Simulation arguments

Joe Carlsmith

February 18, 2022

> *"If Nature herself proves artificial, where will you go to seek wildness? Where is the real out-of-doors?"*
>
> – C.S. Lewis

(Content warning: weird)

Summary

- This post examines "simulation arguments"—i.e., arguments that we should assign significant probability to living in a computer simulation.
- I distinguish between two types. The first type starts with our "standard story" about the world, and tries to argue that *on this* story, unless something prevents the creation of suitably many sims suitably like us, we're probably sims. I think this version fails: on our standard story, none of the sims will make our observations. For example, they won't observe the specific set of books on my bookshelf.
- This sort of objection feels cheatsy—and ultimately, I think, it doesn't work on the best version of the argument. But I use a series of cases to bring out its similarity to less cheatsy-feeling cases: for example, cases in which it appears that all the sims that will ever get made are squid-people with tentacle arms. Simulation arguments, I suggest, should say the same thing about *all* these cases.
- But the "indifference principles" that classic simulation arguments rely on get confused about this. In particular, they end up having to hand-wave about which things we count as "knowing," and/or about which "reference classes" are "admissible." This confusion makes it seem like classic arguments require an unprincipled kind of "selective skepticism"—one that risks undermining itself.
- I think we do better to focus on a second type of simulation argument, which relies centrally on a claim about what sorts of *priors* it's reasonable to have over "structurally similar" worlds that contain your observations. In particular, instead of saying "on your standard story, you're probably a sim," this version says: "if *that's* your story, you should put lots of weight on these alternative, structurally similar stories, too—on priors, such stories are similarly likely, and they contain your observations, too. But on *those* stories, you're a sim."
- I think this second type of argument works better—and that it brings more to the table than a standard skeptical "what if." Indeed, I think that classic formulations of the simulation argument are implicitly leaning on something in the vicinity. But bringing out the underlying assumptions about "reasonable priors" more explicitly helps to avoid the confusion that classic formulations often create.

- That said, Type 2 arguments raise their own questions—for example, about how the relevant notion of "structural similarity" should be understood and applied, and about serious flirtations with "epistemic pascal's muggings." These questions (and other heuristics) leave me wary of Type 2 arguments, too.
- I also discuss two further simulation arguments. The first is from Thomas (2021), who argues that we're almost certainly sims, because *conditional* on being non-sims, the expected ratio of sims to non-sims is very high. I think this argument rests on learning the lesson of the classic argument in one breath, but forgetting this lesson in the next.
- The other argument is that even conditional on finding yourself (apparently) living in earth's early history, finding yourself as an "unusually something" person (e.g., unusually influential, entertaining, etc) should be some kind of substantive additional update in favor of being a sim. I don't currently see how this argument is supposed to work; and on the classic argument, it seems like Donald Trump's credence on "I'm a sim" should be basically the same as that of a janitor in Des Moines (even if there will be many more Trump sims than janitor sims).
- I conclude with a few reflections on "simulation woo," and on whether being in a sim would compromise the world's "otherness."

Thanks to Katja Grace for especially extensive discussion of the issues in this post: the "Type 2" formulation I present was centrally her suggestion. And thanks to Paul Christiano, Owen Cotton-Barratt, Cate Hall, Ketan Ramakrishnan, Carl Shulman, and Teruji Thomas for discussion as well. See Birch (2013), Thomas (2021), and Garfinkel (unpublished) for similar points about "Type 1" arguments. Garfinkel's discussion also gestures at something similar to "Type 2," and it has done a lot to influence my thinking on this topic more generally—influence on display throughout this post.

1 The core of a simulation argument

Consider:

> HIGH RATIO: There are many more sims than non-sims in an epistemic position "like ours."

"Like ours" hides key gnarly-ness. For now, though, let's set that aside, and assume anyone having human-ish, 21st-century-ish experiences meets the standard.

I think of simulation arguments as resting on a core claim about reasonable credences with respect to HIGH RATIO, namely:

> CORE CONSTRAINT: Conditional on HIGH RATIO, we're probably sims.

That is: you're not allowed to think that we're non-sims in a HIGH RATIO world.

If we accept this constraint, then one game is to haggle about whether HIGH RATIO is true. Maybe, for example:

1. Sim can't be conscious.
2. Sims with human-ish, 21st-century-ish experiences take too much compute, even for a technologically mature civilization.
3. Too few civilizations reach technological maturity.
4. Too few technologically mature civilizations decide to create lots of sims with human-ish, 21st century experiences.
5. Something else, such that HIGH RATIO is false.

Following Chalmers (2022), let's call claims like these "sim blockers." On Chalmers' formulation of the argument, then, either there are sim blockers, or you're probably a sim. Bostrom (2003) goes a bit further, and attempts to rule out all sim blockers except 3 and 4. But it's the same idea.

Once you're haggling about sim blockers like 1-4, though, there's a fairly clear path to substantive credence on "I'm a sim": namely, on a standard picture of the world, claims like 1-4 don't look probable enough to get you below, say, 10% on being in a sim (Chalmers is at >25%, and early Bostrom seems to be at ~33%—though in recent interviews, he explicitly "punts" on probabilities). Indeed, I know some people who are well above "probably" on being sims, partly due to skepticism about 3 and 4, especially in "big world" cosmologies where lots of civilizations have a shot at making it to a sim-filled future.

So it seems like we have a simple argument for a pretty strange conclusion. On closer inspection, though, the argument can start to feel slippery. In particular: standard justifications for CORE CONSTRAINT can easily suggest a version of the simulation argument that doesn't, actually, work, and which leads to lots of confusion along the way.

But there's also a better framing available. This framing, though, leads to weirder places. Let me explain.

2 Not just standard skepticism

> *"We accept the reality of the world with which we're presented. It's as simple as that."*
>
> – From The Truman Show

Philosophers often talk about "skeptical scenarios." Maybe, for example, you're a brain in a vat. Maybe your mind is the only thing that exists. Maybe the world was created five minutes ago with the appearance of age.

By construction, a skeptical scenario is one that in some (sometimes controversial) sense, you can't tell you're not in. So it's a scenario in which an epistemic procedure that assigns most of its probability to a "standard story" (e.g., there are no demons, the fossils are real, etc) will lead you astray. Imagine, for example, that you really *are* a brain in a vat, in a lab run by mischievous scientists. Imagine these scientists watching you consider this possibility, but then dismiss it, maybe while babbling about some fancy philosophy thing ("context-dependent knowledge," "causal-role semantics," "externalism about justification"). Don't you look the fool? The scientists are laughing at you.

Still, thems the breaks. Faced with someone who says: "what if you're in a skeptical scenario?", often the best you can do is say: "yeah, what if." By construction, you can't rule it out.

But "can't rule it out" is different from "substantial probability." And assigning *probabilities* to skeptical scenarios is a much harder game—one that isn't the philosophy literature's forte (philosophers tend to be more interested in the relevance of skeptical scenarios to what we "know").

Indeed, naively, when considering skeptical scenarios, you've basically just have to go on priors. After all, if before making observations O, you started out with an X% prior on a skeptical scenario in which you observe O, then observing O should leave you with at least X% on that scenario. But priors are notoriously hard to argue about. We tend to think it reasonable to give a very low prior to claims like "my mind is the only thing that exists." But our stories about *why* are hazier (though see here for one attempt, focused on a particular notion of what scenarios are "simple").

Simulation arguments, though, are supposed to be much more than "what ifs." The point isn't to add yet another skeptical scenario to the roster of "can't rule it outs." Rather, as Bostrom (2011) puts it:

> "[The simulation argument] begins from the starting point that things are the way we believe they are, and then, while granting us that we might be justified in assigning a high initial credence to these beliefs, nevertheless tries to show that we have specific empirically-grounded reasons for revising these initial beliefs in a certain way—not so as to make us generally agnostic about the existence of an external world, but to accept the disjunctive conclusion."

The aspiration, then, is to provisionally *grant* something like a "standard story"—and then to show that it leads us to assign significant credence to HIGH RATIO. Per CORE CONSTRAINT, then, either we need to retract that credence, or we need to give significant probability to being sims.

In this sense, the argument is supposed to have a kind of "gotcha" flavor. "You thought that you could believe X standard thing, and still think you're not a sim. But: gotcha! You can't."

But how, exactly, is this "gotcha" justified? Let's distinguish between two approaches.

3 Type 1 arguments: "On that story, you don't know where you are."

The first approach, which I'll call "Type 1," tries to show that absent sim blockers, our standard story involves drawing a map of the world on which we don't know "where we are." That is: we draw a small number of observers labeled "non-sims having human-ish, 21st century experiences," and a giant number labeled "sims having human-ish, 21st century experiences," and then we step back and realize: "sh**, I can't tell where I am on this map—and based on the ratio of sims to non-sims with experiences like mine, I'm probably a sim." Gotcha.

In this sense, a Type 1 framing tries to show that our situation is like the following case:

> SIMS WITH RANDOM NUMBERS: You wake up in a white room, with the number 3 written on your hand. A sign in front of you reads. "I, Bob, created nine sims, and one non-sim, all in white rooms, all with identical signs. Then, for each observer, I drew a number (without replacement) out of a hat containing the numbers 1-10, and wrote it on their hand. No one else exists, other than me, the nine sims, and the one non-sim."

The sign, here, is supposed to be like the "standard story" (absent sim blockers). The numbers are supposed to be like the differences between the various human-ish, 21st-century-ish experiences posited by the standard story map—differences that don't provide any evidence about one's simhood.

And indeed, something like CORE CONSTRAINT looks good here. Specifically: if you condition on the sign telling the truth, it's natural to put your credence on "I'm a sim" at 90%. After all, you don't have "anything to go on" other than (a) your first-person observations (the white room, the number on your hand, etc) and (b) your knowledge that there are nine sims, and one non-sim, all making observations like yours, though with different randomly drawn numbers on their hands. So it's natural to split your credence about "who you are" evenly between each of the observers seeing white rooms, signs, and so on—i.e., to be "indifferent." After all, what else would you do? Why think something else?

Now, some people worry about this sort of move (see e.g. Weatherson (2005)). In worlds with an infinite number of observers with observations like yours, for example, splitting your credence evenly looks naively like it'll give 0 credence to everyone, which can cause problems. This and other issues cause some (for example, advocates of UDASSA) to abandon indifference.

But we don't actually need indifference, here. Rather, what we need is a weaker constraint, which says that in cases suitably like this, then conditional on the sign telling the truth, *you shouldn't be weirdly confident* that you're not a sim. Maybe you're not at exactly 90% on being a sim. But that's the right ballpark. And "who knows, any probability is equally reasonable" (e.g., "99% that I'm the seventh sim Bob created!") looks pretty silly.

So overall, I think we should agree that if our situation is like SIMS WITH RANDOM NUMBERS, then conditional on "the sign telling the truth," we should think we're probably sims.

4 Squid people

But now consider a different case:

> Nine squids, one human: You wake up as a human in a white room. A sign in front of you reads: "I, Bob, created nine sims, and one non-sim, all in white rooms, all with identical signs. The sims are all squid-people with tentacle arms. The non-sim is a human. No one else exists, other than me, the nine sims, and the one non-sim." Conditional on the sign telling the truth, what should your credence be that you're a sim?

Here, 90% is looking dumb. 0% is a better ballpark. After all, conditional on the sign telling the truth, *all the sims are making observations different from your own*. In particular, they're observing tentacle arms.

The problem for Type 1 simulation arguments, though, is that our situation with respect to the "standard story" is actually closer to Squid-person sims than to Sims with random numbers. That is, I think that denying "sim blockers," but accepting the rest of the "standard story," leads to a picture of the world on which there are tons of sims *whose observations are different from our own* (though: less dramatically different than tentacle arms).

To see this, let's first consider a slightly more realistic, but still highly simplified and ridiculous, version of the case.

> Future squid sims: As our science progresses, it becomes increasingly clear that the universe is finite, and that humans are the only intelligent species that inhabit it. What's more, it really looks like humanity is on track to reach technological maturity, and to end up in a position to run zillions of simulations. The stable world government, though, has engaged with the simulation argument, and after some confusing philosophical discussion, it passes a law that everyone expects to be strictly enforced: namely, that all simulations that are ever run can only involve squid-people with tentacle arms. No sims of humans will ever be allowed. However, some of the squid-person simulations will be of squid-people evolving on planets somewhat like earth, and they will look, from the inside, like the early history of a "squid civilization"—a civilization that will one day be in a position to run zillions of its own sims.

Here, if we deny "sim blockers" like "humanity fails to reach tech maturity after all," and we accept the "standard story" on which humanity will go on to create zillions of simulations of squid sims, then we are indeed accepting a picture of the world on which there are tons more sims than non-sims. But this picture *does not* put us in a position analogous to Sims with random numbers. Rather, it puts us in a position analogous to Nine squids, one human. That is, the sims that the standard story is positing are sims with observations *different* from our own (see also Garfinkel on what we should think, if we find ourselves running a bank of Charlie Chaplin simulations).

Indeed, even absent "sim blockers," neither the Bostrom nor the Chalmers version of the simulation argument tries to say that there's some sort of problem with thinking that you're a non-sim, here. Rather, both focus solely on a version of High ratio that only grants sims with *human-type* experiences the status of being in an epistemic position "like ours." And as a rhetorical choice about what version of the argument will feel most naively compelling, I think this makes sense: "you can't think that you're an early human non-sim in a world with lots of squid sims" seems a lot less intuitively forceful than e.g. "you can't think that you're an early human non-sim in a world with lots of highly-realistic human ancestor simulations." But actually, I'll suggest, these stand or fall together. And on the Type 1 argument, they fall.

5 Will our descendants ever simulate my bookshelf?

To see this, consider a series of cases that get more like our actual case. For simplicity, we'll keep the "small world" assumption that (on the standard story) the universe is finite, and that humans are the only intelligent life within it. But we'll remove the government's binding commitment to never run sims with humans in them. Consider, instead:

- Low-res sims: The government settles on some way of measuring the "resolution" of our experiences, and it declares that all human sims will be run at "lower resolution"—such that anyone from our world who transitioned into a sim world would notice a marked reduction in the vividness and detail of their perceptions.
- Hot and happy sims: The government declares that all future human sims have to be extremely attractive, entertaining, and happy—much more so than any existing humans.
- Flawed ancestor sims: The government places no restrictions on which sims can be run. Indeed, it actively plans to run lots of extremely detailed "ancestor simulations," which it will try as hard as possible to make accurate. However, even with tons of effort, lots of stuff will be different and/or lost to history. For example, most of such ancestor simulations will consist centrally of people who never existed; when these simulations include sim versions of actual historical figures, the personalities of these figures will need to be inaccurately reconstructed based on whatever patchy evidence survives through to the post-human future; and the same sort of inaccurate reconstruction will apply to everything else, too—rooms, restaurants, conversations, bookshelves, clothing, and so on.

All of these, I suggest, are actually analogous to Future squid sims. That is, in all of these cases, conditional on no sim blockers, drawing up the "standard story" of the world involves positing lots of future sims *with observations different from your own*.

The last case—Flawed ancestor sims—is the trickiest, and it's very close to the version that Bostrom focuses on (he doesn't include the "small world" constraint, but the argument ought to work regardless). And it is indeed tempting to think that, in a standard story world with lots of ancestor sims, the historical details we see around us are in some sense "independent" of whether we are sims or non-sims, such that conditional on the existence of such ancestor sims, our situation starts to look like Sims with random numbers with a true sign—i.e., everyone has *different* observations, but the differences aren't evidence about whether you're a sim or not.

But this temptation derives, I think, from subtly construing the "standard story" at a higher level of abstraction than it actually operates on. That is, the standard story is not a map of the world in which the humans (whoever *they* are—*if* they even exist) make some unspecified set of 21st-century-ish observations, then go on to create ancestor simulations. Rather, we build out the standard story's "map" by *starting* with what we see around us.

Thus, for example, let's say, in my current situation, that I start drawing my map of the world by drawing a human named Joe, in a room with XYZ specific books on the bookshelves in his room—the books that I, right now, can see. From there, I draw this Joe's past, his future, the Bay Area, California, Earth, and so on. And in that earth's future, let's say that per FLAWED ANCESTOR SIMS, I start drawing various bits labeled "ancestor sims run by post-human civilization." However, per the "slightly flawed" constraint, let's say that I'm in a position to assume that *none* of the ancestor sims in Joe's future feature another Joe with that same specific set of XYZ books on his shelves. Rather, the specific make-up of the historical Joe's bookshelf will be lost to history.

Now suppose I finish my map, and step back to take a look. Per the Type 1 argument, am I hit with a "Gotcha! You don't know where you are on this map—and probably, you're in one of the bits labeled 'ancestor sims.'" No, I'm not. After all, I look around me, and there are XYZ books, shining in their specificity. And there is a little note on the map, pointing to the ancestor simulation bits, reading "none of the Joe-like humans in these ancestor simulations have XYZ books on their shelves—if there are even any Joe-like sims at all."

What's more, this is the *same sort of note* posited by the standard story in cases like FUTURE SQUID SIMS, LOW-RES SIMS, and HOT AND HAPPY SIMS. In all these cases, we draw out a map of the world that starts with humans making our specific observations, and which puts in the future of those humans some large numbers of simulations. All of those simulations, though, have notes on them, indicating some feature of the sims that differentiates their observations from those of the original humans—e.g., "observing tentacle arms," or "noticeably lower-resolution perceptions" or "hotter and more entertaining." The differences are more stark than in FLAWED ANCESTOR SIMS. But the epistemic significance of the notes seems analogous.

To be clear: the point here isn't that our standard story map includes a little pointer saying "this is me." Rather, it's that we start drawing our standard story map by drawing someone making our observations. But unless there's someone else in the world making exactly these observations, such observations are *enough* to let us locate ourselves on the map, once we've drawn it. So Type 1 arguments fail: equipped with our "standard story" map, we aren't "lost" with respect to our place in the world, even if the people making our observations have lots of sim in their future.

6 Do the random light speckles in my visual field prove I'm not a sim?

Perhaps you're thinking: bleh. This feels like a lame objection to the simulation argument. And basically, I agree: I think pointing at the books on my bookshelf is in fact a bad way of arguing that I'm not a sim, in FLAWED ANCESTOR SIMS. But is pointing to the fact that I'm a human a bad way of arguing that I'm not a sim, in FUTURE SQUID SIMS? Less clear. My point is that we should say the same thing about both cases—and that Type 1 arguments get a grip on neither.

To bring out the problem with the "bookshelf" counterargument, though, let's look more closely at the limiting case of genuinely indistinguishable observations:

> PLANNING ON INDISTINGUISHABLE SIMS: You are a simulation scientist. You've been working on a technology that will scan a non-sim's body and brain, and then create sims with experiences subjectively indistinguishable from those of the scanned non-sim. The scanner operates by continuously scanning anyone who is in a certain white room in your lab, such that it can recreate any of the experiences that occurred while inside. Inside this room you've placed a red button, with a sign on it that says: "If you are a non-sim, this button will create nine sims with experiences exactly like yours, facing a button and a sign that look just like this one. If you're a sim, though, pressing the button won't actually create any new sims—that would take too much compute." You enter the white room. You are currently planning to press the button.

Here, it looks very plausible to me that to the extent that you buy the sign's story (you wrote it—or at least, you think you did), then conditional on being in the white room and planning to press the button, you should think you're probably a sim. And it also seems like the "standard story" runs into a more direct sort of Type 1 "Gotcha! If this is your map of the world, you don't know where you are."

That is, suppose that once you're in the white room, you start drawing out a "standard story" map by drawing a non-sim scientist entering the white room, making your current observations, and then pressing the button. You then start drawing nine sims, all with experiences subjectively indistinguishable from the ones you're having right now. Now, I think, if you step back and look at your map, you should be getting confused. *This* is the sort of map that doesn't allow you to pin down which observer you are: you've got nine sims, and one non-sim, all making the exact same observations you're making now. Conditional on this map being accurate, then, and with nothing else to go on, it seems natural to split your credence about "where you are" roughly evenly between them all—and hence, to put ~90% on being a sim (see Elga (2004) for more on this type of reasoning).

If we say that, though, should the fact that the experiences are *exactly indistinguishable* really make a difference? Consider a slightly altered version of the case:

> PLANNING ON SIMS WITH DIFFERENT LIGHT SPECKLES: Same set up as before, except that the scanner is *slightly* imperfect. In particular: it can't exactly reproduce, in the sims, the specific patterns of random light speckles in the visual field of the non-sim. Rather, the sims see their own, distinct random patterns, which the non-sim never saw.

Suppose that in this case, you find yourself in the white room, planning to press the button. Can you reason as follows? "Well, consider this particular little speckle of light I just saw near the top of my visual field. When I draw my 'standard story' map, I first draw a non-sim scientist with this sort of speckle in her visual field. Then I draw nine sims in her future, with a little note on all of them saying 'does not see that particular speckle of light.' Then I step back and—voila! I still know where I am. So, on my standard story, I'm confident I'm not a sim."

To me this look pretty silly. Certainly, it seems like the type of reasoning that is going to screw the sims over, hard (though this is true in all of the cases I've considered—and it's true in skeptical scenarios more generally). But also: it seems strange to posit some important discontinuity between Indistinguishable sims and Sims with different light speckles. Suppose, for example, that as you're striding confidently towards the button in Sims with different light speckles, you notice a little note from one your grad students pinned to the scanner, which says: "I fixed the scanner! Now it perfectly captures the non-sim's light speckles." Should that note really be some sort of big update about whether you're a sim?

One feels like: no. You should say the same thing about both these cases. But Planning on sims with different light speckles is basically the same case as Flawed ancestor sims (light speckles = bookshelf contents)—which in turn, I've argued, is basically the same case as Future squid sims (and Low res sims, and Hot and happy sims). If the "standard story" gets us worried that we're sims in Planning on indistinguishable sims, then, it seems like it should get us worried in all these cases.

7 Maybe indexical dogmatism?

Now, one option here is to stick to your guns and say: "I'm not even worried in Planning on indistinguishable sims." And indeed, there is a way of telling the "standard story" that makes this seem at least somewhat intuitive. Suppose, for example, that the indistinguishable sims are going to be run on a set of laptops in the office next to the white room—on the left, if you're facing the button. Faced with the button, then, there's something intuitive about pointing to the left and saying: "On my standard story, the sims are going to be run *over there*. But I'm not *over there*. Rather, I'm *here*. Thus, on my standard story, I'm not a sim."

Similarly, you might say: "Look, on my standard story, *I'm going to create whatever sims will exist*. But I know that I'm not any of the sims that I create. Thus, on my standard story, I'm not a sim." (See Crawford (2013) for more in this vein.)

Indeed, the simulation argument is often casually framed in a way that seems vulnerable to this type of objection. Thus, for example, it can feel like the argument is saying: "In the future, absent sim blockers, it seems like our descendants will create tons of ancestor simulations. But if that's true, then probably we're one of those simulations." But now one feels like: huh? You said that the sims were in *our* future, created by *our* descendants. So you just called me the child of my children. You just pointed ahead of us in time, and said "probably we're over there."

In the context of sims with genuinely indistinguishable observations, these sorts of objections are drawing a subtly different kind of "standard story" map: one that includes a little pointer to the non-sim saying "this is me," which *doesn't* get justified via appeal to the non-sim's observations. That is, this sort of standard story starts out knowing where you are, and then it refuses to get lost, even in the presence of other observers making exactly the same observations you are. Let's call this "indexical dogmatism."

Now, you might think that "indexical dogmatism" is like saying that you're the non-sim in a case like:

> IDENTICAL SIMS: You wake up in a white room. A sign in front of you says: "I, Bob, created nine sims and one non-sim, all with exactly identical observations."

Here, conditional on believing the sign, it seems wild (to me) to "stick to your guns" (what guns?) and say that you're the non-sim.

But indexical dogmatists would say that the case is more like:

> IDENTICAL SIMS WITH MISLEADING SIGNS: You wake up in a white room. A sign in front of you says:
>
>> "Claim 1: I, Bob, created nine sims and one non-sim, all with exactly identical observations.
>> Claim 2: You're the non-sim."

Here, the signs are getting more f***-ed up. In particular, if you believe the *whole* sign, then you should believe Claim 2, and think that you're the non-sim (Claim 2 is analogous to the little pointer, in the indexical dogmatist's 'standard story map," which says "this is me"). True you should think that there are a bunch of other sims out there, making exactly your observations, but getting lied to by their sign; so conditional on the sign telling the truth, there's a sense in which you "got lucky"—luck that can feel epistemically unstable to posit. But here the indexical dogmatist's move is to say: "Look, I got into this whole 'what if I'm a sim' game because I started believing my sign. But if I believe my sign, I'm not a sim."

Similarly, in PLANNING ON INDISTINGUISHABLE SIMS, you might argue: "Look, on my standard story, that button—that is, the button in front of me — is going to create sims. Now, either that button is going to create sims, or it isn't. If it is, then I know I won't be one of the resulting sims, since I created them. And if it isn't going to create sims, then why am I worried about my world satisfying HIGH RATIO? I only got into that worry by thinking that my button would work." (See also arguments to the effect that if we're going to create sims, that's actually evidence that we're *not* sims, because the sims won't generally be allowed to create new sims—that's too much compute.)

I don't, ultimately, think that this is good reasoning. But it's a type of reasoning that I think it's important for simulation arguments to grapple with, and which often arises as a source of confusion.

Indeed, sometimes Bostrom really *courts* this confusion. Consider his (2011) response to the charge that the simulation argument is self-undermining, because if you're in a simulation, the empirical evidence about neuroscience, computation, and cosmology he appeals to isn't a reliable guide to what's going on in universe simulating us. Here, Bostrom appeals to a disjunction:

> "A. If we are in a simulation, then the underlying reality is such as to permit simulations, it contains at least one such simulation, and [we are sims] is true.
>
> B. If we are not in a simulation, then the empirical evidence noted in the simulation argument is [sic] veridical taken at face value, suggesting that a technologically mature civilization would have the ability to create vast number of simulations; and consequently, by the simulation argument, there is a very high probability at least one of [most civilizations go extinct prior to technological maturity; most technologically mature civilizations don't make lots of ancestor sims; we're sims] is true."

Here, B looks confused. In particular: in none of the "we're not sims" worlds are we sims. So none of your credence on the B horn, here, should end up on the "we're sims" hypothesis. That's denying B's antecedent. (There are ways of reconstructing what Bostrom is saying here that make more sense—but I think his presentation, at least, muddies the waters.)

What's more, the indexical dogmatist wants to say that the "standard story" *includes* "we're not sims"—or at least, it includes a pointer saying "we're these people," hovering over some non-sims that might have sims in their future. So trying to get to "we're sims" via the indexical dogmatist's standard story seems like a non-starter—the same sort of non-starter as arguing for "we're sims" via B.

8 Selective skepticism

Now, maybe you don't like indexical dogmatism. Indeed, I'm not a fan: in Planning on indistinguishable sims, it seems very natural, to me, to get pretty worried about being a sim that a scientist "like you" created.

But I think the indexical dogmatist's questions about "why are you believing one part of the 'standard story,' and not this other part" are important—and they apply even if we require that the "standard story" be stated entirely in third-personal terms. Indeed, if anything, they become more forceful.

Consider, for example:

> Signs that lie about animals: You wake up as a human in a white room. A sign in front of you reads:
>
>> "Claim 1: I, Bob, created nine sims and one non-sim, all in white rooms.
>>
>> Claim 2: The sims are all squids, and the non-sim is a human.
>>
>> Claim 3: The sims see a different sign—one that claims that all the sims are lion-people, and the non-sim is a squid."

From the perspective of the "standard story," this case, I claim, is basically just FUTURE SQUID SIMS. And I've argued that we should say the same thing about FUTURE SQUID SIMS that we say about FLAWED ANCESTOR SIMS, and that we should say the same thing about FLAWED ANCESTOR SIMS that we say about PLANNING ON INDISTINGUISHABLE SIMS.

But here's a bad argument about this sort of case: "Either the sign is telling the truth, or it's lying. If the sign is telling the truth, then probably the sign is lying, and I'm a sim. If the sign is lying, I'm probably one of the sims the sign told me about—the sims getting lied to by their signs. Either way, I'm probably a sim."

In fact, both horns of the disjunction are bad here. The first horn is bad because, if the sign is telling the truth, it's not the case that the sign is probably lying. Yes, there are nine *other* (simulated) signs that are lying. But conditional on this sign telling the truth, you shouldn't assess its probability of lying by randomly sampling from the signs in the world.

The second horn is bad because conditional on the sign lying, it's not at all obvious that you should conclude that you're probably a sim getting lied to by their sign. In fact, scenarios where there are lots of sims getting lied to by signs are salient *centrally because you started believing the sign*. If you stop believing the sign, why should you think that the most likely "lying sign" worlds are "I'm a sim" worlds?

So it can seem like a non-"What if?" argument for "I'm a sim," here, requires a kind of "selective skepticism" about the "standard story" (see Birch (2013) for more). That is, it can't believe the *whole sign*, or it will admit that it's not a sim—the sims, after all, are squids. And it can't *dismiss* the whole sign, or it's thrown back to reasoning entirely "on priors" about whether it's likely to be a sim, without any sort of active evidence for this hypothesis.

Rather, to get a positive argument for "I'm a sim" going, it needs to thread some sort of weird needle. In particular, it needs to go in for Claim 1, and then believe Claim 3 enough to conclude that "the sims get lied to in Claim 2" (but not enough to conclude that the sims see a Claim 2 about lion people) — and then reach the conclusion that in fact, Claim 1 is true, but Claim 2 and Claim 3 are false (though Claim 3 is "sorta true"), and you're probably a sim. But why reason this way in particular?

As an example of reasoning that requires a move like this, consider Chalmers (2022, p. 19 of the online appendix). Faced with an objection similar to the "self-undermining" objection Bostrom discussed above, Chalmers responds with a similar disjunction:

> We can reason: (1) either our evidence about computer power is heavily misleading, or it is not, (2) if our evidence about computer power is heavily misleading, we're probably in a simulation (as that's the most likely way for this evidence to be misleading), (3) if our evidence about computer power is not heavily misleading, we're probably in a simulation (by the original argument), so (4) we're probably in a simulation.

But this looks a lot like saying, in SIGNS THAT LIE ABOUT ANIMALS, "Either Claim 1 is false, or it's true. If it's false, then probably our sign is lying to us, which is the kind of thing that, according to the sign, happens to sims. But if Claim 1 is true, then probably we're in a sim—after, nine out of ten non-Bob observers are in sims." But (even setting aside the issues for "If Claim 1 is false, we're probably sims"), one feels like: "what about Claim 2?"

This sort of "selective skepticism" issue is closely related to why Type 1 simulation arguments fail. That is, Type 1 simulation arguments try to say that if we believe the whole sign, we don't know where we are. But this only works if our sign posits other observers with exactly identical observations (and if we reject indexical dogmatism). In all the other cases, including the real world, getting confused about where you are requires "forgetting" some information you thought you knew—information like "all the sims are squids." And it's not clear why we should forget this, but remember everything else (see Garfinkel (unpublished) and Birch (2013) for more on this).

This sort of issue also leaves appeals to "indifference principles" mired in epistemic muddiness. Thus, for example, Bostrom (2003) writes that his indifference principle applies even to observers with distinct observations, "provided that you have no information that bears on the question of which of the various minds are simulated and which are implemented biologically." OK, but how do we tell what "information" we "have"? Presumably, for example, Bostrom wants to say that I don't "have" the information that none of the sims see my bookshelf. But in FUTURE SQUID SIMS, do I "have" the information that all the sims are squids? If I do, what makes the difference? And if not, why not—and why do I still "have" the information that the ratio of sims to non-sims is X, or that there are sims at all?

Thomas (2021) is admirably precise about a similar problem. He defines his indifference principle relative to an "admissible reference class," where "admissible reference class" just means that, once you know that you're in the reference class, the rest of your evidence doesn't tell you anything further about whether you're a sim vs. a non-sim. For this to work in Future squid sims, the reference class needs to be pitched at quite a high level of abstraction—e.g., "observers who seem early in their civilization's history"—and we need to declare that "the sims are all squids" isn't part of our evidence. But why are we hollowing out our world model in this particular, intermediate way—a way that draws a bunch of non-sim humans with squid sims in their future, then erases the information about which species is where? Why do we believe Claim 1, in SIGNS THAT LIE ABOUT ANIMALS, but forget Claim 2?

One answer, here, is that *by the sign's own lights*, Claim 2s written on signs in this world are unreliable; but we don't have comparably undermining evidence with respect to Claim 1s. And I do think that something like this asymmetry drives our intuition, at least. If Sally tells us "Claim 1: p. Claim 2: q. Claim 3: I lied to everyone else about Claim 2," it's natural to think that p is at least a better bet than q. But we might also start to wonder about all of these claims—and indeed, about Sally.

What's more, we can run versions of the case that make believing Claim 1 even trickier. Consider:

SIGNS THAT LIE ABOUT ANIMALS AND NUMBERS: You wake up as a human in a white room.

A sign in front of you reads:

> "Claim 1: I, Bob, created nine sims and one non-sim, all in white rooms. Claim 2: The sims are all squids, and the non-sim is a human. Claim 3: If you're a sim, you see a different sign—one that lies, in Claim 2, about the animals involved in this situation, *and* which lies, in Claim 1, about the number of sims vs. non-sims."

Here, it seems especially silly to argue, on the basis of some kind of "indifference principle," that somehow "starting with the sign" should lead you to think that you're in a scenario with nine sims, and one non-sim, but where it's 90% that you're a sim. In particular: in *addition* to the sign specifying that the sims are squids, it also specifies that *the sims see signs that lie about the ratio of sims to non-sims*. If you believe Claim 3, then unless you think you're a non-sim, you can't look to Claim 1 to tell you the ratio of sims to non-sims. You need to rely more "on priors."

9 Are you a sim who just happens to be right about the basement?

My impression is that discussions of the simulation argument sometimes try to skate over issues in this vein, by focusing on cases that are as much like PLANNING ON INDISTINGUISHABLE SIMS as possible. That is, one often ends up talking a lot about sims whose basic scientific picture of their world (with respect to neuroscience, cosmology, computer science, etc) *just happens* to also reflect the truth about the basement: sims, that is, who are drawing "maps" that resemble the "true map" as much as possible (ancestor sims are a classic example, here). Such a focus makes it easier to think that a Type 1 strategy can work: everyone in the "reference class," you're encouraged to think, is drawing the same map, and the only question is "where you are."

But this sort of rhetorical strategy can end up seeming weirdly confident that if we're sims, we're sims *whose basic scientific picture just happens to accurately represent a basement universe we've never had contact with* (call these "epistemically lucky sims"). That is, it can feel like the argument is saying: "On your standard picture of the world, you had at least 30% of your 'things are normal' probability on a world with a high ratio of epistemically lucky sims 'like you' to basement non-sims like you, right? Ok, then, let's run CORE CONSTRAINT, and put almost all of that 30% on being an epistemically lucky sim."

Thus, if you started out with 1% on something as f***-ed up as being a sim, you end up at ~30:1 odds that your basic picture of the universe is still roughly right (though not *entirely* right—for example, your bookshelf is in a very different place; hence the "selective skepticism"). It's just that the universe you're right about isn't the one you see around you. Rather, it's somewhere else; somewhere you've never been.

But in addition to driving your credence towards a fairly arbitrary subset of possible sims (often, we could've easily run argument using a different set), this sort of move clashes with a different intuition: namely, that conditional on being a sim (or maybe, more broadly: conditional on being in a scenario as epistemically f***-ed up as living in a sim), we're unlikely to "just happen" to have a basically correct picture of the basement universe.

Now, it's not *crazy* to think that we're the lucky sort of sim. Indeed, lots of the fictional worlds *we* create (*Friends*, *Dunkirk*, even *Harry Potter*) bear a reasonably close resemblance to our own (though future advanced civilizations might be more imaginative). And it's not actually clear that the epistemic relationship that simulated cosmologists in an ancestor simulation have to the real cosmos is *defectively* "lucky." After all, if the simulators first looked at their basement cosmos, and then put the truth about it into the simulation, then in some sense the simulated cosmologists "would have seen different (simulated) cosmological evidence," had the basement cosmos been different (though we can also imagine cases where, instead, the simulators randomly generated different cosmologies—this looks dicier).

Still, it can feel like people who go around arguing that we're decently likely to be sims, on the basis of the empirical evidence they get from the world that surrounds them, are exhibiting some kind of blithe overconfidence in being sims who are right about the science of the basement universe (whether "luckily" or no). If pressed, they often fall back on the sort of disjunctive argument Chalmers gave above—namely, "well, either our science reflects the science of the basement, or it doesn't; and either way, we're probably sims." But then they go back to talking as though they're type of sim whose world lets them know what's going on in the basement. And this feels like it's in tension with the contrary intuition that with respect to knowing stuff about basement science, most sims are screwed (this is the sort of intuition that "Claim 3 says that if you're a sim, Claim 1 is lying" is meant to evoke).

Now, to be clear, Core constraint does not mandate this sort of "I'm an epistemically lucky sim" view. To the contrary, Core constraint just tells you that you're not allowed to believe that you're a non-sim in a High ratio world. So faced with empirical evidence suggesting that you're in a High ratio world, it's possible to simply to treat Core constraint a kind of "reductio" on the idea that your empirical evidence is reliable in this respect—rather than to conclude that in fact, you're probably a sim whose empirical evidence tells you about the situation in the basement.

What's more, to the extent you begin to move portions of your credence away from "my empirical evidence is reliable, and I'm a non-sim in a High ratio world," you don't (*pace* Chalmers) need to *give* that credence to "I'm a sim." You could, of course, give it to "sim blockers" (including, "my evidence is unreliable" sim blockers like "we're wrong about the compute required for sims, and/or how much future civilizations will have access to). But even if you don't go that route, there are all sorts of "my empirical evidence is unreliable" scenarios to consider—including various zany skeptical scenarios that you had initially ruled out (brain in vats, evil demons, solipsism, Boltzmann brains, etc). "I'm a sim" should be *in there*, sure. But let's not let the salience of sims, in this particular dialectical context, prompt an over-quick jump to: "conditional on my empirical evidence being unreliable, I'm probably a sim." That sort of claim requires substantively additional argument.

10 Type 2 arguments: "If that's your story, you should be taking other stories seriously, too."

With these problems and confusions in mind, I want to turn to what seems to me a better framing of the simulation argument, which I'll call the "Type 2" framing. I don't think this framing is actually very different from the standard presentation in the literature—indeed, I expect that many people will say: "dude, that's just what the argument has always been." And maybe so (I'm going to skip diving in too hard on exegesis). Regardless, though, amidst the morass of problems I've just discussed, I've found this particular way of thinking about the issue clarifying—and I think it has some under-appreciated implications.

Type 1 framings, as I've presented them, say, "If that's your story, you don't know where you are." Type 2 framings, by contrast, say: "If that's your story, you should be giving lots of weight to these other, *structurally similar* stories, too—stories that you might've mistakenly failed to seriously consider, but which reflection makes clear are comparably likely, on priors, to the type of story you wanted to tell." That is, where Type 1 framings try to get you *lost* in a given world you already believed in, Type 2 framings try to get you to *expand* the set of worlds you're considering. But the expansion in question isn't just motivated by a standard skeptical "what if?" Rather, it's motivated by plausible constraints on how opinionated, on priors, you're allowed to be about worlds that fit the same abstract mold.

11 What would Bertha think?

To see how this works, let's look at a simplified version of FLAWED ANCESTOR SIMS—one where the non-sim humans only run one ancestor simulation.

Suppose that you haven't yet been born into the world, or made any "observations:" rather, you're sitting on the fluffy clouds of the Bayesian heaven, waiting to be en-souled, and wondering what your world will be like. Your designated Guardian Angel—Bertha — approaches to give you a quiz. "Suppose I were to tell you," she says, "that the world fits the following description: there is one non-simulated version of the 21st century, and one extremely realistic (but still imperfect) simulated version. Both have the same number of people in them."

"OK...", you say.

"Now," she says, "consider the following bookshelf." Here she reaches into her pocket, and pulls out a photograph of my bookshelf. "Suppose that this bookshelf will be observed by a human-like creature named 'Joe' somewhere in this world. In A-Type worlds, this occurs in the non-sim 21st century; in B-Type worlds, it occurs in the simulated one. Are A-Type worlds substantially more likely than B-Type worlds?

You squint at the books. They look like standard 21st-century-ish books to you. Nothing about them suggests "sim" vs. "non-sim"—and nor does the name "Joe."

"No," you answer. "A-Type and B-Type worlds are roughly equally likely."

"Good," says Bertha. "Make sure your prior reflects that." And she walks away.

A week later is your "ensouling" ceremony. Bertha hugs you goodbye. You promise her that you'll stay true to your prior, and update only when the evidence warrants. Everyone cries. Then off you go, into the world. Unfortunately, though, once you're actually born, you forget all your long hours of Bayesian tutoring, including your conversation with Bertha. You wake up as a baby named Joe, in what appears to be the 21st century. As you grow up, you start filling out your map of the world, starting with what you see around you—but even as you try to keep a fairly "open mind," you spend most of your time on local detail, and you don't tend to consider hypotheses that aren't made very salient or practically urgent.

One day, you're sitting by your bookshelf, watching the news. The President of the stable global government makes an announcement: "Fellow citizens: for some reason, we've decided to make a binding commitment to definitely run one and only one super realistic ancestor simulation, if we make it to technological maturity (which it looks like we will). However, that simulation is going to be imperfect—and in particular, everyone's bookshelves are going to get lost to history." Awed at the scope and unexplained specificity of this ambition, you get out your map of the world. You look, briefly, at the picture of your house, your bookshelf, your friends, and yourself. Then, in the future of all those things, you start drawing a hazy blob labeled "super realistic ancestor simulation."

You feel a twinge of strange doubt. Won't those ancestor sims be unable to tell that they're in a sim? And if so, are you, maybe... But then you remember what the President said about bookshelves. You locate your own bookshelf on the map, and then add a little note by the ancestor simulation: "Does not contain Joe's bookshelf." Then you sit back in your chair and breathe a sigh of relief.

Suddenly Bertha appears in a clap of thunder. Furious, she jabs her finger at your map of the world. "What kind of world is this?" she asks.

In a flash, it all comes back to you—the eons in the clouds, the painstaking assignment of priors, the strange quiz. "Uhh...an A-type world?"

"Indeed," she says. "An *A-type* world. And what did we say about A-type worlds?"

"Um, uh...that they're roughly as likely, on priors, as B-type worlds?"

"Indeed," she says, then grabs a piece of paper from your desk. She scribbles a copy of your map's depiction of your house, your bookshelf, your friends, etc—but then, instead of drawing an ancestor simulation in the future of these things, she draws a dotted line around all of them, and then a label reading "Ancestor simulation," which she double-underlines. Then she draws, in the *past* of all these things, a hazy blob labeled "Historical, non-simulated 21st-century earth."

She steps back, and points to her new map. "What kind of world is this?" she asks.

You're getting the idea. "A B-type world," you say, sullen.

"Indeed," she says. "A *B-type* world. So on priors, how likely is *this* map to be accurate, vs. *that* one."

"About equal," you say.

"And since being born, have you gotten any evidence that would differentiate between them?"

You gulp: "No." All your observations occur in both of these worlds, the same number of times (and there are the same number of observers to boot, so SSA-ish worries don't get going).

"So how should your posterior probability on the first map compare to your posterior probability on the second?"

"They should be equal," you say.

"And what sort of probability were you about to place on the A-Type world?"

You gulp harder. "70%," you admit.

Bertha glares. "And what sort of probability were you about to place on the B-Type world?"

"Um...I wasn't really thinking about it."

"*Harr-umph*," she says. Your cheeks burn with shame.

So the idea, here, is that before seeing anything—and definitely, before going in for some kind of "standard story" — you and Bertha work to assign comparable priors to structurally similar worlds (e.g., A-type worlds and B-type worlds). And because these worlds all contain your observations (and the same number of observers overall), you won't, later, actually have any reason to favor some over others. So this imposes a tight constraint on your posterior probabilities—one quite similar to CORE CONSTRAINT above.

To see this more clearly, let's look at the case with multiple sims. Imagine that in Bayesian heaven, Bertha instead draws a map of a world with one non-sim earth, and nine ancestor simulations, labeled 1-9. Pointing at the non-sim earth and at the first ancestor simulation, she asks: "are Joe-bookshelf-observations any more likely to show up on the non-sim earth than in ancestor simulation 1?"

"No," you say.

Then she asks: "are Joe-bookshelf observations any more likely to show up in the first ancestor simulation, vs. the second?" And again, you say no—and you say the same about head-to-head comparisons between each ancestor simulation.

"Ok," says Bertha. "Let's call the world where the bookshelf shows up in the non-sim 'A1.' And let's label different 'bookshelf-in-a-sim' worlds according to the ancestor simulation that the Joe-bookshelf observations show up in. In B1, they show up in ancestor simulation 1; in B2, in ancestor simulation 2. It sounds like you've told me that an A1 is just as likely as B1, which in turn is just as likely as B2, and so on." You nod.

(The argument, here, is analogous to the following case. God flips a coin. If heads, he creates one person in a red jacket, or nine people with blue jackets. If tails, he creates one person with a blue jacket, and nine people with red jackets. Now consider a given set of "red jacket observations." Conditional on those observations occurring (or if you prefer: conditional on waking up with a red jacket), what should you probability be on heads? Both SIA and SSA agree on 10%, here—a verdict that becomes even clearer if you first imagine not seeing your jacket color, such that you stay at 50-50, and then update on your jacket color, which really seems like it should be information about the coin. As I discuss in my post on anthropics, this case is a kind of anthropic "square one"—though certain funky views, like Bostrom's take on Sleeping Beauty, start getting it wrong.)

And now, back on earth, if the President announces nine ancestor simulations, and you start to draw nine ancestor simulations on your A-type map, Bertha will also draw *nine* new maps—B1-B9. And she'll remind you that, back in Bayesian heaven, you said that each individual B-type map had the same prior probability as your A1 map. And thus, because your observations show up in all these maps, even after updating on your observations, you just *can't* go above 10% on the Joe-in-the-basement map. If you're at 5%, you should have at least 45% on Joe-in-the-sim; if you're at 1%, you should have at least 9%; and so on.

And once there are zillions of ancestor sims, Joe-in-the-basement becomes effectively impossible to put much weight on. Hence CORE CONSTRAINT.

12 What does this say about squid sims?

Perhaps all this sounds boringly familiar. Isn't this basically just the reasoning that underlies the standard "indifference principles" that simulation arguments rely on? Maybe—but often, these indifference principles are defined relative to a specific (and suspiciously arbitrary) reference class like "human-type experiences," so they get confused when they encounter cases like FUTURE SQUID SIMS. I think that focusing on underlying constraints on priors over structurally similar worlds helps bring out a more fundamental (and flexible) dynamic.

To see this, let's look at FUTURE SQUID SIMS more closely. What would Bertha say about this sort of case?

Well, imagine that back in the Bayesian clouds, Bertha had given you the following quiz: "Suppose that the world fits the following description: one planet of animal-1 simulates nine planets of animal-2, who each receive misleading evidence that they will later simulate animal-3 (in fact, they'll get shut off before then). And now consider two types of worlds. In C-Type worlds, humans simulate nine planets of squid-people, who wrongly think they'll simulate some other type of animal. In D-Type worlds, some other type of animal simulates nine planets of humans, who wrongly think that they'll simulate squids. Are C-Type worlds substantially more likely than D-Type worlds?"

Suppose you squint at the humans, and at the squid people. Nothing about humans makes them seem intrinsically much more likely to show up in an "actually sims squids" slot, vs. an "gets misleading evidence that they'll sim squids" slot. And nothing about squids makes them seem intrinsically much more likely to show up in an "actually gets sim-ed," vs. a "misleading evidence that they'll get sim-ed" slot. (Let's assume that the squid people are in some sense an equally realistic output of an evolutionary process in their simulated world. And let's set aside considerations like "squids are clearly such a ridiculous thing to simulate that if you find yourself planning to simulate squids and only squids, you should update towards being in a thought experiment.")

So you say to Bertha, "C-type and D-type worlds seem about equally likely."

"OK," says Bertha. "Now consider a given set of humans-apparently-simulating-squids observations. Is this set of observations any more likely to show up on the non-sim planet of a C-type world, vs. the first sim planet in a D-type world?"

"Nope."

"And what about on the first sim planet of a D-type world, vs the second?"

"Nope, same probability."

"So the prior on a C-type world with those observations is the same as the prior on each of D1, D2, etc?"

"Yep."

But now we run the whole story again: you forgetting your in-the-clouds prior and accepting the standard story, Bertha's fury, new maps scribbled on new pages, and so forth.

Now, perhaps you want to question whether C-type and D-type worlds actually are comparably likely, on priors. And indeed, I think it would be great to have a story on which they are not—a story that helps explain why cases like FLAWED ANCESTOR SIMS seem more intuitively compelling than FUTURE SQUID SIMS. My best guess explanation for this intuition, at the moment, is simply that FUTURE SQUID SIMS relies on a more abstract and hollowed out specification of the "structure" that both worlds have to fit. That is, "non-sim humans run super realistic ancestor simulations" pins down the relevant initial set of worlds much more than "animal-1 simulates animal-2, who receives misleading evidence about simulating animal-3." So it's easier to feel confident about what types of things are more or less likely to show up in which type of slots (though as I'll discuss below, the relatively likelihoods here ultimately don't matter much, if the number of sims is large enough).

Overall, though, I think that this kind of Type 2 argument allows us to validate and explicate my earlier argument that we should be saying the same type of thing about FLAWED ANCESTOR SIMS and FUTURE SQUIDS SIMS—though perhaps, less confidently (or perhaps not: see below). But this is also a pretty weird (and in my opinion, underemphasized) upshot. Naively, one might've thought, "all the sims are squids" would be a pretty good counter-argument to "I'm a sim"—and I doubt that simulation arguments would've gotten the same traction if they started with the squid case. But I think it's the same logic—and we should be clear about where it leads.

Another, closely-related advantage to Type 2 arguments is that they provide a justification for the type of "selective skepticism" that simulation arguments often get accused of. Such selective skepticism is mandated by the sort of prior you and Bertha fixed on, back in the clouds. To *refuse* such skepticism is equivalent to having some kind of ultra-strong prior on your species, or your bookshelf, or your light speckles, in particular, showing up in the basement of a sim-filled world, relative to structurally similar worlds where they show up in one of the sims. And Bertha's point is just: before making any observations, and before becoming ensconced in some kind of "standard story," such a prior would be unreasonable.

13 What if the cases aren't structurally similar?

Now perhaps you're thinking: "Joe, if I'm getting worried about being a sim in FUTURE SQUID SIMS, am I also supposed to be worried in cases like NINE SQUIDS, ONE HUMAN, too?" Recall that in NINE SQUIDS, ONE HUMAN, you wake up as a human in front of a sign saying that there are nine squid sims, and one human non-sim, and that everyone sees the same sign. That is, according to the sign, the squids are getting told that they're sims.

This case sounded a bit like FUTURE SQUID SIMS, insofar as the "standard story" was positing sims you know you're not. But it's actually more like:

> FUTURE SQUID SIMS WITH LITTLE TAGS: Same case as FUTURE SQUID SIMS, except that we're planning to give all the future squid sims little tags in their visual field reading "you're a sim."

Are we supposed to run simulation arguments in this sort of case, too?

Maybe, but it's going to get messier, and constraints related to "structural similarity" are going to do less work. In particular, we're going to have to assign priors to:

- *E-type worlds*, where animal-1 simulates animal-2, and then tells animal-2 that they're sims, vs.
- *F-type worlds*, where animal-1 simulates animal-2, and doesn't tell animal-2 that they're sims, and gives animal-2 misleading evidence that they'll one day simulate animal-3 and tell animal-3 that they're sims.

And it's less clear what you said to Bertha about this sort of comparison, back on the clouds.

We can say something similar in response to the charge that all this talk of priors is just going to take us back to the land of standard skepticism. Consider a solipsism world that features only your mind and its specific observations, floating in nothingness, with nothing to explain it; and compare this to a "normal" world where your observations are explained by a relatively simple set of stable physical laws, governing an external reality in which you're embedded. These worlds have a notably different structure, and it's not at all obvious what you're supposed to say to Bertha about how they compare. If you had some story, then, about why a solipsism world with your observations is less likely, on priors, than a "normal" world with your observations, you're allowed to keep telling it.

(As an example of such a story: with nothing to explain or constrain the observations, it seems like your prior on solipsism worlds is going to have to cover *every possible set of observations*—e.g., every sequence of "pixels" in a visual field the size of yours—in an unbiased way, such that once you update on having relatively stable and consistent observations, your credence on solipsism should take a massive hit. This problem plausibly arises for Boltzmann brains as well, though that case is a bit more complicated. And in some sense, it feels like a problem for all skeptical scenarios that start with the observations that a normal world predicts you'll make, and then transports them to a skeptical scenario that doesn't seem to favor any particular set of observations over others.)

So Type 2 arguments aren't trying to say that you need to assign comparable credence to all worlds that contain your observations, or that in some generic sense we need to keep non-trivial prior credence on various zany skeptical hypotheses. Rather, the claim is more specific: namely, that you need to assign comparable credence to *structurally similar* worlds that contain your observations. When we combine this with the large numbers of sims in "lots of sims" worlds, we end up with something very close to CORE CONSTRAINT.

14 Type 2 messiness

So overall, I find the Type 2 framing a helpful way of approaching the simulation argument. But it, too, has problems.

In particular, assessing "structural similarity" seems pretty messy, and dependent on the type of description you're giving of a world. Thus, for example, if you want, you can describe FUTURE SQUID SIMS WITH LITTLE TAGS as a world where "Animal-1 simulates Animal-2, who thinks some stuff about stuff" and leave it at that. In some sense, E-type worlds and F-type worlds both fit this structural schema. But we don't want to say that we are therefore required to give them comparable credence. And the same goes for attempts to subsume skeptical scenarios and normal scenarios under descriptions like "some observers think some stuff."

Clearly, some kind of taste is required here. Indeed, without it, we seem likely to run into the sorts of problems that plague attempts to be "indifferent" between different seemingly-symmetric hypotheses. Consider, for example, the "Box factory paradox":

> A factory produces cubes with side-length between 0 and 1 foot; what is the probability that

> a randomly chosen cube has side-length between 0 and 1/2 a foot? The tempting answer is 1/2, as we imagine a process of production that is uniformly distributed over side-length. But the question could have been given an equivalent restatement: A factory produces cubes with face-area between 0 and 1 square-feet; what is the probability that a randomly chosen cube has face-area between 0 and 1/4 square-feet? Now the tempting answer is 1/4, as we imagine a process of production that is uniformly distributed over face-area. This is already disastrous, as we cannot allow the same event to have two different probabilities [...]. But there is worse to come, for the problem could have been restated equivalently again: A factory produces cubes with volume between 0 and 1 cubic feet; what is the probability that a randomly chosen cube has volume between 0 and 1/8 cubic-feet? And so on for all of the infinitely many equivalent reformulations of the problem (in terms of the fourth, fifth,... power of the length, and indeed in terms of every non-zero real-valued exponent of the length). (From Hájek (2012)).

(See also the water/wine paradox, and others.) I haven't really looked into these problems much, but I would not be at all surprised if they start to bite Type 2 arguments pretty hard. If Bertha comes to you with questions about boxes, for example, naïve appeals to "structural similarity" are going to lead you into trouble fast. So perhaps you should be more broadly wary.

Now, Bostrom (2003) claims that his "indifference principle" is immune to these paradoxes, because it only applies to "hypotheses about which observer you are, when you have no information about which of these observers you are" (p. 8). And indeed, box-factory paradoxes aside, the type of reasoning at stake in your conversation with Bertha about FLAWED ANCESTOR SIMS sounds fairly good to me (and even more so, for PLANNING ON SIMS WITH DIFFERENT LIGHT SPECKLES). But note that the immunity Bostrom posits seems most solid if we lean into a subtle suggestion that the simulation argument is really a TYPE 1 ARGUMENT—that is, an argument in which we have fixed on a map, but don't know "who we are" within it. Actually, though, outside of cases with genuinely indistinguishable observers living in the same world, the simulation argument requires that multiple maps come into play—and sometimes, maps that start to look quite different (e.g., "humans simulate squids" vs. "dogs simulate humans"). Type 2 arguments try to wear this on their sleeve—but they are correspondingly open about a certain type of vulnerability, and appeals to "structural similarity" offer only limited protection.

To get a flavor of where problems might show up, consider a case like HOT AND HAPPY SIMS, where we plan to only make sims that are way hotter and happier than us. Faced with the chance to put credence on this "standard story," a Type 2 argument considers the alternative, structurally similar hypothesis that despite our acne and flab and terrible suffering, we are sims that are hotter and happier than some (extremely unfortunate) set of basement people. That is, it considers worlds that fit the structure: "least hot/happy people simulate more hot/happy people, who mistakenly think that they will simulate most hot/happy people." And it wonders, for the given absolute level of hotness/happiness we actually observe around us, about the probability, on priors, of that level playing the role of "least," "more," or "most."

But I expect that the only way to think sensibly about this question is to also have some kind of "absolute prior" over how hot/happy people *tend to be*, period, across all possible worlds. And this feels similar to the sense in which I expect that dealing sensibly with "Box factory" cases (I haven't looked into the standard resolutions in the literature) will require some kind of absolute prior over *what sorts of boxes factories make*, across all possible worlds. In particular: "level of hotness/happiness" is a continuous quantity, and continuous quantities seem like an area where epistemic procedures with an "indifference-y" flavor run into trouble especially fast. (We can say similar things about "levels of resolution" in Low res sims.)

I also expect that such "absolute priors" will be needed if we want our approach to simulation arguments to be able to handle cases that fit jankily with a "structural similarity" framing. Thus, consider:

> Sims who think they can't sim: The government is planning to make tons of sims, but to make it seem, to the sims, like making sims is impossible.

Here, we can, if we wish, describe this as a case where "Some people simulate other people, and deceive them." And we can wonder, conditional on the world having this structure, whether we should think it more likely that we are deceivers of the type described by the standard story—or whether, perhaps, we are (somehow) the deceived. But these hypotheses aren't nicely symmetric: if we are being deceived, it's in a different way than the sims we're planning to deceive (notably, for example, creating sims appears possible to us). So we need some "absolute prior" about which sorts of deceptions are likely to show up, where. (We try to swamp the messiness here by appealing to the numbers of sims, but this has an "epistemic pascal's mugging" flavor that I'll discuss in a moment.)

Indeed, ultimately, there's nothing special about "structural similarity." Appeals to it are just an attempt to get some grip on reasonable priors about things (thanks to Katja Grace for emphasizing this point in discussion). To really assess a given case, you need to go back to Bayesian clouds, talk to Bertha, and actually think about what sort of opinionated it's reasonable to be, before you make any observations. At the end of the day, it was about priors the whole time (how could it not be?).

In fact, I think that being clear about the role of priors here is useful, insofar as it helps shed light on how much to expect the philosophical dialectic surrounding simulation arguments—a dialectic that often focuses on "disjunctions" rather than overall probability assignments, and which gets mired in tangles about "selective skepticism," "admissible reference classes," "what you count as knowing," and all the rest—to inform your overall credences about being a sim (and especially, a sim of one type vs. another). For example: if you wake up in a world with some f***-ed up sign talking about how "(1) there are a zillion sims but (2) you're not one of them but also (3) I lied to all the sims and also (4) the sims are baseballs and also (5) you're not conscious but the sims are and also (6) actually there are a no sims unless (7) this sign is a lie," the thing to do, I expect, is not to start saying things like "well, either this sign is telling the truth, or it's lying," or "either this sign's claim about the zillion sims is true, or it's false," or "either I'm a sim, or I'm not." Nor, indeed, should you start thinking about what worlds would be "structurally similar" to some version of the world posited by the sign, to the extent there is one. Rather, the thing to do, for better or for worse, is to *do the hard, messy work of being a Bayesian*. That is, to have a prior over worlds, to think about what sorts of worlds would make observing signs like this more or less likely, and to update. It's the same old game—and even if they succeed, simulation arguments don't play it for you. Rather, they just rule out certain High ratio moves.

15 Wait, is this an epistemic pascal's mugging?

I also want to flag another worry about Type 2 arguments: namely, that their logic flirts very hard with (even if it doesn't strictly imply) a kind of "epistemic pascal's mugging."

To see this, let's first return to your conversation with Bertha about Future squid sims — the conversation in which Bertha asks you compare C-type worlds (worlds where humans simulate squids, who wrongly think they'll simulate something else) with D-type worlds (worlds where something simulates humans, who wrongly think they'll simulate squids). And suppose that for some idiosyncratic reason, you decide that you think C-type worlds are quite a bit more likely than D-type worlds (maybe, to you, humans just don't *seem* like a "gets simulated" type of species).

But now we notice: the claim that C-type and D-type worlds are comparably likely isn't actually necessary here. Rather, what matters is that, to the extent that C-Type worlds are x times more likely than D-Type worlds, and there are y times as many relevant sims as non-sims in both worlds, y needs to be much larger than x (see Ben Garfinkel's doc for more on this sort of ratio).

Suppose, for example, that you think C-type worlds are 10x more likely than D-type worlds (but you continue to admit that the conditional probability of a given set of human observations is the same on a C-type "basement humans" planet vs. a D-type "sim humans" planet). But let's say that the broader "animal-1 simulates animal-2, who mistakenly thinks they'll simulate animal-3" structure, here, involves one animal-1 planet running 1000 animal-2 planets (and each animal-2 planet thinking that they'll run 1000 animal-3 planets). So when you show Bertha your C-Type "standard story" map, she draws 1000 D-type maps, each with your observations in a different sim. You say: "Ok, but I said that each one of these D-Type maps is 10x less likely than my C-type map"— and she says, "Yes. But where does that leave you overall?" You frown. Hmm. You give 10 units probability to the C-type world, and 1 to each of the D-type worlds—so, 1000 overall to the D-type worlds. Sh**: it looks like you still have to have ~100x the credence on a D-Type world that you have on a C-Type world. And obviously, if there are zillions of sims, C-Type worlds are right out, even if you thought that they were way more likely on priors.

Now, on its own, and even in cases with the same number of observers, this kind of reasoning can lead to pretty weird places. Consider the hypothesis that you're living in a type of simulation that it *really* doesn't seem like you're in: for example, a futuristic cooking show (you're an absolutely terrible cook, and you don't pay any attention to cooking). Even if your observations are *way* more likely to show up on an early-history basement planet than on a futuristic cooking show, you can easily end up thinking that you're much *more* likely to be living in a futuristic cooking show than in the early history of a basement civilization that will go on to create sufficiently many, sufficiently diverse simulations (though obviously, if the simulations are sufficiently diverse, then living in a cooking show shouldn't be your overall best guess).

Indeed, if this civilization is big enough and does enough zany stuff, you can end up thinking it more likely that you live in some kind of wild, *non-simulated* skeptical scenario—for example, that you're a literal brain in a vat, or that you're in some sort of terraformed biological Truman-show situation, or that your planet was created by nano-bots a few thousand years ago with the appearance of age—then that you're in the basement early history of such a civilization. (Though at a certain point, we do need to start talking about the logistics these scenarios imply, and trying to get quantitative about probabilities.)

Maybe you think: yeah, well, that's basic probability theory for ya. But reasoning with this broad flavor, construed in a Type 2 way, gets even weirder when we start to move towards cases with different numbers of observers. To see this, consider:

> SIMS SEEM SUPER UNLIKELY: You live on 21st century earth. Modern science says that the universe is almost certainly finite and entirely devoid of life, except for humans. It appears to you that simulations are prohibitively computationally expensive to run, even for very advanced civilizations. The brain appears to work via crazy quantum microtubules that each have their own libertarian free will. Also, your stable global government has made a binding commitment to never run any sims, ever, and the universal human consensus, professed by all babies as soon as they can think, is that running sims would be a moral and epistemic horror. Also, there is a giant asteroid heading towards earth which will almost certainly kill everyone; and even if you make it through, scientists are confidently forecasting a vacuum collapse in a few centuries that will wipe out all life in the universe in an entirely correlated way.

Now suppose that you're back on the Bayesian clouds. Bertha asks you: "Consider a world where SIMS SEEM SUPER UNLIKELY is the whole basement story. Call this a G-type world. And now consider, instead, a world where an advanced civilization with tons of resources runs a zillion zillion simulations of apparently G-type worlds, plus tons of other stuff. Call this an H-type world. Which is more likely: a G-type world, or an H-type world?"

"Um... I dunno, they're not very structurally similar."

"Ok," says Bertha. "But an H-type world isn't, like, more than a zillion times less likely than a G-type world, right?"

"Fair enough. H-type worlds sound pretty weird, but a zillion-fold prior penalty seems over-confident."

"And you agree, I assume, that the conditional probability of a given set of G-type observations is roughly the same on a basement G-type planet vs. a simulated G-type planet in an H-type world?"

"Seems right."

But now, by the same logic as before, it looks like conditional on making observations like those in SIMS SEEM SUPER UNLIKELY, you're at least a zillion times more likely to live in an H-Type world than a G-Type world.

And here we might start thinking: hmm. In particular: is it really so impossible to just straightforwardly believe that the world is the way that our best science says it is? And if this is where the logic underneath the best form of the simulation argument ultimately leads, why were we bothering to talk about all that empirical evidence about brain compute and cosmology and Matrioshka brains? Why, indeed, were we even considering disjuncts like "everyone goes extinct before reaching technological maturity," or "everyone decides not to run sims," and looking to our own universe for evidence about them? If, no matter what, we're going to end up concluding that we're ludicrously more likely to be sims than to be non-sims in a G-Type world, then any evidence we get that points in the direction of an G-Type world wouldn't be enough of an update.

(Indeed, if we condition on being conscious (should we?), naïve versions of this sort of argument seem like they might allow you to rule out *metaphysical* hypotheses like "sims can't be conscious," regardless of the apparent strength of the arguments or the philosophical consensus, purely on the basis of the fact that G-Type worlds are ones where sims are conscious. And this seems an additionally strange move—one that ignores not just the empirical evidence, but the armchair evidence, too.)

In fact, we can run versions of this argument that seem like they could cause you to be much more confident you're in some other sort of non-sim skeptical scenario (e.g., a "demon farm" world, where zillions of evil demons deceive zillions of people into thinking that they're in G-type worlds; or a "world-created-five-minutes-ago-with-the-appearance-of-age farm" world) than that you're in a bona fide G-type world. And we can do this across a very wide range of evidential situations. It can feel like traditional skepticism has struck again, pascal's-mugging style. Maybe you don't, ultimately, expect to be in a demon farm world: but if these arguments go through, it ends up a better bet than "things are normal," for many readings of "things are normal" that we previously thought believable.

That said: let's not be hasty. In particular: as typically stated, the simulation argument studiously avoids making any comparisons between scenarios with different numbers of observers (this allows Bostrom's "weak principle of indifference" to remain anthropically innocuous, since SIA and SSA agree in equal-numbers cases). And we can try to limit the scope of Type 2 reasoning to reflect this sort of restraint.

Indeed, worlds with different numbers of observers have, built in, an important type of structural dissimilarity—one that leaves it open, in some sense, what sort of prior to place on them. Thus, you can, if you wish, give H-type worlds a *zillion zillion zillion*-fold prior penalty, relative to G-type worlds—and it's not clear that Type 2 arguments have any grounds for objection.

And beyond this (or perhaps, to justify it), you can also try to wheel in other sorts of anthropic tools to punish H-type worlds. Maybe you start saying SSA-like things about disliking big worlds when you know where you are (though: do you, in Bayesian heaven? And I don't, personally, like saying SSA-ish things). Maybe you start saying UDASSA-ish things about the bits it would take to specify an H-type world (though: what if the physics of the H-type world is pretty simple?), or one of the observers within it. Or maybe something else I'm not thinking of, but which readers can suggest.

Indeed, while I haven't explicitly invoked SIA, and I don't think that Type 2 reasoning strictly requires it (one could try to imagine an SSA-ish conversation with Bertha), the type of "epistemic pascal's mugging" argument I've just given bears a very close resemblance to the "Presumptuous Philosopher"-type cases that plague SIA. And no surprise: I'm a bit of an SIA type, and the "Type 2" reasoning I've been describing feels to me spiritually similar to what currently seems to me like the best justification for SIA-ish reasoning: namely, something like "more people like you" → "more opportunities to generate your observations."

Thus, for example, what makes it so much more likely that you're a sim in a "single something-animal planet simulates 1000 human planets" world than in a "single human planet simulates 1000 squid planets" world is that whether simulated or no, each of these "human planets" is ~equally likely to generate your observations (enough to outweigh any difference in prior). And this sort of reasoning has a very natural (but presumptuous) extension to comparisons between "single something-animal planet simulates 1000 human planets, who simulate no one" vs. "single human planet simulates no one." So while it's possible to ignore this extension and focus on more innocuous forms of Type 2 reasoning (forms sufficient to get claims like Core constraint going), I worry that this pushes under the rug something that indicates a problem.

(Also, certain applications of "structural similarity" involve positing different numbers of observers. Thus, suppose that it looks like we have the resources to run a million sims, and we're planning to give each of them evidence that makes it look like they can only run a thousand. How should we think about the possibility that actually, in the next level up, someone else is running a billion?)

This type of "epistemic pascal's mugging" issue currently leaves me wary, overall, of how fast Type 2 reasoning might just run into SIA-style problems; and about the issues in this space more generally.

16 Expected ratios of sims to non-sims

We see a very similar sort of "epistemic pascal's mugging" flavor in a recent and more provocative formulation of the simulation argument, offered by Thomas (2021), which attempts to reach the surprising conclusion that actually, we are *definitely* sims—because conditional on being non-sims, the chance that we go on to make zillions of sims is non-trivial.

Thomas's argument (as I understand it—it's possible I'm missing something) works via the following basic dynamic. Consider three hypotheses:

1. Humanity creates zillions of sims, and we're non-sims.
2. Humanity creates zillions of sims, and we're sims.

3. Humanity does not create zillions of sims, and we're non-sims.

(We can also add "4. Humanity does not create zillions of sims, and we're sims"—but I'll set that aside.)

If we accept the logic of the simulation argument (Type 2 will work for this), then the probability of 2 needs to be a zillion times larger than the probability of 1. And if we accept our standard empirical picture of the world, then the probability of 1 *isn't* a zillion times smaller than the probability of 3: that is, conditional on us being non-sims, the idea that we go on to create zillions of sims isn't totally out of the question. Thus, the probability of 2 needs to be close to a zillion times larger than the probability of 3—and thus, since 2 also swamps 1, of 1 and 3 combined. That is (and even including 4), we have to be sims.

It's a cool argument. But naively, one thinks: huh? Even assuming that humans are alone in the universe, on my standard empirical picture of the world—a picture on which it seems *totally possible* that humanity *doesn't* go on to create zillions of sims—I can somehow conclude that actually, either it does, or we're sims in some other even weirder situation? (And note that Thomas's argument doesn't invoke some particular anthropic theory like SIA, which prompts such "huhs?" all over the place anyway.) And when we consider scenarios like Sims seem super unlikely, this sort of "huh?" gets all the stronger.

Suppose, for example, that the super-duper forecasters in the Sims seem super unlikely world are all "standard empirical picture types" who condition on not being sims, and they say that there's a one in zillion chance that actually, all this anti-sim evidence is misleading, and we'll go on to create a zillion zillion sims (and don't even get me started on the probability that we get hypercomputers and create *infinite* sims). Assuming that the forecasters are perfectly calibrated at this level of precision (yes, I know, the case is very realistic and has no problems), does that basically wrap it up? Can the asteroid-prevention agency relax its efforts? Should the global government start worrying about a pro-sim coup?

I don't think so. The problem is that Thomas's argument requires that we accept the lesson of the simulation argument in one breath, but fail to learn this lesson in the next breath. In particular: if we buy the simulation argument's core constraint that our credence on 1 ("zillions of sims, but we're non-sims") needs to be a zillion times smaller than our credence on 2 ("zillions of sims, and we're sims"), then *we shouldn't try to keep our standard empirical picture of the world*, even conditional on not being sims.

Thus, suppose that prior to learning about the simulation argument, you were at ~100% that we're not sims, and ~1% on "we create zillions of sims" (so, ~1% on 1, and ~99% on 3). Then you learn that your credence on 1 needs to be a zillion times smaller than your credence on a new hypothesis, 2. Ok, well, you can't give 2 a *zillion percent credence*, so you're going to have to make some adjustment to your overall probability distribution. But nothing mandates that this adjustment keep the ratio between 1 and 3 constant, or even near-constant. To the contrary, getting *radically* less excited about 1 seems a much more attractive response (especially in cases like Sims seem super unlikely) than shrinking both 1 and 3 by a similar factor, to make room for certainty on 2.

Yet Thomas's argument rests on the idea that conditional on not being a sim, you've kept dream of 1 alive. And if you do this, and you accept CORE CONSTRAINT, then course you're certain you're a sim (and I'm happy to grant Thomas's argument in this conditional form: if CORE CONSTRAINT is right, then either 1 is off the table even conditional on being a non-sim, or you're a sim). But whole point of CORE CONSTRAINT is to kill the dream of 1. Trying to hold on to it is a route to sim-hood.

Of course, killing the dream of 1 is its own form of "huh?"—and it has its own flavor of philosophical presumptuousness. But it's less presumptuous, I think, than being certain you're a sim in SIMS SEEM SUPER UNLIKELY.

What's more: if we look closely, we can see that the type of "epistemic pascal's mugging" Thomas's argument evokes is actually quite similar to the H-type, "what if an advanced civilization is simulating a zillion zillion worlds where sims seem super unlikely" problem I discussed in the previous section. The key step, in both cases, is to set up a "not ludicrously more likely than" comparison between a SMALL BUT PLAUSIBLE world—that is, a world where the SIMS SEEM SUPER UNLIKELY evidence is correct—and some other, much less plausible world that features tons of sims. In Thomas's case, this is a world where we're not sims, but we create them; in the previous section's case, it's any given H-type world, where our observations show up in one sim in particular. And from there, in both cases, the rest is just a matter of showing that this particular implausible world is just as likely as each of a zillion others, where your observations show up somewhere else. Thus the SMALL BUT PLAUSIBLE world disappears.

17 Sims in big worlds

Thus far, I've been focusing on cases in which humans are (apparently) the only species in the universe. This assumption makes things simple—but relaxing it also gives "I'm a sim" worries more bite.

In particular: if there are many civilizations in the universe, there are more opportunities for the universe to fit a schema involving the claim "someone runs a zillion sims." And this sort of schema gets troubling (though note that the number of basement civilizations matters as well).

Actually working through this gets a bit complicated, especially once we start talking about worlds where the humans have made a firm commitment to never run sims (and there's an asteroid heading towards earth regardless), but it seems like e.g. the lizard people in the next galaxy over have a 10% probability of running a zillion simulated lizard cooking shows, featuring lizard sims with little "you're a sim" tags in their visual field (how much less likely are you to be a basement human in that sort of world, than a sim in some vaguely-but-not-totally structurally similar world?). I'm going to skip these sorts of issues for now. But they're an important counterpoint to "maybe we just go extinct," or "I dunno, maybe humans just decide not to run sims."

I'll note, though, once we start talking about *infinite worlds* (which, if we're getting into

SIA-ish vibes, or if we just like various mainstream cosmologies, we will), I basically expect everything to get super gnarly and probably break. Bostrom "deliberately sets aside" infinite worlds in the original paper (see his section 6 discussion in the FAQ), but suggests in the FAQ that we might appeal to something about the limiting fraction of sims vs. non-sim observations in expanding hyperspheres. My impression, though, is that making moves in this vicinity work is an unsolved problem (albeit, one that cosmology has to face more generally).

That said, I don't think we should say something like: "Stuff breaks in infinite worlds, so we can just believe whatever we want." Indeed, if we *do* live in an infinite cosmology, there is presumably some answer to the various epistemic questions that the "measure problem" implicates—and one that we might expect will capture the type of "normality" that we (hopefully) do OK on when we act like we're in a finite case (thanks to Carl Shulman for discussion).

Still, breaking on infinite worlds is another note of caution.

18 Lessons from Boltzmann brains

Also, there's this stuff about Boltzmann brains: i.e., observers generated by random fluctuations in a sufficiently big world, which make all possible sets of observations some very large number of times. Boltzmann brains raise issues notably similar to sims. In particular, they fall out of various seemingly-plausible "standard stories," but positing them involves sharing the world with some very large number of observers making observations similar to your own, except in a wacky skeptical setting—thereby prompting the concern that conditional on such standard stories, you're overwhelmingly likely to be in such a wacky setting. Indeed, with Boltzmann brains, Type 1 arguments actually get a grip—in a Boltzmann brain-y world, there are multiple observers making exactly your observations.

I haven't looked into this, but my impression is that options re: Boltzmann brains include:

1. Continuing to believe in Boltzmann-brain-favoring cosmologies, but penalizing the hypothesis that you're a Boltzmann-brain on the basis of the fact that your experiences are coherent/stable/non-disintegrating. (But then: aren't there zillions of Boltzmann brains with these memories of coherence, who are making this sort of move too?)
2. Updating away from Boltzmann-brain-favoring cosmologies, in virtue of the fact that your experiences are coherent/stable/non-disintegrating. (This is maybe my favorite option, due to it plausibly falling out of spreading out my credence on a Boltzmann brain world between all the Boltzmann brains with my observations in that world, and then all that credence getting cancelled when I don't disintegrate. But it also feels a bit empirically presumptuous—and obviously, in a Boltzmann-brainy-world, tons of Boltzmann brains are making this sort of update.)
3. Updating away from Boltzmann-brain-favoring cosmologies in virtue of the fact that believing in them is "cognitively unstable." (This is the route favored by Carroll (2017).

It's similar to trying to say, about sims, that "sure, my empirical evidence suggests that I'm a sim; but if I were sim, then this evidence wouldn't be reliable—so I guess, somehow, this means I'm not a sim?"—though it adds the claim that "and therefore, my empirical evidence is unreliable." And we might wonder, as we do with sims, whether it could make sense to focus on "epistemically lucky Boltzmann brains." And also: is that a way to do cosmology? Sounds a bit philosophy-ish to me...)

4. Trying to devise (or update towards) a funky anthropic theory according to which, in a universe will way more Boltzmann brains than normal observers, you're *still* more likely to be a normal observer—perhaps because the Boltzmann brains are stuck in the library of babel, and hence are harder to "find" by the arbitrary Universal Turing Machine that determines who you are likely to be to. (But this view has lots of its own weird problems.)

Presumably there are other options as well—and perhaps readers who've thought more about it can inform me of their own confident solutions to the Boltzmann brains problem. And indeed, I'm optimistic that there's some solution, here. In fact: I find it ~impossible to actually get worried that I'm a Boltzmann brain, and that my experience is about disintegrate: that thing where this sort of view just keeps making the wrong prediction seems too compelling (yes, yes, the Boltzmann brains say that too: and yet, at the end of the day...). But "this sort of view keeps wrongly predicting that I'm about to disintegrate" isn't something we can say about sims (though actually, if we start thinking that sufficiently many sims are short-lived...)

Still, sim worries and Boltzmann brain worries have clear structural similarities. Indeed, one can imagine framing options 1-3, here, in terms of a "Core constraint" that either you don't live in a Boltzmann-brainy-world, or you're probably a Boltzmann brain. And one can imagine arguing against 4 on the basis of whatever arguments motivate a sim-like indifference principle. Indeed, conversely, we can imagine (though I don't feel very tempted by this) trying to tell 4-like stories about why we're not sims—stories on which sims are somehow intrinsically harder to "find" than basement people (though note that we don't have the library of babel to appeal to).

But I also want to point at a higher-level similarity: namely, that even if you don't currently see a flaw with some argument for putting significant credence on being a Boltzmann brain, this is the kind of conceptually-tricky, suspiciously-clever, specific-thing-some-people-just-recently-thought-up, big-number-driven, weird-conclusion argument that it seems wise to approach with some caution—and to avoid swallowing whole after some cursory encounter. And I want to say the same thing about the simulation argument. Maybe you think that one is substantially stronger than the other; or maybe not. But they trigger, for me, some similar heuristics.

19 But aren't I improbably cool?

I want to close with a brief discussion of a *different* type of simulation argument: namely, that even conditional on finding yourself apparently living in earth's early history, if you find that you are an "unusually something" person (e.g., unusually influential, entertaining, successful, etc) by "earth's early history" standards, then this is some sort of important *additional* update in favor of being a sim. Bostrom mentions this type of argument in various podcasts; and see also Chalmers on "sim signs."

But how, exactly, is the argument supposed to run? In the context of the classic simulation argument, for example, the update at stake easily falls into the noise (thanks to Carl Shulman and Paul Christiano for discussion). Thus: suppose that you're Donald Trump. And suppose that conditional on no sim blockers, you expect that humanity will go on to run a zillion "all of earth's early history" simulations (each featuring one Donald Trump), *and*, let's generously say, a zillion "Donald Trump" simulations for *each* ancestor simulation (so: ~a zillion zillion Trumps in total). Call this an "Ancestor sims and Trump sims" world.

It's true that, per Core constraint, Trump has to think that it's at least a zillion zillion times more likely that he's a sim than that he's basement Trump in an "Ancestor sims and Trump sims" world. But compare this conviction with the "I'm a sim" credences of a janitor in 21st-century Des Moines. The janitor should think he's at least a *zillion* times more likely to be a sim than to be a 21st-century janitor in an "Ancestor sims and Trump sims" world—which is already a lot.

What's more: when we step back to look at their overall credences, here, it doesn't seem like Trump's extra zillion is going to make much of a difference. Suppose, for example, that both of them put 60% credence on "sim blockers." In that case, if we accept the basic structure of the original argument, they have to each give ~all of their remaining credence to "I'm a sim." It's true that technically, the ratio of the remaining tiny sliver that the Janitor is allowed to keep on "No sim blockers and I'm in the basement" vs. "I'm a sim" differs from Trump's by a factor of a zillion—but regardless, both of these slices of probability pie are squeezed into the 40% on "no sim blockers." So we're fighting over slivers. And if we lower their probability on "sim blockers," or get rid of it entirely, the same argument holds.

Indeed, this argument holds in pretty much the same way even if there were, like, *a hundred* (or even ten) full ancestor sims, and a zillion zillion zillion Trump sims. Just the hundred is enough to get the janitor struggling to put much weight on being a non-sim in an "Ancestor sims and Trump sims" world. Trump's extra "no way" is small potatoes.

Now, perhaps you could try to get some other active argument going that Trump should have a lower probability on "sim blockers" than the janitor. But it's hard to see how this could fall out of Core constraint, at least. Core constraint, after all, just tells you that you can't think that you're a non-sim in a High ratio world. Once both Trump and the janitor learn that lesson, the simulation argument doesn't tell them where to go from there.

Maybe we try to say something SSA-ish about reference classes? E.g., in a "no sim blockers" world, Trumps are a much larger fraction of the reference class of "people overall"? But if we're doing SSA-ish reference-class stuff, we're also plausibly doing Doomsday-ish updates *towards* sim blockers—especially given that in a world with a zillion zillion

Trump sims, there are presumably ludicrous numbers of post-human-God-knows-whats (though: are they in the reference class? bleh, bleh.)

Maybe we say something SIA-ish? E.g., there are more Trumps in non-sim-blockers world, so Trump updates hard towards such worlds, and becomes certain he's in them, and therefore a sim? But then: so, too, does the janitor.

Or consider the following argument, which is closest to my best-guess about what sort of argument is actually psychologically operative. Hypothesis 1: "Things are normal." Hypothesis 2: "Something weird is going on." Probability of being Trump conditional on hypothesis 1: low. Probability of being Trump conditional on hypothesis 2: higher I guess? Thus, being Trump is an update towards "something weird is going on." And what's the most salient way something weird might be going on? I dunno, something about simulations? That's what my friends talk about, anyway.

There's presumably more to say, here, and more options for trying to reconstruct a form of reasoning that would justify the type of thought above. I mention this reconstructive task partly because I have some hazy suspicion that for some subset of people engaging with simulation stuff, "aren't I improbably cool" arguments for being a sim exert a psychological force that substantially exceeds their philosophical credentials—especially once we've taken into account the classic argument, and the bare fact that you live in earth's early history at all. Maybe there's something importantly extra there (and *conditional* on being in a sim, being Trump might tell you something about what type): but I encourage advocates to actually formulate it, and to subject it to scrutiny.

20 Against simulation woo

> *"It feels like the whole world revolves around me somehow."*
>
> – Truman

This scrutiny seems important partly because I think simulation stuff can throw otherwise sensible people back into some kind of hazily superstitious and "woo-y" mode—and I mean woo-y, here, in the bad way, not the "taking spirituality seriously" way. I mean woo-y, that is, in the sense evoked when X tells you that it's "too much of a coincidence" that she just happened to run into Y at that place; or the way a friend of mine used to talk, in a certain tone, about "weird things happening" related to someone who had died; or the way you might conclude, when everyone on a plane dies horribly except your daughter, that God is good and loves you extra special. It's a woo-y-ness that starts, subtly or not-so-subtly, to bend the raw and vast otherness of the world back towards the cramped and janky narratives of the self; to make the world smaller, so that it better fits some story; a story centrally structured by various needs and desires and random associations; a story, that is, that you just casually made up—with the self, suspiciously, near the center.

In some sense, the impulse is understandable—the possibility of being in a sim throws the "what sort of situation are we in" doors open quite a bit wider than the standard fare.

It's not an area we're used to thinking about clearly and seriously, and it's natural for a variety of (sometimes questionable) psychological and epistemic impulses to rush in to fill the gap—especially given the analogies with various types of religious thinking. And this laxity finds fertile ground in the way the topic can feel, socially, like the type of "weird stuff" area where one is licensed in saying or thinking whatever one pleases, halfway as a joke—but perhaps, in the back of your mind, with a more subtly real "what if?"

Indeed, for a long time, my most visceral reaction to discussion of simulations was some combination of suspicion and repugnance, prompted by the sense in which the topic seemed to harness, amplify, and excuse various impulses to process the world through a certain kind of self-oriented lens—a lens that seemed to me not just false, but *really importantly* false; not just false, but the polar opposite of true; false about the core thing; false like a pilgrim walking in the wrong direction.

Here I think of a quote from Lewis, on describing his early attitude towards Supernaturalism:

> In order to breathe freely I wanted to feel that in Nature one reached at last something that simply *was*: the thought that she had been manufactured or 'put there,' and put there with a purpose, was suffocating...To find that it had not simply happened, that it had somehow been contrived, would be as bad as finding that the fieldmouse I saw beside some lonely hedge was really a clockwork mouse put there to amuse me, or (worse still) to point some moral lesson. The Greek poet asks: "If water sticks in your throat, what will you take to wash it down?" I likewise asked, "If Nature herself proves artificial, where will you go to seek wildness? Where is the real out-of-doors?" To find that all the woods, and small streams in the middle of the woods, and odd corners of mountain valleys, and the wind and the grass were only a sort of *scenery*, only backcloths for some kind of play, and that play perhaps one with a moral—what flatness, what an anti-climax, what an unendurable bore!

This isn't quite what I found repugnant about simulations, but it's related—some compromising of the world's otherness, it's "out there"-ness, which had been and is, for me, so central—both ethically and epistemically (though as ever, like it or not, the "what's actually true" bit doesn't care what you vibe with).

From another perspective, though, this is a strange reaction: if you're in a sim, the world is in some sense quite a bit *more* "other" and "out there" than you might otherwise have thought. The raw wind still howls—louder, indeed, than seems at all comfortable. But it does so in some larger and even stranger space—a space, perhaps, that you can see less of, but which is no more small and "about you" for that. Quite the contrary.

21 Wrapping up

I haven't covered the topic of what it would make sense to *do*, if we're sims, or if we start to put real non-ha-ha credence on being sims. For many, this is the most pressing question. Indeed, it's a question a certain type of person wants to ask *before* engaging with any of the epistemics—and hopefully, to find, fast, that the topic can be dismissed on the grounds that it lacks any practical upshot.

This kind of "Come come, does this actually make any difference to anything? I knew it: no. Great, I'll ignore it," reaction has always seemed strange to me. Serious arguments for ridiculous reorientations in your understanding of your overall existential situation do not come along every day. This doesn't mean we have to jump in and just buy this argument (as I've tried to emphasize, I think various of the philosophical issues here remain gnarly—and even if the argument seems to you airtight, there is wisdom, I think, in moving slowly and carefully through this sort of terrain). But ignoring it entirely until it tells you to buy different groceries seems, to me, to betray some lack of existential curiosity.

What's more, "either sim blockers, or we live in a computer simulation created by some advanced civilization" seems like the kind of thing that might, actually, be useful to know, if it's true. It's a bit like the way knowing foundational mathematics or physics can be useful later on—except, plausibly, more obvious. And indeed: there are discussions out there of what sorts of immediate practical upshots simulation hypotheses might have (see e.g. Hanson (2001), Tomasik (2016), Greene (2020)). To me, the most interesting questions have to do with humanity's long-term future—but these also bring up a bunch of additional issues in ethics, decision theory, and anthropics, which I haven't really worked through, and which I won't try to bite off here.

For now, I'll simply note that as with the other especially weird posts on this blog (e.g., 1, 2), but maybe more than ever, I'm feeling very bad if someone comes to me and tells me that they're doing a bunch of weird stuff—and especially, that they're doing a bunch of less altruistic stuff, or stuff less oriented towards protecting humanity's long-term future, or stuff that exhibits less basic sanity or psychological stability or personal responsibility—because of something about simulation arguments. Indeed, I feel pretty bad about learning that a bunch of people are going to spend a bunch of time that would've otherwise gone to altruistic pursuits thinking about simulation arguments instead. I wrote this post as a personal (and to some extent, academic) project, and I think that the topic raises interesting, substantive, and potentially important questions. But I don't think that "getting to the bottom of it" should be a top priority — not with the other problems staring us in the face (and see Chalmers (2022) on why problems in sims are just as "real" and ethically pressing).

Indeed, even if you start to think that you're decently likely to be a sim, I think there's an altruistic case for "being the basement person you would've wanted to see in the world"—both from an acausal perspective (correlating with the true basement people, and/or fulfilling commitments you would've wanted yourself to make about what you would do, given "early history" observations), and from a more straightforward causal perspective (if you're in an influence-compatible early history basement after all—for example, because of some sort of sim blocker, or because swallowing the simulation argument whole wasn't, actually, a good idea—you can plausibly do an especially large amount of good). This isn't to say that thinking you're decently likely to be in a sim would make no difference (and this isn't a topic I've thought much about)—but I think it might well add up to more "normality" than a cursory look would suggest.

Finally: regardless of whether you buy simulation arguments or not, they are a reminder that the world we see and take for granted is only a part of the world—only, ultimately, a certain type of "zone"—and that in principle, the overall "situation" could in fact be many different ways, not all of which we are accustomed to considering. We do, in fact, need "priors"—and indeed, capacious ones, adequate to include worlds that are bigger and stranger than some standard story (is it so standard, if you step back and look?). Dealing well with such worlds is a delicate art. But it's one that the simulation argument, whether sound or not, reminds us to learn.