision theory, copy-altruism, and so on, all of which could well change the practical game entirely in terms of "what does this imply," and maybe restore various types of "normality." For example, as I gestured at above above, if you're updateless in a way that counteracts some of SIA's implications (for example, if you're a copy-altruistic EDT-er who self-modifies to avoid fifth-ing), you may end up acting like an (EDT-ish) SSA-er in lots of cases, even if you like SIA in principle (thanks to Carl Shulman, Paul Christiano, and Katja Grace for discussion).

That said, as with the acausal wackiness, I don't think "cool let's just ignore this stuff entirely" 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 an *extremely 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 (e.g., "Ok but tell me right now what practical difference this is going to make"), it seems useful to have on hand. A more accurate basic picture of your existential situation as a whole, for example, seems pretty robustly worth having. And some of us are also just curious.

What's more, doing anthropics *badly* has costs. You end up confused about the doomsday argument, for example. You end up reasoning badly about the fine-tuning of the universe. You end up wondering whether your metaphysical essence is compatible with being a chimp, or a bacterium. At the very least, we need some sort of anthropic hygiene, to avoid making these sorts of errors. And the line between "avoid basic errors" and "make actually-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 especially sure what it implies in the real world. We should try to figure out a better theory (the need to handle infinite cases seems especially pressing), and perhaps there is one out there already (as I've said, I haven't looked hard at the alternatives). In the meantime, we should tread carefully, but stay interested in understanding the implications of the theories we have.