Simulation arguments

Joe Carlsmith

September 2022

I Introduction

Call someone who lives in a computer simulation a "sim," and someone who does not, a "non-sim." Some argue that we should have high confidence that at least one of the following is true: either it's not the case that most people with experiences broadly similar to our own (for example, experiences of living on something like 21st century earth) are sims, or we are sims.[1]

This essay formulates and defends what I see as the strongest version of this argument, and then explores some of the complexities and uncertainties that it leads to. I begin by describing the original argument in Bostrom (2003), and by offering a more general set of conditions that an argument like Bostrom's should satisfy. I then distinguish between two versions of such an argument—a Type 1 version, which rests on various empirical claims (in particular, about the computational power available to advanced civilizations); and a Type 2 version, which does not.[2] Bostrom's is a Type 1 argument.[3] But Type 1 arguments face objections—in particular, "selective skepticism" objections (why accept the relevant empirical claims about computational power, but not other common-sensical claims that would provide evidence that we're not sims?), and "self-undermining" objections (if we're sims, why think that our evidence supports the relevant empirical claims, applied to the reality simulating us?).

These objections may be answerable. But in my opinion, they distract unnecessarily from the core of what makes simulation arguments interesting. Loosely stated, this core (as I formulate it) consist of two claims: (1) condi-

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

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

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

tional *only* on having e.g. 21st-century-experiences and on living in a world where most people with such experiences are sims, our credence on being sims should be high (I call this claim "Observer-class indifference"); and (2) adding in the rest of our evidence, while continuing to condition on most people with such experiences being sims, shouldn't change this credence (I call this claim "Admissibility"). Type 2 arguments focus on these core claims directly, and they treat questions about the likelihood of various empirical claims as secondary (though obviously, important to our credences overall). For this reason, I suggest, Type 2 arguments are superior to Type 1 arguments, and less vulnerable to the objections above. More importantly, though, I think they have real force—force I attempt to elucidate.

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

- *How should we adjust our overall credences—including our credence on being sims—in light of these arguments?* Here I describe a number of options, all of which have problems. I then focus in particular on an argument from Thomas (2021), to the effect that we should be basically certain that we're sims, because conditional on being non-sims, the expected ratio of sims to non-sims is high. I suggest that we need not accept this conclusion, but that it points at the need for caution in trying to preserve common-sensical credences *conditional on being non-sims*, if we accept Type 2 simulation arguments.
- *To what range of sims and scenarios do Type 2 arguments apply?* Here I suggest that these arguments apply to a wider (and stranger) range of cases than the literature has focused on. For example, I suggest that they prevent you from giving non-trivial credence to being an early-history non-sim human, living in a world where most people with early-history experiences are simulated squid-people with tentacle arms (and that this is true even though your evidence is incompatible with being a squid-person). Charting the full limits of the Type 2 argument's applicability, though, looks challenging.

I close by briefly mentioning a few other outstanding uncertainties—about infinite worlds, about the implications of anthropic principles like the Self-Indication Assumption ("SIA") and the Self-Sampling Assumption ("SSA") for simulation arguments,[4] and about the practical implications of taking such arguments seriously.

[4]See Bostrom (2002a) for an overview of these principles.

II Bostrom's argument

Bostrom's (2003) simulation argument is the most prominent in the literature. He introduces it with two central assumptions: first, that suitably sophisticated sims can have conscious experiences like ours; and second, that technologically advanced civilizations (Bostrom calls these "post-human") would have enough computational power to run enormously many such sims with a trivial portion of their overall resources.[5] Bostrom justifies the first assumption by appeal to what he calls the "substrate independence" thesis in the philosophy of mind, and the second assumption via appeal to additional arguments about the limits of computation, the resources available to advanced civilizations, and the computation required to simulate a human-like mind and the necessary environment.

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

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

(Bostrom's argument for this disjunction is not airtight, but it won't be my focus here.)[6]

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

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

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

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

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

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

Thus, concludes Bostrom, we should have high credence that at least one of 1, 2, or SIM is true.[7] In the original paper, he suggests splitting our credence evenly between them; and in a 2008 FAQ, he assigns SIM "something like in 20%-region, perhaps, maybe" (though in more recent interviews he explicitly "punts" on probabilities).

III Set up

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

But this calculation isn't airtight. For one thing, as Chalmers (2022, appendices) points out, in principle post-human civilizations could also create many *non-sims* with pre-post-human experiences (for example, non-sims on terraformed planets, non-sim brains in vats, etc)—non-sims that this equation wouldn't capture. What's more, as Bostrom and Kulczycki (2011) point out, it could be that H, the average number of people that live in a pre-post-human stage of a civilization, is much larger for civilizations that do not end up creating ancestor sims than for the ones that do. Thus, for example, if there are two civilizations, A and B, and A has a pre-post-human population of 10 people, and goes on to create 1000 ancestor sims of 10 people each (so, 10,000 sim people total), but B has a pre-post-human population of a billion people, and it goes on to create no ancestor sims, then the equation above gives the wrong result for f_{sim} (f_p is 1/2, f_i is 1, N is 1000, so the equation says that f_{sim} should be 500/501, but actually it's ~1/100,000—i.e., 10,000/(10,000+1B+10)). Bostrom and Kulczycki (2011) discuss "patches" meant to address this issue. For my purposes, though, and especially for Type 2 arguments, the details of how we calculate f_{sim} are not central. Rather, what matters is the probability of the more basic disjunction: either it's not the case that $f_{\text{sim}} = \sim 1$, or we're sims.

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

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

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

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

I'll approach the questions here from a Bayesian perspective.[8] In particular, I'll assume that we come to these questions equipped with a prior probability distribution, which I'll denote Pr, over both *de dicto* hypotheses (that is, roughly, hypotheses about the nature of the objective world) and *de se* hypotheses (that is, roughly, hypotheses about my *location* within an objection world—e.g., which person I am, where and when I exist, etc).[9] I'll denote the *de se* proposition that I am a sim as "*i*S", the *de se* proposition that I am a non-sim as "*i*NS", and the *de se* proposition that I have total evidence E as "*i*E." And I'll assume that my overall probability on a given proposition p, given my total evidence, is the conditional probability $\Pr(p \mid i\text{E})$.[10]

As a generalization of Bostrom's "human-type," I'm also going to make use of the notion of an "observer class," which will denote a set of observers

[8] In particular, my approach and terminology are both heavily influenced by the approach in Thomas (2021), which I see as the most rigorous in the literature.

[9] See Lewis (1979) for a classic discussion. More specifically, we can think of an objective world as a fully specific *de dicto* hypothesis, and we can think of a centered world as a triple $\langle w, s, t \rangle$ where w is an objective world, s is a subject in that world, and t is a time.

[10] Here I do not mean to assume any particular views about the relationship between your evidence and your knowledge.

such that "I am an O" or "*i*O" is true and a part of my evidence (I call this condition MEMBERSHIP below), but which can also include people who do not have my evidence.[11] To get a flavor for this notion, consider the following case.

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

Let's assume, for the moment, that the truth of the sign's claims is part of your evidence (questions about this will become important later). Now consider the observer class "people who wake up in white rooms, seeing signs like this one." *i*O, for this observer class, is part of your evidence, as is the fact that no one else in this observer class has your evidence—you see a 3 on your hand, whereas they do not. Following Bostrom, I'll denote the fraction of sims in the observer class as f_{sim}, and I'll call a world where $f_{\text{sim}} = \sim 1$ a "high fraction" world.[12]

Because *i*O is weaker than *i*E, we can imagine conditioning the prior solely on *i*O and on f_{sim} being some fraction x, while ignoring the rest of our evidence. Thus, for example, in the case above, we can imagine forgetting about the number on your hand, and simply asking: "Conditional on being a person who wakes up in a white room seeing a sign like this one, and conditional on 90% of such people being sims, what's the probability that you're a sim?". The intuitive case for Bostrom's "bland indifference principle" (BIP) above begins with the idea that at the very least, in this sort of circumstance, your answer should be 90% (formally: $\Pr(i\text{S} \mid i\text{O} \text{ and } f_{\text{sim}} = x) = x$). Some might dispute this, but I won't do so here.[13]

Importantly, though, for Bostrom's full BIP to hold, it also needs to be the case that adding in the rest of your evidence doesn't make a mean-

[11] Thomas (2021) calls this a "reference class," but this terminology calls to mind Bostrom's own use of "reference class" in the context of his work on the "Self-Sampling Assumption" (SSA) in e.g. his (2002a); and I've argued in the previous chapter that this sort of reference class is problematically mysterious. But my use of observer classes will be importantly different. In particular, Bostrom's use of reference classes in the context of SSA requires that there be one particular (and worryingly arbitrary) reference class that governs the sorts of updates SSA makes about the probability of living in a given world (or at least, the simple version requires this; we can imagine more complicated versions, where e.g. it's vague what the reference class is). But observer classes do not have this problem, because we do not have to fixate on a particular observer class. Rather, we can simply say (as I do below) that the constraints imposed by the simulation argument apply to *any* observer class satisfying certain conditions.

[12] When the specific ratio of sims to non-sims matters, I also sometimes focus directly on that. But usually talking about the fraction being ~1 is sufficient.

[13] See Weatherson (2005) for some relevant worries.

ingful difference to this probability (formally, and assuming "no difference" instead of "no meaningful difference" for the sake of simplicity, $\Pr(i\text{S} \mid i\text{E} \text{ and } f_{\text{sim}} = x) = \Pr(i\text{S} \mid i\text{O} \text{ and } f_{\text{sim}} = x)$. Let's say that an observer class is "admissible" if this further condition holds.[14] This allows us to capture Bostrom's condition that "we don't have any information that indicate [sic] that our own particular experiences are any more or less likely than other human-type experiences to have been implemented in vivo rather than in machina."

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

In my opinion, one of the most unnecessarily confusing parts of Bostrom's argument is that it focuses on one very specific observer class (namely, human-type observers) and one specific type of simulation (namely, ancestor simulations)—a choice that can seem arbitrary, which fails to address the question of how far reasoning of this broad type extends, and which raises questions about whether the argument will fall victim to the types of problems that plague appeals to "reference classes" in other contexts—for example, anthropic reasoning.[15] In fact, I think, the best formulation of the argument does not single out a particular observer class; rather, it applies to *any* observer class that satisfies a certain set of conditions, namely:

- Membership: The fact that you're in O is part of your evidence.[16] (Formally: $\Pr(i\text{O} \mid i\text{E}) = 1$)
- Observer-class indifference: Conditional only on being in O and on the fraction of sims in O being x, your credence on being a sim should be x. (Formally: $\Pr(i\text{S} \mid i\text{O} \text{ and } f_{\text{sim}} = x) = x$)

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

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

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

- ADMISSIBILITY: This credence shouldn't change once you incorporate the rest of your evidence. (Formally: $\Pr(i\text{S} \mid i\text{E and } f_{\text{sim}} = x) = \Pr(i\text{S} \mid i\text{O and } f_{\text{sim}} = x)$)

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

> CORE CONSTRAINT: For *any* observer class that satisfies MEMBERSHIP, OBSERVER-CLASS INDIFFERENCE, and ADMISSIBILITY, you cannot assign non-trivial probability to being a non-sim in a world where almost everyone in that observer class is a sim (that is, to the conjunction of iNS and $f_{\text{sim}} = \sim 1$).

As a purely formal matter, CORE CONSTRAINT looks good to me. By OBSERVER CLASS INDIFFERENCE, conditional only on iO and $f_{\text{sim}} = \sim 1$, your credence on iS should be ~1. So by ADMISSIBILITY, conditional on *all* your evidence and on $f_{\text{sim}} = \sim 1$, your credence on iS should be ~1 as well.[17] Thus, there's no room left for non-trivial credence on the conjunction of iNS and $f_{\text{sim}} = \sim 1$.

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

On its own, I see CORE CONSTRAINT as uncontroversial. But it's also, on its own, relatively uninteresting. In particular, we haven't yet said anything about which observer classes satisfy all of MEMBERSHIP, OBSERVER-CLASS INDIFFERENCE, and ADMISSIBILITY, or about whether those observer classes are such that we have reason to take the hypothesis that $f_{\text{sim}} = \sim 1$ at all seriously. Thus, for example, consider the observer class "observers with experiences exactly identical to your own." You might well grant that this sort of observer class satisfies all three of these conditions, and thus that you cannot place meaningful credence on being a non-sim

[17] Indeed, strictly speaking we do not need MEMBERSHIP as a condition. I include it, though, because the fact that you are a member of relevant observer class is central to the intuitions that simulation arguments draw on, and because I think it's easier to think about the step from OBSERVER-CLASS INDIFFERENCE to ADMISSIBILITY (and in particular, about the question of whether, conditional on iO and on $f_{\text{sim}} = \sim 1$, iE is more likely conditional on being a non-sim vs. a sim) if your membership in the observer class remains constant.

in a world where most people with exactly your experiences are sims. But you don't have any obvious motivation for taking seriously the hypothesis that there are a large number of observers having *exactly* your experiences, most of whom are sims—a hypothesis much more exotic than e.g. the hypothesis that technologically mature civilizations create lots of ancestor simulations (or other simulations) more generally.[18] Even a highly detailed ancestor simulation, after all, would presumably not replicate a historical non-sim's experience *exactly*—a point that Bostrom concedes.[19] So you can plausibly accept CORE CONSTRAINT about an observer class like "observers with experiences exactly identical to your own," without making very meaningful alterations to your everyday worldview overall.[20]

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

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

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

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

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

I. "If there are no sim blockers, most humanlike beings are sims.

II. If most humanlike beings are sims, we are probably sims.

III. So: if there are no sim blockers, we are probably sims" (p. 98).

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

IV Type 1 and Type 2 arguments

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

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

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

- A realistic simulation of the experience generated by a human can be created using approximately 10^{14}-10^{17} operations per second.[21]
- "The main computational cost in creating simulations that are indistinguishable from physical reality for human minds in the simulation resides in simulating organic brains down to the neuronal or sub-neuronal level."
- "We can use 10^{33}-10^{36} operations as a rough estimate" of the cost of a realistic simulation of human history (this is assuming that detailed simulation of the environment is not required).
- "A rough approximation of the computational power of a planetary-mass computer is 10^{42} operations per second."[22]

Let's call this set of claims Comp (for "computer power"). A Type 1 version of Bostrom's argument treats Comp as highly likely, given our evidence (that is, it accepts that $\Pr(\text{Comp} \mid i\text{E}) = \sim 1$).[23] It runs as follows:

A Type 1 version of a Bostrom-like argument:

I. Comp is likely, given our evidence. (Formally: $\Pr(\text{Comp} \mid i\text{E}) = \sim 1$)

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

[22] These quotes are all from section III, p. 3-6.

[23] It also accepts that $\Pr(\text{Sims can be conscious} \mid i\text{E}) = \sim 1$, but for simplicity I'm going to pass over this in what follows.

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

III. The observer-class "early-seeming people" satisfies MEMBERSHIP, OBSERVER-CLASS INDIFFERENCE, and ADMISSIBILITY.

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

V. Thus, it's very likely that at least one of the following is true: 1, 2, or *i*S (Formally: $\Pr(1, 2, \text{or } i\text{S} \mid i\text{E}) = \sim 1$).[24]

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

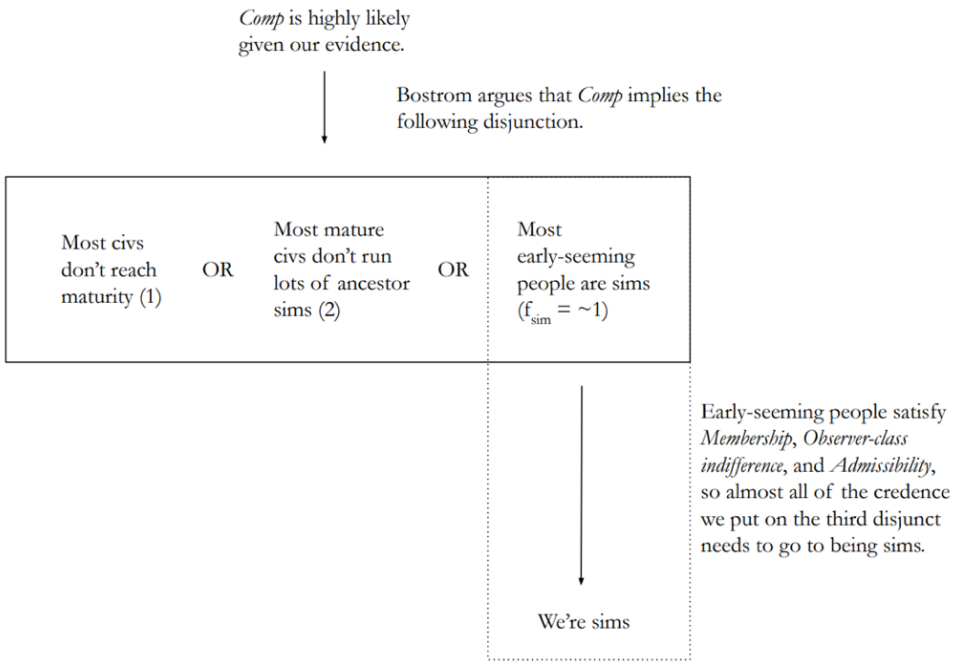

Figure 1: A Type 1 version of a Bostrom-like argument

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

A Type 2 version of a Bostrom-like argument:

[24] As discussed earlier, Bostrom's formulation of the conclusion is that at least one of 1, 2, or "we're probably sims" is true; but I've altered the formulation here to avoid confusions about conditional and unconditional probabilities. I'm also generally assuming, for simplicity, that $\sim 1 \times \sim 1 = \sim 1$.

VI. The observer-class "early-seeming people" satisfies MEMBERSHIP, OBSERVER-CLASS INDIFFERENCE, and ADMISSIBILITY.

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

That is, in picture form:

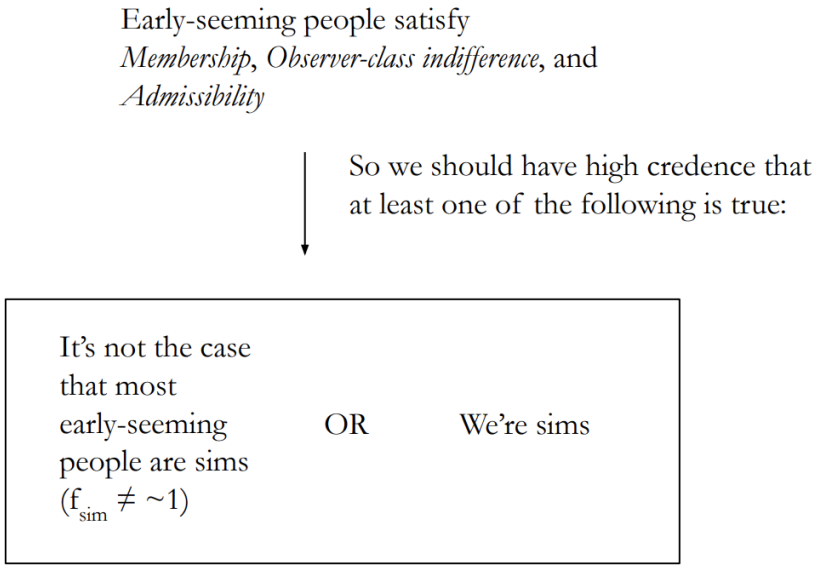

Figure 2: A Type 2 version of a Bostrom-like argument

Bostrom's is a Type 1 argument—or at least, the text of the paper strongly suggests a Type 1 interpretation. In particular, Bostrom (2005b) writes that "The simulation argument relies crucially on non-obvious empirical premises about future technological abilities" (p. 95), and the structure of the paper closely mirrors the Type 1 structure above. And indeed, many in the literature—e.g. Birch (2013), Besnard (2004), Garfinkel (unpublished), Tegmark (2014)—have read Bostrom in a roughly Type 1 way.[25]

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

Type 1 arguments, though, face objections. I'll focus on two: what I'll call the "selective skepticism" objection and the "self-undermining" objection. These objections aren't necessarily fatal to a Type 1 argument, but they can get confusing—and the confusion they create, I'll suggest, isn't necessary. We should just focus on Type 2 arguments instead.

V Selective skepticism

The selective skepticism objection to Type 1 arguments runs as follows.[26] Consider some further set of claims which, if you include them in your evidence, cause ADMISSIBILITY to fail; let's call claims of this form "admissibility blockers." Clear examples here might include: "I have two unsimulated hands," "no ancestor sims will see exactly the books on my bookshelf," and "if there are any sims, they're all in my future"—since if any of these claims are included in your evidence, then even conditional on most early-seeming people being sims, your evidence rules out being a sim, and thus ADMISSIBILITY fails.

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

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

Here, Bostrom suggests that the ancestor simulations he has in mind are centrally what are sometimes called "short-cut simulations"—that is, simulations that do not fully reproduce the empirical dynamics of the universe of the early-history civilization, but which instead only do so to the extent required to fool the inhabitants of the sim.[27] In this sense, if all you

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

[27]See Chalmers (2022), p. 94-96, for more discussion. In my opinion, this dimension

know is that you're an early-seeming person in a world where almost all early-seeming people are ancestor sims of the sort Bostrom has in mind, but then you learn that e.g. the stars you see in your telescopes are real, or that the microscopic structure of the inside of your planet exists, or that the universe you see around you contains quantum phenomena that would be extremely computationally expensive to simulate, this is a strong update towards being a non-sim—and in this sense, even such mundane, everyday scientific claims would be admissibility blockers as well.[28]

The selective skepticism objection argues that:

> PARITY OF EVIDENCE: Some admissibility blockers are on evidential footing that is comparable to or stronger than the claims in COMP.

That is, faced with a Type 1 argument, PARITY OF EVIDENCE suggests that the overall situation looks like this:

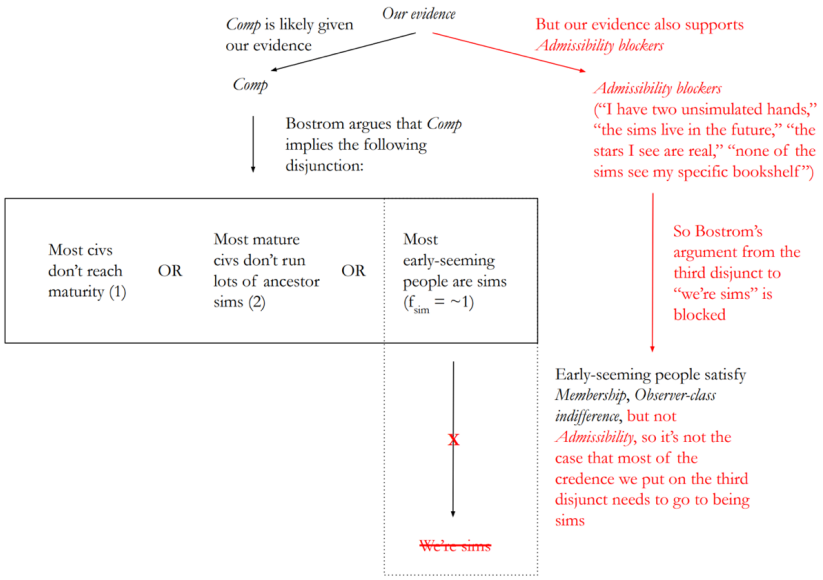

Figure 3: Selective skepticism objections to a Bostrom-like Type 1 argument

Now, by everyday standards, it does indeed seem that claims like "the

of Bostrom's discussion is underemphasized. That is, in my experience, casual readers of Bostrom often end up assuming that if they are sims, they live in a simulated world where their conventional scientific worldview is at least true *of that world*. Actually, though, the sorts of simulations that Bostrom has in mind are much more revisionary—e.g., they involve fake stars, fake scientific experiments, possible edited-memories if anyone ever notices something amiss, and so on.

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

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

> IMPERFECT ANCESTOR SIMS: You are a mediocre office worker living in a pre-post-human civilization, who doesn't tend to think about big picture topics much. Rather, you get most of your information on them from the stable global government's official scientific authorities, who thus far have never (to your knowledge) led you astray. One day, you're sitting by your bookshelf and watching TV, when a broadcast from the official scientific authorities comes on. They say:
>
> "People of Earth: we have a few announcements:
>
> - Announcement 1: Our super-duper forecasters are saying that it's 99.9% likely that we're going to make it to technological maturity, that this government is going to remain stable until then, and that whatever intentions we commit to now will be enacted later.
>
> - Announcement 2: The universe is finite, and we're the only life that exists anywhere or that will ever develop on its own. Also, the physics of the universe we see around us involves quantum phenomena that would be extremely computationally expensive to simulate, the stars we see in our telescopes are real, and the physical microstructure of the earth continues to exist when no one is looking.
>
> - Announcement 3: It'll take around 10^{36} operations to run a detailed simulation of all of our cognitive history—an 'ancestor simulation.' (That is, assuming we take lots of shortcuts on the environment and don't simulate quantum phenomena, stars, galaxies, or the earth's physical microstructure—if we had to do that, running ancestor sims would be out of the question.) And we're going to have tons of planet-sized computers that can run at least 10^{42} operations *per second*, so running tons of detailed ancestor simulations is going to be extremely easy in the future.
>
> - Announcement 4: In light of this, we've decided to make the following binding commitment: come technological maturity, we're going to run a billion such ancestor simulations (but that's it: no more, no less). However, we're not going to be able to recreate our history *exactly*. Rather, lots of details are going to be forever lost. For example, none of the sims are going to have the same books on their shelves as you do.
>
> - Announcement 5: We're going to turn off the ancestor simulations before they reach technological maturity—so even though they will likely *think* that they're going to run simulations themselves (and their global governments will make announcements to this effect), they actually won't."

Obviously, this case differs from our own situation in a number of re-

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

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

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

VI Self-undermining

The self-undermining objection is related but distinct.[29] It runs as follows. Consider:

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

SIM IGNORANCE seems intuitively plausible, especially once we bring to mind that for Bostrom's argument to work, the claims in COMP need to specifically apply to the level of reality simulating us—a place that we, if we're sims, have never seen, touched, or been to.[30]

But SIM IGNORANCE is in tension with the structure of a Type 1 argument—or at least, with the credence on *i*S that the argument is supposed to argue for. That is, a Type 1 argument works by arguing that COMP is very likely given our evidence, then arguing that our credence *on* COMP needs to be divided between 1, 2, and $f_{\text{sim}} = \sim 1$, and then arguing that the portion

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

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

of *that credence* that goes to $f_{\text{sim}} = \sim 1$ needs to go almost entirely to *i*S as well. Thus, the credence that ends up on *i*S, as a result of the argument, is specifically credence on the conjunction of Comp and *i*S. That is, to the extent a Type 1 argument tells us to put credence on being sims, it tells us to put credence specifically on being sims in worlds where Comp is true of the level simulating us. But if that's *all* the credence on being sims that we end up with, then Sim ignorance fails: conditional on our evidence and on being sims, Comp is treated as certain.

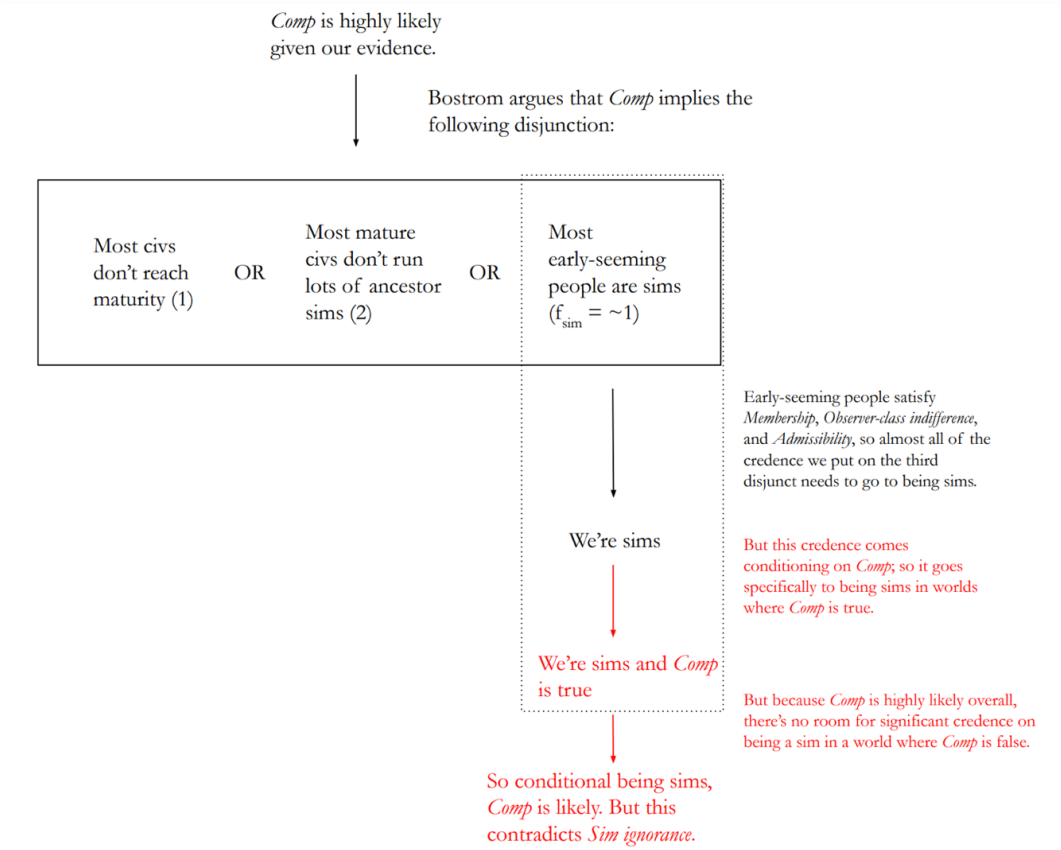

Figure 4: A self-undermining objection to a Bostrom-like Type 1 argument

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

at less than 1/99, which is quite a bit lower than Bostrom wants to go.

VII Arguing for Admissibility directly

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

Now, advocates of Type 1 arguments may have replies to these objections. Perhaps, for example, Type 1 arguments could argue against Parity of evidence, on the grounds that hypotheses to the effect that $f_{\text{sim}} = \sim 1$ undermine or defeat our evidence for claims like "the stars I see are real," but not our evidence for claims like "a rough approximation of the computational power of a planetary-mass computer (in unsimulated reality) is 10^{42} operations per second."[31] And perhaps they could reject Sim ignorance, despite its initial intuitive appeal, and say that actually, if we're sims, we're most likely to be sims who are right about claims like Comp.[32]

Below I'll discuss replies like this in a bit more detail. But I don't think we need to rest the viability of the simulation argument, in general, on the viability of replies like this, because we don't have to treat Comp as highly likely. Rather, we can run a Type 2 argument instead: one that just claims, directly, that Membership, Observer-class indifference, and Admissibility holds for the observer class "early-seeming people," and thereby concludes, per Core constraint, that we can only have extremely low probability on the conjunction of *i*NS and $f_{\text{sim}} = \sim 1$—regardless of our views on Comp.

Membership isn't in dispute, and I'm happy to grant Observer-class indifference, as are many of Bostrom's objectors (if *all you knew* is that you're an early-seeming person, and that x is the fraction of early-seeming people that are sims, it seems to me very reasonable to put x credence on being a sim). So the key claim here, as I see it, is that the observer class

[31] Some of the comments in Bostrom (2005b), for example, seems to me to suggest this sort of response.

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

"early-seeming people" satisfies ADMISSIBILITY: that is, recall, that if you had *all* your evidence and you knew that x is the fraction of early-seeming people that are sims, your credence on being a sim should be the same as if you only knew that you're an early-seeming person and that x is the fraction of early-seeming people that are sims (i.e. $\Pr(i\text{S} \mid i\text{E} \text{ and } f_{\text{sim}} = x) = \Pr(i\text{S} \mid i\text{O} \text{ and } f_{\text{sim}} = x)$). Let's look more closely at this claim now.

ADMISSIBILITY falls out of MEMBERSHIP and OBSERVER-CLASS INDIFFERENCE if we also grant the following:

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

That is, once you know that you're in the observer class and that the fraction of sims in the observer class is x, the rest of your evidence doesn't tell you anything about whether you're a sim or a non-sim—rather, it's equally likely either way.[33] The aim here is to get at the more general analog of the sort of reasoning you do in SIMS WITH RANDOM NUMBERS about the evidence you get from the "3" you see written on your hand. In particular: because (on God's story at least, which we grant for the purposes of the case) you're equally likely to have ended up with a "3" written on your hand conditional on being a sim vs. a non-sim, adding in the fact that you have a "3" written on your hand doesn't update your probability on being a sim vs. a non-sim, once you've taken into account the fact that 90% of people who wake up in white rooms seeing the sort of sign you're seeing are sims. The claim is that the same sort of dynamic applies—at least roughly—to the rest of our evidence, once we update on being early-seeming people and on the fact that most early-seeming people are sims. That is, MEMBERSHIP, OBSERVER-CLASS INDIFFERENCE, and NO UPDATE hold for the class early-seeming people, so ADMISSIBILITY holds as well.

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

VIII No update from your experiences

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

> SIMPLE INTERNALISM: Your evidence is constituted by your experiences. (Formally: $i\text{E} = i\text{Q}$)

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

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

> NO UPDATE FROM YOUR EXPERIENCES: Conditional on being in the observer class, on the fraction of sims in the observer class being x, and on being a sim, the probability of having your *experiences* is the same as it would be conditional on being in the observer class, on the fraction of sims in the observer class being x, and on being a non-sim. (Formally: $\Pr(i\text{Q} \mid i\text{S and } i\text{O and } f_{\text{sim}} = x) = \Pr(i\text{Q} \mid i\text{NS and } i\text{O and } f_{\text{sim}} = x)$)

Conditional on MEMBERSHIP and OBSERVER-CLASS INDIFFERENCE, this principle gets you to high credence on being a sim, conditional only being in the observer class, most people in the observer class being sims, and having your experiences in particular. And then SIMPLE INTERNALISM says that having your experiences in particular is all the evidence you've got. So conditional on all your evidence and on most people in the observer class being sims, you're still at high credence on being a sim.

Should we accept NO UPDATE FROM YOUR EXPERIENCES, applied to the observer class "early-seeming people"? To me it seems plausible that we should, at least in fairly idealized cases like IMPERFECT ANCESTOR SIMS. Thus, e.g., it's not as though the books on my particular bookshelf are any more likely, on priors, to show up on the bookshelf of a sim vs. a non-sim; rather, they're just a set of 21st century books, with nothing, on priors, especially sim-y or non-sim-y about them. In this sense, I should treat my seeing these books in particular the same way I treat seeing the number "3," in particular, in SIMS WITH RANDOM NUMBERS. And the same, plausibly, can be said of all my experiences, once we condition on iO and $f_{\text{sim}} = \sim 1$—that is, that these experiences are not as any particular indication of sim-hood vs. non-sim-hood.

We can restate the intuition here, and the eventual upshot, in terms of your absolute priors over having different experiences in different scenarios. Thus, consider a more general version of the description of the world offered by the government in IMPERFECT ANCESTOR SIMS. On this description, there are n early-seeming non-sim people on a non-simulated planet (call this planet 0), and a billion sets of n early-seeming sim people, each on a different simulated planet (call these planets 1-1,000,000,000), all of whom hear announcements from their government like announcements 1-5 above, and all of whom have fake stars, simulated hands, and no sims in their future. Let's call any world that fits this description a "Z world," and let's call the proposition that you have experiences Q on planet y in a Z world $i\text{ZQ}_{\text{y}}$.

One way of drawing out the intuition in favor of NO UPDATE FROM YOUR EXPERIENCES, applied to a case like IMPERFECT ANCESTOR SIMS, is to note that on priors, and conditional on living in a Z-world, it seems equally likely that you have experiences Q on any one of these planets. That is,

$\Pr(i\text{ZQ}_0) = \Pr(i\text{ZQ}_1) = \Pr(i\text{ZQ}_2)$... If your experiences Q are, for example, hearing a particular set of government officials announce that you aren't a sim but that there are lots of sims in the future, these experiences simply aren't any more likely to show up on the non-sim planet vs. any given sim planet. By hypothesis, in Z-worlds, *all the governments* (sim and non-sim) announce that their listeners are non-sims but that there are lots of sims in the future. Thus, finding yourself with experiences Q *can't* simultaneously update you towards high probability on living in a Type Z world, *and* high probability on being a non-sim—all of your credence on Type Z worlds needs to stay split equally between each of $i\text{ZQ}_n$, and there are a more than a billion of these, all mutually incompatible, in play. And if $i\text{E} = i\text{Q}$, then your experiences are all the evidence you have. So your posterior credence on $i\text{ZQ}_0$ (the only Z-type scenario where you're a non-sim) is (dramatically) capped.

Overall, then, No update from your experiences looks pretty good to me, at least in cases like Imperfect ancestor sims.[34] So if we accept Simple internalism, then No update follows, as does Admissibility.

IX Reject No update?

To my mind, the most philosophically interesting way of denying Admissibility is to grant No update from your experiences, but to deny No update. Thus, for example, in a case like Imperfect ancestor sims, and conditional on living in a Z-type world, you can concede that your experiences are equally likely to occur on the non-sim planet vs. any given sim planet (that is, $\Pr(i\text{ZQ}_0) = \Pr(i\text{ZQ}_1) = \Pr(i\text{ZQ}_2)$...), but deny that you would have the same evidence, if you were having those experiences as a sim vs. a non-sim. In particular, you might say that if you're a non-sim, then one or more of the admissibility blockers we discussed above (e.g., "I have two unsimulated hands," "the stars I see are real," "all the sims live in the future," "none of the sims see my bookshelf," etc) are either included in your evidence, or supported by it in a way that simple internalism cannot account for. And indeed, in the real world, we tend to treat claims like "the stars I see are real" like they are on quite solid evidential footing. Type 2 arguments, then, plausibly require denying this sort of common-sense, at least if we're also going to put substantive credence on most early-seeming people being sims in whose mouth such claims are false.

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

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

even if they might have been in other circumstances.[35] And this seems plausible to me. Maybe I can normally be confident that I am not an envatted brain. But if I learn that the aliens are going around envatting many people's brains while they're sleeping—enough, indeed, that most human brains are envatted—then this confidence needs to alter.

Now, strictly, this is not enough for the present purpose, because you may not have learned that you live in a Z world—rather, you may only have some positive credence c that you live in a Z world.[36] But if we accept the claim that *if* you condition on living in a Z world, then you should have high credence on being a sim (just as, if you condition on the aliens envatting almost everyone, you should have high credence on being a BIV), then it seems like whatever overall credence c you put on living in a Z world, conditional on your evidence, should mostly go to being a sim—and our conception of evidence will need to accommodate this.

Wading too deeply into these waters, though, is beyond my purpose here.[37] I'm happy to grant that there is theoretical daylight between *No update from your experiences* and No update, that various epistemologies may accept the former but resist the latter, and that the simulation argument works most smoothly against the backdrop of epistemologies that are fairly happy to grant the latter once the former is in place (Simple internalism is an especially clear example).[38]

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

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

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

[38]If we open ourselves to doubting beliefs like "I have two unsimulated hands" and "the stars I see in my telescope are real," though, does the simulation argument retain its dialectical interest relative to discussions of more standard skeptical threats in the literature? Yes. Simulation arguments rest specifically on claims about how to relate epistemically to hypotheses where most observers of a certain type are in a skeptical scenario, and where, at least naively, we have some empirical reason to take seriously a hypothesis of this form. More standard skeptical discussions have neither of these features. And while simulation arguments require that on priors, and conditional on living in a Z world with experiences Q, you're equally likely to live on any of the planets

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

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

If I were in this case, I would feel intuitively uncomfortable resting easy with the belief that I'm a non-sim about to press the button.[39] And an epistemic procedure that licenses such confidence would lead the sims, at least, very much astray—sims that I would be actively expecting to share the world with (not some purely hypothetical set of sims, living in a pos-

(e.g. $\Pr(i\text{ZQ}_0) = \Pr(i\text{ZQ}_1) = \Pr(i\text{ZQ}_2)$...), they do not include any comparable constraints relating your prior on being a brain in a vat with experiences Q (call this *i*BIVQ) to your prior on having experiences Q in worlds with no BIVs at all. Even after having experiences Q, then (and even accepting SIMPLE INTERNALISM), a hypothesis like *i*BIVQ can remain arbitrarily less likely than a hypothesis like "I'm a non-sim in a Z world."

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

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

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

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

sible world I have no reason to put much credence on). And I would feel this same discomfort if, say, I thought the button was going to be pressed with only 10% probability (say, if a ten-sided dice came up 1)—that is, I would feel like I couldn't put 10% on the button getting pressed, while also putting ~100% on being a non-sim.

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

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

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

For this reason, even in the absence of a worked-out epistemology that tells us whether or not to make the transition from NO UPDATE FROM YOUR EXPERIENCES to NO UPDATE and thus to ADMISSIBILITY, Type 2 arguments seem to me independently forceful and attractive. That is, to me it seems quite plausible that even for fairly inclusive observer classes like "early-seeming people," MEMBERSHIP, OBSERVER-CLASS INDIFFERENCE, and ADMISSIBILITY will hold, and thus, that the upshot of CORE CONSTRAINT will apply: you can't put substantive credence on the conjunction of *i*NS and $f_{\text{sim}} = \sim 1$.

X Where should we end up overall?

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

- How should we adjust our overall credences—including our credence

on being sims—in light of Type 2 arguments?

- To what range of cases and observer classes do Type 2 arguments apply?

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

Let's say that prior to considering the simulation argument in its entirety (but after reflecting on cosmology, the current landscape of existential risks, the possible motivations for running simulations of early-seeming people, and the existing empirical evidence about the computational power available to technologically mature civilizations), you had the following pattern of credences:[40]

Starting distribution

a) *i*NS: ~100%

b) *i*S: ~0%

c) Comp: 99%

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

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

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

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

Bostrom-like distribution

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

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

c) Comp: 99%

d) "The stars I see are real": **~80%**[41]

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

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

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

We can think of this sort of adjustment as attempting to preserve as much as possible of your previous picture of the objective world, while allowing that conditional on (e), your confidence in your location within that world (and in some aspects of your previous picture of the objective—for example, whether the books on your shelves appear on an early-history non-sim planet, or on a sim planet) becomes undermined—and with it, your confidence in claims like "the stars I see are real." And this is, indeed, a salient possible end-state: albeit, one that requires rejecting the idea that you should be least as confident in (d) as in (c) (call this "*Stars* $\geq$ *Comp*"—it's one way of cashing out PARITY OF EVIDENCE above), and that conditional on being a sim, COMP is not likely (i.e., SIM IGNORANCE).

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

Apparently-f_s-is-small distribution

a) *i*NS: ~100%

b) *i*S: ~0%

c) COMP: 99%

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

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

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

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

Trying-to-capture-Sim-ignorance distribution

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

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

c) COMP: **~80%**

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

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

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

XI Are we almost certainly sims?

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

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

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

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

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

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

iv. You're a sim in some other situation.[42]

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

More generally, whenever you give c credence to being a non-sim in a world where the ratio of sims to non-sims is R, you are committing yourself

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

to giving Rc credence to being a sim in such a world (this is why, according to CORE CONSTRAINT, c has to be less than $1/R$ overall). So if c is much more than a $1/R$ fraction of your total credence on iNS, then iNS gets swamped by iS.[43]

The key premise in this argument is the claim that conditional on being a non-sim, the expected ratio of sims to non-sims in the observer class is high. And this premise looks initially plausible. After all, we're used to assuming that we're non-sims, and the empirical facts Bostrom points to (i.e., COMP) make it hard to rule out futures where our descendants create very large numbers of sims in various plausibly admissible observer classes—even if we're substantially more skeptical of Bostrom's empirical case than Bostrom is.

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

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

In such a case, are we supposed to conclude that actually, despite all appearances to the contrary, we are overwhelmingly likely to be sims? This is the conclusion that would fall out of Thomas's argument, in combination with the credences that the super-duper forecasters give conditional on being non-sims. But it seems a strange lesson. Is it really so hard to keep some credence on being a non-sim, in a world where it appears overwhelmingly likely that no sims will ever be created?

I think that this is indeed the conclusion you should reach, *if* you accept the credences that the super-duper forecasters give conditional on being non-

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

sims. But you need not accept such credences. And indeed, if you don't, you can explain why: namely, that the super-duper forecasters *aren't taking simulation arguments into account*.[44] After all, as the previous section made clear, such arguments generally require that we *revise* something about how we would have otherwise apportioned our credences—and a very salient revision (indeed, one more made more salient by the implication that Thomas's argument highlights) is to *drastically* reduce our credence in a high ratio of sims to non-sims, conditional on being non-sims.

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

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

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

Preserve your credences conditional on being a non-sim

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

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

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

c) COMP: **~100%**

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

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

Or to put it another way, Thomas's argument points at a very substantive additional constraint on the overall credences you need to end up with after considering simulation arguments: for all observer classes that satisfy MEMBERSHIP, OBSERVER CLASS INDEPENDENCE, and ADMISSIBILITY, *either* the expected ratio of sims to non-sims in those classes needs to be low conditional on being a non-sim, *or* you have to be basically certain that you're a sim. And naively, a high expected *R* conditional on being a non-sim seems very reasonable; as does placing at least some substantive credence on being a non-sim. But so does not being basically certain that you're a sim. So something has to give.

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

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

Let's turn to another thorny issue.

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

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

> SQUID SIMS: The situation is just like IMPERFECT ANCESTOR SIMS, except that according to the announcements you hear, the government authorities do not decide to run any ancestor simulations. Indeed, they decide that they will never, ever simulate any humans. Rather, they decide to run a billion simulations of *squid people* with tentacle arms, living in squid civilizations that haven't yet reached technological maturity—and not to run any other sims. What's more, while the authorities expect the simulated squids to plan on running simulations of some yet-different animal civilization (and

to have global governments that make announcements to this effect), they are going to turn off the squid sims before such sims make it to technological maturity.[45]

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

Now, in my experience, many people's intuitive reaction in this case is that Admissibility does not hold, and that it's fine to believe the government's entire story.[46] And perhaps, ultimately, there will be a way of justifying this intuition. Personally, though, I'm skeptical: my own best guess is that a Type 2 argument blocks believing the government's whole story here, as well.

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

Here's an intuition pump that might be helpful. Consider the set of worlds that fit the following schema, for some set of three animal types A, B, and C (e.g. lions, humans, and squids). On planet 0, there is one set of n early-seeming non-sim A-animals, whose government says "all the early-seeming A-animals are non-sims (so you're non-sims), but there are a billion sets of n early-seeming sim B-animals in the future"; and then, in

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

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

the future of planet 0, and on each of planets 1-1,000,000,000, there are n early-seeming sim B-animals, all of whose governments say "all the early-seeming B-animals are non-sims (so you're non-sims), but there are a billion sets of n early-seeming sim C-animals in the future" (when actually, no animals of type C exist). Call any world that fits this description—for some values of A, B, and C—a V-world.

Conditional on living in a V-world and on having the experiences described in Squid sims—e.g., looking down at human hands, listening to a government announcement about future squid sims—is it any more likely that you live on Planet 0 vs. Planet 1? I don't see why it would be. Humans, after all, are not more likely, on priors, to play the role of a type A animal, in a V-world, vs. a type B animal; and the same holds for squids with the respect to type B and type C. That is, "humans simulate squids, who wrongly think they'll simulate some other type of animal" is no more likely, on priors, than "some type of animal simulates humans, who wrongly think that they'll simulate squids." And conditional on it being humans on any given planet, your experiences in particular seem equally likely. So to me it seems plausible that your credence in living on Planet 0 should be equal—or at least, roughly equal—to your credence in living on Planet 1, and same for Planet 2, 3, and so forth. Thus, conditional on living in a V-world and having your experiences, the usual Type 2 logic applies, and if we accept it in the previous cases (e.g., Z worlds above), then it seems like you can't have gotten strong evidence that you live on Planet 0 in particular. But the world you hear your government describing, in Squid sims, is effectively a V-world where you live on Planet 0. So you can't have strong evidence for the government's whole story.

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

In fact, Sims with different light speckles, Imperfect ancestor sims, and Squid sims seem to me sufficiently analogous, structurally, that

my best guess is that they stand or fall together—that is, that Type 2 simulation arguments either work in all these cases, or in none of them. My best guess is: all of them—and I think that if you say "none of them," then you should probably start resting easy with the belief that you're a non-sim in INDISTINGUISHABLE SIMS as well (though this seems to me quite strange), given its similarity to SIMS WITH DIFFERENT LIGHT SPECKLES. That said, I haven't tried to exhaust all of the possible disanalogies here, and it's possible that there's some other line to be drawn (ideally, in my opinion, between IMPERFECT ANCESTOR SIMS and SQUID SIMS, as I do think that applying Type 2 arguments to SQUID SIMS is somewhat counterintuitive, and it would be nice to explain why).

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

> SIMS WITH LITTLE TAGS: A case like IMPERFECT ANCESTOR SIMS, but the stable global government also announces that it's going to put a little tag in the visual field of all the future sims, which says "you are a sim"—but where such sims have 21st-century-like experiences beyond this. You have no tag in your visual field.

Are you allowed to put substantial credence on the government's announcements being true—including the bit where you're a non-sim? That depends on whether ADMISSIBILITY holds for any observer classes that you and these sims-with-tags are both a part of—for example, the observer class "people who have 21st-century-like experiences, whether or not they have little tags." This observer class, too, plausibly satisfies OBSERVER CLASS INDIFFERENCE—so the main question is whether, once you condition on being in this observer class and on $f_{\text{sim}} = \sim 1$, the rest of your evidence—e.g., listening to your global government announce a plan to simulate lots of sims with little tags, not having a little tag yourself—is any more or less likely, conditional on being a sim vs. a non-sim. But just as, in SQUID SIMS, you can't rest easy with "all the sims are squids" (since the experience of the government announcing that all the sims are squids seems similarly likely conditional on being a sim vs. a non-sim, given *i*O and $f_{\text{sim}} = \sim 1$), neither can you rest easy with "all the sims have little tags." Rather, the question is whether government announcements to the effect that all the sims have little tags, combined with the absence of little tags in your experience, are a strong sign you're a non-sim, given *only* that you have 21st-century-experiences, and that almost everyone with such experiences is a sim. And this seems to me quite unclear.

And this same unclarity applies in a whole panoply of more exotic cases—for example, ones in which the government announces that e.g. it will only run sims of people on cooking shows; or of people in extremely violent and entertaining video games; or of people for whom running sims seems

impossible; or of universes governed by physical laws dramatically different from our own; or where the government announces that humans will never run sims, but the lizard people the next planet over are getting ready to run lots of ancestor sims.[47] Importantly: in all of these cases, we need to take care not to conflate the frequency of our own experiences among sims vs. non-sims, *in the world that the government posits*, with the likelihood, *on priors*, of our having those experiences given that we're sims vs. non-sims, conditional only on *i*O and $f_{\text{sim}} = \sim 1$. After all, our priors, here, cannot be set by the frequencies *within the world*—because which world obtains is precisely the question. And this makes ADMISSIBILITY correspondingly hard to reason about.

What's more, though, ADMISSIBILITY isn't actually required. Rather, the argument continues to work, in a roughly similar way, even if your particular experiences are a strong update towards being a non-sim, after you condition on *i*O and $f_{\text{sim}} = \sim 1$. Thus, for example, suppose you are convinced that experiences as boring and mediocre as yours are massively more likely to be had by a non-sim with 21st century experiences than a sim with 21st century experiences, conditional on *i*O and $f_{\text{sim}} = \sim 1$ (perhaps because you think entertainment value by far the most likely explanation of someone running sims). Still: *how much* more likely? If, for example, *i*O and $f_{\text{sim}} = \sim 1$ leaves you at a billion to one on *i*S, then even if your boring experiences are a million to one update towards *i*NS, you'll still be at a thousand to one on *i*S at the end of the day. So if the ratio of sims to non-sims is large enough, ADMISSIBILITY can fail very badly, and you'll still lose your ability to place much credence on the conjunction of *i*NS and $f_{\text{sim}} = \sim 1$.[48]

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

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

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

XIII Wrapping up

In the last few sections, I've discussed a number of complications and uncertainties that arise if we start to take Type 2 arguments seriously. In particular: we need to find an overall credence assignment that respects Core constraint for all observer classes that satisfies Membership, Observer-class indifference, and Admissibility (a set that may be quite large, and which seems challenging to analyze), while also navigating the pressure created by principles like *Stars* $\geq$ *Comp* and Sim ignorance, and by arguments like Thomas's (2021) for high expected ratios of sims to non-sims, conditional on being a non-sim.

In closing, I'll also briefly note a few other issues. First: simulation arguments—including Type 2 versions—work best in finite worlds. Indeed, Bostrom "deliberately sets aside" infinite worlds in his original paper. In a later FAQ, he suggests that handle them, we might appeal to the limiting fraction of sims vs. non-sims in expanding hyperspheres—but this sort of approach faces significant problems.[49] Granted, finding a way to talk sensibly about the fraction of observers with X experiences in infinite worlds is a problem for cosmologists more generally, but the existence of this problem adds an open question about how to best apply simulation-argument-style reasoning to infinite cosmologies.[50]

Second, I haven't been discussing the bearing that anthropic principles like the Self-Indication Assumption ("SIA") and the Self-Sampling Assumption ("SSA") might have on hypotheses involving simulations—but pretty clearly, there are implications. SIA updates towards worlds where there are more observers with evidence like yours; whereas SSA updates towards worlds where the observers with evidence like yours are a larger fraction of some reference class.[51] Notably, though, sim-filled worlds plausibly involve more observers in epistemic situations like yours—thereby

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

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

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

prompting SIA to update towards them (and especially: towards worlds obsessed with simulating you in particular). But they also involve lots of civilizations reaching technological maturity; and if the average non-sim population in a technologically mature civilization outnumbers the average sim population (plausibly, after all, most of the resources go to the actual *citizens* of the civilization itself, rather than to running sims), and most of the non-sim population has technological-maturity-indicating experiences quite unlike our own, then on SSA, finding yourself without technological-maturity-indicating experiences is a large update *against* worlds like this—since conditional on living in such a world, you should've expected to be a non-sim member of technologically mature civilization, rather than either a sim, or a non-sim very early on in history.[52] And I expect other implications of SIA and SSA for debates about sims as well, beyond these examples.

Finally, I haven't, here, tried to tackle the practical implications that fall out of the discussion—and in particular, that would fall out of starting to give serious credence to being a sim. Some writers on this topic have explored the possibility that, for example, research into whether we live in a simulation itself risks causing the simulators to turn us off;[53] that the simulation argument should make you act more selfishly;[54] and that it should make you act on shorter time horizons and with less concern for the long-term future.[55] Topics like these seem to me well worth further investigation.

All in all, then, there's still a lot to work out. Still, as far as I can tell, the basic thrust of the simulation argument has real philosophical force and interest—especially when interpreted in the Type 2 manner I've argued for here (that is, as not resting on the likelihood of any particular set of empirical claims). Perhaps it does not, ultimately, work—but I don't think its failures are at all obvious, and I expect that teasing them out, if they exist, will itself be an instructive exercise. After all, serious arguments for such dramatic re-orientations in our basic understanding of our existential predicament do not come along every day. We should pay attention when they do.

And whether we buy simulation arguments or not, they are a reminder that the world we see and take for granted is only a part of the world; and that in principle, our overall existential situation could in fact be many different ways, not all of which we are accustomed to considering. Ultimately, we need priors—and indeed, capacious ones, adequate to include worlds that

[52] Or, let's assume, a non-sim in some weirder situation—i.e., a fake, terraformed non-sim world created by an advanced civilization. And of course, as with all conclusions drawn from SSA, this one depends on the reference class.

[53] See Greene (2020).

[54] See Hanson (2001).

[55] See Tomasik (2016).

are bigger and stranger than what we take to be normal (is it so normal, when you step back and look?). Dealing well with such worlds is a delicate art. But it's one that simulation arguments, whether sound or not, remind us to learn.[56]

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