Anthropic reasoning in infinite worlds

Evan Gunter

October 21, 2020

Abstract

Only worlds which contain observers can be observed. This tautology is the root of anthropic reasoning. Although incorporating this tautology into reasoning which doesn't account for this selection bias does profoundly change the credences we should hold, the full power of anthropic reasoning is revealed when moving from the logically required to the probable. We are probably "typical" observers. Intuitively, this should change our credences even more. For example, it lets us change our credences about hypotheses on which there are *few* observers like us, rather than *no* observers like us. Some such hypotheses deal with the number of spacetime dimensions: although, on various assumptions about possible physics, it is very hard or impossible to demonstrate that a number of spacetime dimensions is completely devoid of life, it is sometimes doable to show that it will be very sparsely populated. However, we cannot change our credences without a formalization of this type of reasoning. Several such precise formalizations of anthropic reasoning have been proposed. There are two which are most prominent: the Strong Self-Sampling Assumption (SSSA) and the Self-Indication Assumption (SIA). I show that these are both special cases of a more general principle. SSSA is much more widely accepted. However, I argue that it is (1) less intuitive and more arbitrary, (2) less predictive, and (3) no less flawed than SIA. Also, existing formalizations of anthropic reasoning fail spectacularly in cases like our world: infinite cases. I discuss ways to extend anthropic reasoning to work in these cases. Finally, I apply the developed formalizations to the case of the dimensionality of spacetime.

1 Introduction

There are some questions about our universe that standard, causal scientific methods may be unable to answer. Why did so many factors align to make Earth habitable? Why is the vacuum energy so much lower than the predicted value?[7] Why are there three spatial dimensions and one time dimension?[2] These facts may lack causal explanations. For example, there is presumably no single causal reason explaining why Earth is the appropriate distance from a star, has a sufficiently non-eccentric orbit, is far from gamma ray sources, and has a magnetosphere, and has all of its other characteristics that seem plausibly to be necessary for life.[1] But we should not be at all surprised by this seeming coincidence. Without the factors making life on Earth possible, there would be no life, and hence we would not be here to ask the question. We are confused when we treat Earth as drawn randomly from the set of planets, but not when we treat it as drawn from the set of habitable planets.[2] Likewise, if the vacuum energy were as high as predicted by current theories, complex structures, and hence life, could not exist;[3] and, as argued by Tegmark [2], if there were a different number of spatial or time dimensions, life might not be able to exist.[4]

Reasoning that treats what we observe as drawn at random from a set of experiences, rather than being drawn at random from states of the world, is called anthropic reasoning. As an anthropic reasoner, I have

[1] This assumes that the Earth was not designed by some higher power. Indeed, the type of reasoning discussed here—anthropic reasoning—removes the force behind many arguments for gods, by providing an alternate explanation.

[2] At least as long as the set of habitable planets is likely to be nonempty, which it is in large universes such as our own.

[3] Of course, our theories may change to predict a much lower vacuum energy, in which case an anthropic explanation is somewhat redundant.

[4] In these cases, our confusion is not fully resolved since we may predict that there should be *no* observers, in which case the fact that we exist is still a shocking coincidence—but even in these cases, our conclusions are significantly changed by noting that some characteristics of our observations could not have been otherwise. For example, it may suggest that we live in a multiverse with varying values of the parameter in question; then a world with the appropriate value of the parameter for life is likely to exist.

some "prior" credences about what my experience will be like. Then, I "update" these credences to reflect the information I gain by knowing that my current experience has the characteristics that it does.[5] The information I gain this way is weak—I only learn from one data point: my own current experience. A close analogy is trying to learn about typical size of the fish in a lake from a single catch. However, it can be informative if my experience has *a priori* special characteristics which my prior credences hold to be very unusual—just as how I do learn something about the sizes of fish in the lake if the one fish I catch is 10 meters long. If my experience has an *a priori* special characteristic—for example, that it seems highly structured and intelligible, rather than chaotic and random—then that provides evidence that a large number of experiences have this special characteristic. Although there are several ways to formalize anthropic reasoning even in the finite setting it is generally clear that it can be formalized without causing any excessively counterintuitive outcomes. However, things are much different in the infinite setting. It is not clear that anthropic reasoning in this setting can be intelligibly formalized at all, and even if it can be, numerous problems arise in the attempt. I'll consider two of these problems.

The first problem is that the universe is most likely infinite, and there are most likely an infinite number of observers within it. This makes it difficult to draw conclusions from the assumption that my current experience is drawn randomly from some set of experiences. If my prior credences hold that there are some lakes with mostly small fish but a few big fish, and some lakes with mostly big fish but a few small fish, then, if all lakes contain a finite number of fish, on catching a large fish I conclude with high probability that this lake contains mostly large fish. But, if all the lakes contain an infinite number of fish, it's not clear what I should conclude, or even if my prior credences are meaningful: every lake contains an infinite number of both large and small fish, so I can't find the probability of catching a large fish by dividing the number of large fish by the total number of fish. More broadly, it may be that, for multiple hypotheses, there are an infinite number of experiences exactly like my current experience. I have intuitions that some hypotheses render my current experience much more likely—for example, the hypothesis that I am a life form living on a planet, rather than a "Boltzmann brain" arising from chaotic fluctuations in near-empty space—but, if the universe is spatially infinite[6], then there are infinitely many observers with exactly my experience on both of these hypotheses. Is there a principled way to compare these different infinities? Is the total number of experiences like mine which would exist given some hypothesis even the right quantity to compare?

The second problem with infinities in anthropic reasoning is that, if we accept what Bostrom[1] calls the Self-Indication Assumption (SIA)—one formalization of anthropic reasoning—then (barring logical contradiction) I should believe, with probability 1, that the universe is infinite. This is because one implication of SIA is that my existence is evidence for the existence of many observers—I am more likely to exist if there are many observers (just as how I'm more likely to catch a fish if the lake is full of them). So, if there are infinitely many observers given some hypothesis, and only finitely many given its negation, then the fact that I exist leads me to conclude with probability 1 that the hypothesis holds in my universe. Bostrom discusses this consequence in the "presumptuous philosopher" thought experiment ([1], p.124)—where a philosopher adhering to SIA believes with certainty that the universe is infinite, even though the physical evidence is unclear. Bostrom claims that this objection is fatal for SIA, and suggests as an alternate formalization the Strong Self-Sampling Assumption (SSSA), which avoids this problem at the cost of holding that experiences don't behave like fish—your prior credence of existing must always be 1, whereas your prior credence of catching a fish depends on the number of fish in the lake. Is there a way to avoid the chain of reasoning that leads to these undesirable consequences for SIA?

I'll consider these two problems in this essay. First, I motivate a general formalization of anthropic reasoning, of which SIA and SSSA are special cases. Then, I discuss some reasons why SIA is better-motivated than SSSA—and well-motivated compared to other possible special cases—in both finite and infinite contexts. Then I turn to the issues with infinities that arise in all current formalizations of anthropic reasoning, and propose some ways in which these issues may be resolved—while also showing that SIA is better suited than

[5] I use scare quotes because of course there is no time before I know the contents of my current experience. The logic of Bayesian reasoning applies here, but I never actually hold my "prior" credences—I just know what they would have been, if I didn't have the evidence from my current experience.

[6] And we are correct that the universe does not have some large-scale order making it such that not all of the physically possible structures are actual. For example, if the universe were periodic—for example, if moving 10^{25}km in the x, y or z directions brought you to a location with exactly the same physical structure—then it would not contain all possible physical structures.

SSSA for these adaptations. I use the dimensionality of spacetime as a motivating example, showing how these different epistemological choices have strong effects on our conclusions.

2 Anthropic reasoning

In order to discuss the issues with anthropic reasoning when infinities are involved, we need to formalize anthropic reasoning. In particular, we would like to have an equation for the posterior probability of a certain hypothesis after updating on our current experience, given our priors about which worlds, and which experiences within those worlds, are more likely. First, we need some vocabulary. An *observer-moment* is the experience of some observer during a period of time.[7] Each possible observer-moment belongs to a particular possible world. If σ is an observer-moment, then we let w_σ denote the possible world to which σ belongs. We don't know which possible world we belong to, i.e. which possible world is actual: if we did, then we would already know the truth of all the hypotheses about our world. Instead, our current experience is described by an *observer-type*: a set of observer-moments which are subjectively indistinguishable.[8] We do know that we belong to a certain observer-type, but we don't know which member of the equivalence class we are. So, as a first step towards finding the posterior probability of a hypothesis, we want a formula for the prior probability that we are a certain observer-moment; then we can sum over all the observer-moments belonging to worlds in which a certain hypothesis holds in order to find the prior probability of a hypothesis. Suppose we are finding the probability that we are observer-moment σ. The prior probability should be proportional to the probability that w_σ is actual—we are more likely to be observer-moments belonging to more likely worlds. It should also be proportional to the "weight", "measure", or prior probability we assign to the observer-moment itself, here denoted as $\mu(\sigma)$ for observer-moment σ. This should be separate from the prior probability of the overall world since, if there are two observer-moments in the same possible world, we may still have a prior that we are more likely to be one than the other.[9] Why denote this with a new symbol, μ, instead of just something like $P(\sigma|w_\sigma)$, i.e. the prior probability of my current observer-moment being σ given that w_σ is actual? After all, if μ were just $P(\sigma|w_\sigma)$, then the prior probability of our observer-moment being σ would just be the prior probability of the world w_σ being actual times this conditional probability. However, it is very useful to allow μ not to be a conditional probability. I discuss two reasons.

Firstly, it allows us to make some more natural choices of notation. Bostrom holds that there is a set of experiences—a "reference class"—from which our experiences as being uniformly drawn.[10] Bostrom claims that the appropriate reference class does not include all observers. For example, we may think that there are some aliens that are observers, but are so different from us that we could not have been one of these aliens; so, we exclude these aliens from our reference class. To encode this, we set μ to be a positive constant for experiences in our reference class and 0 for experiences outside. We also might want to encode much more fine-grained distinctions using μ. For example, we might think that, all else being equal, we are more likely to be observer-moments that are more lucid—i.e. fully awake adult humans rather than sleeping humans, insects, or rocks. There is no strict binary classification between observers and non-observers—for example, a fertilized human egg, since it has no brain functions, is not an observer, whereas an adult human clearly is; but one develops smoothly into the other, with no sharp dividing line where observerhood is attained. So, since insects and rocks also lie somewhere along this continuum, it would be very arbitrary to grant

[7] As we will see later, the exact length of this period doesn't matter very much.

[8] Again, nothing much turns on the exact definition of "subjectively indistinguishable"; it could be experiences that are identical (aside from haecceity; in other words, there may be two experiences whose qualities are the same, but which are still different experiences, for example because they arise from different regions of spacetime; these are considered identical on this definition), or experiences for which the observer has the same knowledge state, or something in between. All that matters is that α does *not* include any pairs of observer-moments which are able to judge that they differ from each other in their qualities. On some views, this can allow experiences that differ drastically in their actual character to both be included in α, if they only differ in their "inscrutables"—their non-structural properties. The structural properties are those that can be deduced from the structure of the substrate giving rise to the observer-moment. The non-structural properties are all the other properties of the observer-moment. (Proposed non-structural properties include, for example, qualia.) Such properties by definition cannot affect the knowledge state, since the knowledge state is structural: whether someone knows something can be deduced from the physical structure of their brain (on the safe assumption that interactionist dualism is false).

[9] As we will see later, the choice of μ is the most important part of a theory of anthropic reasoning—and indeed, the only part of such a theory which does not follow logically from the laws of probability.

[10] Or at least such that we should assume our experiences are uniformly drawn in order to perform anthropic reasoning; there is no commitment to any sort of actual selection mechanism.

them zero observerhood; instead, we may want to grant them extremely small amounts of observerhood.[11] Since a lucidity prior is about the character of the experience associated with σ, it treats subjectively indistinguishable observer-moments identically; there is a noticeable difference between more and less lucid experiences. But, the conditional probability of two subjectively indistinguishable observer-moments in different possible worlds may not be the same. If world w_1 contains one observer-moment σ of observer-type α and world w_2 contains two observer-moments σ_1, σ_2 of observer-type α, then, since we know that we are an observer with certainty, $P(\sigma|w_1) = 1$ while $P(\sigma_1|w_2) = P(\sigma_2|w_2) = \frac{1}{2}$. However, if we use μ, then we can set $\mu(\sigma) = \mu(\sigma_1) = \mu(\sigma_2)$, capturing the intuitive logic of our choice of prior.

Secondly, using μ will be convenient in the infinite setting. If we take the case above, except w_2 has infinitely many observer-moments σ_k of observer-type α, then for all k, $P(\sigma_k|w_2) = \frac{1}{\infty} = 0$; but this cannot be correct, since now the probabilities to not sum to 1. Although we cannot avoid this problem, we can avoid encountering it for now: we can still, in this case, let $\mu(\sigma) = \mu(\sigma_k) = 1$ without issue.

So, the prior probability that I am observer-moment σ is proportional both to $\mu(\sigma)$ and to the prior probability that w_σ is actual; all that remains is to ensure that this is appropriately normalized. Since there can be exactly one actual world[12], and I can be exactly one observer-moment, the events that each world is actual are all mutually exclusive, and so are the events that I am each observer-moment within some possible world. Since the probability that I am observer-moment σ is proportional to $P(w_\sigma)\mu(\sigma)$, the probability of being a member of the union of all possible observer-moments—which must be 1, since we take as given that I am an observer-moment—is proportional to $\sum_\beta P(w_\beta)\mu(\beta)$, where β is an observer-moment and the sum is over all possible observer-moments. Let $\Omega(w)$ be the class of all observer-moments in possible world w.[13] Note that we can partition the possible observer-moments in this sum by the possible worlds they belong to: $\sum_\beta P(w_\beta)\mu(\beta) = \sum_w P(w) \sum_{\beta\in\Omega(w)} \mu(\beta)$, where $\sum_w$ sums over all possible worlds. We observe that this is the expected total measure of actual observer-moments; we will, for convenience, denote it by E. Since this sum must be 1 when normalized, to normalize $P(w_\sigma)\mu(\sigma)$, we just divide by E. In other words, the prior probability that I am observer-moment σ is the expected fraction of all experiences which experience σ constitutes—which is indeed intuitively what the prior probability that our current experience is σ should be. We can express this mathematically as the *Anthropic Assumption*:[14]

$$P_\alpha(\sigma) = \frac{P_\alpha(w_\sigma)\mu_\alpha(\sigma)}{\sum_w P_\alpha(w) \sum_{\beta\in\Omega(w)} \mu_\alpha(\beta)} = \frac{P_\alpha(w_\sigma)\mu_\alpha(\sigma)}{E}. \tag{AA}$$

Here, our current experience is denoted by observer-type α. Unlike above, this equation has α subscripts on P and μ. This is to indicate that these reflect the credences of the agent performing anthropic reasoning, which is in this case α. However, note that P_α is still a "prior" credence function, i.e. the credence function as it would have been if we had no information about the actual nature of our current experience α. When a world w is an argument of a credence function, it is interpreted as the proposition "world w is actual"; when an observer-moment β is an argument of a credence function, it is interpreted as the proposition "observer-moment β is the actual observer-moment corresponding to my current experience"; when an observer-type α is an argument of a credence function, it is interpreted as the proposition "observer-type α is the observer-type of my current experience".

[11] For this reason, I reject that reference classes as Bostrom describes them (where they are completely binary) are plausible.

[12] A multiverse theorist should consider possible worlds to denote possible multiverses, not possible worlds within a possible multiverse.

[13] Note that I use Ω in a slightly different way from Bostrom—his $\Omega(w)$ includes all the observer-moments in w which are *in my reference class*, whereas mine includes all of them. This is because I introduce μ as a prior over experiences; this makes my formulation more general than Bostrom's, since Bostrom's can be recovered by setting $\mu = 0$ to all observer-moments outside my reference class, and a constant value to all observer-moments within my reference class.

[14] We see that this equation is not sensitive to the exact time considered to be a "moment", since changes in length can be compensated for by changes in the measure μ. For example, suppose that each observer-moment β corresponds to the same length of time and has the same measure $\mu(\beta) = c$. Then, if we halve the time of a "moment", we can maintain the original result by also halving μ: if $\Omega(w)$ is the original set of observer-moments, $\Omega'(w)$ is the new set of observer-moments where each observer-moment in $\Omega(w)$ has been split in two, μ is the original measure, and μ' is the new measure, we have $\sum_{\beta\in\Omega(w)} \mu(\beta) = |\Omega(w)| \cdot c = 2 \cdot |\Omega(w)| \cdot \frac{c}{2} = |\Omega'(w)| \cdot \frac{c}{2} = \sum_{\beta'\in\Omega'(w)} \mu'(\beta')$. The theory can also handle observer-moments of different temporal lengths, since μ allows longer moments to be weighted more heavily. However, we generally assume that all observer-moments are of a fairly short duration, so that an observer at any one time can know what their observer-moment consists of. We can even think of observer-moments as being instantaneous; then the sum over β in the equation for the Anthropic Assumption (AA) becomes an integral, $\int_{\Omega(w)} \mu(\beta)d\beta$. This idea is discussed in more detail in section 4.

Now, we use this assumption to deduce a generalized version of the "Observation Equation" proposed by Bostrom[1]. Below, let P_α^{An}—where the "An" stands for "anthropic"—denote the "posterior" credence: the credence of α after taking into account the information given by the nature of the observer-type α, i.e. the information that we have about the world based on what our subjective experience is. Let h denote some hypothesis. Let Ω_h denote the class of observer-moments for which h holds.

$$\begin{aligned}
P_\alpha^{An}(h) &= P_\alpha(h|\alpha) && \text{Definition of } P_\alpha^{An} \\
&= \frac{P_\alpha(h \cap \alpha)}{P_\alpha(\alpha)} && \text{Conditional probability} \\
&= \frac{\sum_{\sigma \in \Omega_h \cap \alpha} P_\alpha(\sigma)}{\sum_{\sigma \in \alpha} P_\alpha(\sigma)} && \text{Split into observer-moments} \\
&= \frac{\sum_{\sigma \in \Omega_h \cap \alpha} \frac{P_\alpha(w_\sigma)\mu_\alpha(\sigma)}{E}}{\sum_{\sigma \in \alpha} \frac{P_\alpha(w_\sigma)\mu_\alpha(\sigma)}{E}} && \text{Apply the AA} \\
&= \frac{\sum_{\sigma \in \Omega_h \cap \alpha} P_\alpha(w_\sigma)\mu_\alpha(\sigma)}{\sum_{\sigma \in \alpha} P_\alpha(w_\sigma)\mu_\alpha(\sigma)} && \text{Cancel } E \qquad \text{(GOE)}
\end{aligned}$$

Here I have named the final result the Generalized Observation Equation (GOE), since it encompasses Bostrom's Observation Equation (i.e. the SSSA Observation Equation) and the SIA Observation Equation as special cases. Bostrom's preferred choice of μ, deriving from the Strong Self-Sampling Assumption (SSSA) ([1] p.162), yields $\mu_\alpha(\sigma) = 0$ if σ is not in the reference class of α, and otherwise $\mu_\alpha(\sigma) = \frac{1}{|\Omega_\alpha \cap \Omega_{w_\sigma}|}$, where Ω_α denotes the reference class of α. In other words, Bostrom takes the total within-reference-class observer-measure of each possible world to be equal. On this choice of μ, the GOE becomes the SSSA Observation Equation ([1] p.173). My preferred choice of μ is to take $\mu(\alpha(\sigma)) = 1$. This yields a version of the Observation Equation deriving from what Bostrom calls the Self-Indication Assumption (SIA) ([1] p.66). So, in SIA, the total (within-reference-class) observer-measure of each possible world is proportional to the number of (within-reference-class) observers in that world.

This formalism is more general than the two formalisms that Bostrom proposes. I know of no published formalizations of anthropic reasoning that do not take this form for some choice of μ. This formulation also shows that, although SIA was originally developed only to avoid the Doomsday argument,[15] the Observation Equation for SIA takes on a very simple form—a form that, I would argue, is more simple and elegant than the form of the equation for SSSA.

3 Reference class issues in SSSA

There are other ways in which SSSA is less simple and elegant than SIA. SIA does not require a reference class—a subset of all possible observers that we take to be the only observers we could have been—to be useful for anthropic reasoning. SSSA, however, relies crucially on the choice of a reference class. Not only is this choice arbitrary, but many desirable choices do not allow us to make many useful deductions. I will show that two such desirable choices of reference class render SSSA useless—(1) the "minimal reference class" of subjectively identical experiences, and (2) a reference class that excludes a large number of intelligent, non-Boltzmanian observers (observers which arise in normal ways, rather than appearing out of chaotic fluctuations). It is clear why reference class (1) might be desirable—since it has clear, non-arbitrary criteria for inclusion, it is much better-motivated than any other choice of binary reference class (except the "maximal

[15]The Doomsday argument is described in section 5. SIA avoids the Doomsday argument by denying one of the premises of the Doomsday argument: while the Doomsday argument requires that we believe that we are typical humans (premise (2) in section 5), under SIA our "early" birth order provides evidence that we are not typical humans. This is because the total observer-measure of species with many individuals is greater than the total observer-measure of species with few individuals. So, we are more likely to be human if there will be many humans. This effect is of precisely the right size to cancel out the Doomsday argument. This is intuitive, because the measure allocated to each observer-moment in SIA does not depend on the number of other observers in our reference class; μ is always 1, instead of being inversely proportional to the number of observers in our reference class in the world in question. So, it is no surprise that we cannot deduce anything about the number of observers using SIA.

reference class", containing all experiences). Option (2) is less clearly desirable, but reference classes of type (2) are most of the reference classes that are not so broad as to be nearly equivalent to the maximal reference class. Boltzmann brains are observers arising at random in regions of high entropy. Although these observers are extraordinarily rare, modern physics suggests that the universe is spatially infinite and lacking in any complex global structure, so there should be infinitely many Boltzmann brains in the universe. Most[16] of these Boltzmann brains have confusing and chaotic experiences with little similarities to our own, but there are also infinitely many Boltzmann brains which are subjectively identical to humans—even infinitely many Boltzmann brains which are subjectively identical to my current experience. In order to exclude from our reference class a large number of Boltzmann brains (so that it is not essentially equivalent to the maximal reference class) and include almost all non-Boltzmann brains (so that it does not fail to make predictions by the argument I spell out below), we would need to have a very fine-tuned—and hence implausible—reference class.

Now, I show why reference classes of type (1) and (2) render SSSA useless. Below I use the following notation:

- α is the proposition that I experience my current experience (or have my current evidence).
- H is the proposition that the world is hospitable to life.
- I is the proposition that the world is inhospitable to life. (In this case, I'm a Boltzmann brain.)
- N_H is the expected number of actual, non-Boltzmanian observer-moments belonging to my observer-type, given a hospitable universe.
- N_I is the expected number of actual observer-moments belonging to my observer-type, given an inhospitable universe.
- R_H is the expected number of actual observer-moments in my reference class given a hospitable universe.
- R_I is the expected number of actual observer-moments in my reference class given an inhospitable universe.
- E is the prior expected number of observers as defined above.

I assume that there are the same number of Boltzmann brains in the hospitable and inhospitable universes. Below, P_{SSSA} and P_{SIA} denote the probabilities given by Bostrom's two proposed formalizations of anthropic reasoning; as mentioned in section 2, they correspond to setting $\mu_\alpha(\sigma) = \frac{1}{|\Omega_{w_\sigma}|}$ and $\mu_\alpha(\sigma) = 1$ in the GOE.[17]

$$P_{SSSA}(\alpha|H) = \frac{N_H + N_I}{R_H + R_I} \qquad \text{Identical fraction of ref. class (hospitable)}$$

$$P_{SSSA}(\alpha|I) = \frac{N_I}{R_I} \qquad \text{Identical fraction of ref. class (inhospitable)}$$

$$\begin{aligned} P_{SSSA}(H|\alpha) &= \frac{P_{SSSA}(\alpha|H)P(H)}{P_{SSSA}(\alpha)} && \text{Bayes' theorem} \\ &= \frac{P_{SSSA}(\alpha|H)P(H)}{P_{SSSA}(\alpha|H)P(H) + P_{SSSA}(\alpha|I)(1-P(H))} && \text{Split into hospitable, inhospitable cases} \\ &= \frac{\frac{N_H+N_I}{R_H+R_I}}{\frac{N_H+N_I}{R_H+R_I} + \frac{N_I}{R_I} \cdot \frac{1-P(H)}{P(H)}} && \text{Substitute values from above} \end{aligned}$$

[16] On the intuitive definition of "most", noting that all of these numbers are infinite.

[17] The derivation is short and is thus included, but only the final result of this computation will be used in the following discussion.

$$= \frac{1}{1 + \frac{\frac{R_H}{R_I}+1}{\frac{N_H}{N_I}+1}\left(\frac{1}{P(H)} - 1\right)} \qquad \text{Simplify}$$

$$P_{SIA}(\alpha|H) = \frac{N_H + N_I}{E} \qquad \text{Identical \#observers, normalized (hospitable)}$$

$$P_{SIA}(\alpha|I) = \frac{N_I}{E} \qquad \text{Identical \#observers, normalized (inhospitable)}$$

$$P_{SIA}(H|\alpha) = \frac{P_{SIA}(\alpha|H)P(H)}{P_{SIA}(\alpha|H)P(H) + P_{SIA}(\alpha|I)(1 - P(H))} \qquad \text{Bayes' theorem and split cases}$$

$$= \frac{1}{1 + \frac{1}{\frac{N_H}{N_I}+1}\left(\frac{1}{P(H)} - 1\right)} \qquad \text{Substitute and simplify}$$

Note that the results using SSSA and SIA are similar, except that where SSSA has $\frac{R_H}{R_I}$, SIA has 0.

If we use the minimal reference class—the reference class of subjectively indistinguishable observer-moments—then $R_H = N_H$ and $R_I = N_I$. Then

$$P_{SSSA}(H|\alpha) = \frac{1}{1 + \frac{\frac{N_H}{N_I}+1}{\frac{N_H}{N_I}+1}\left(\frac{1}{P_{SSSA}(H)} - 1\right)} = \frac{1}{1 + \left(\frac{1}{P_{SSSA}(H)} - 1\right)} = P_{SSSA}(H).$$

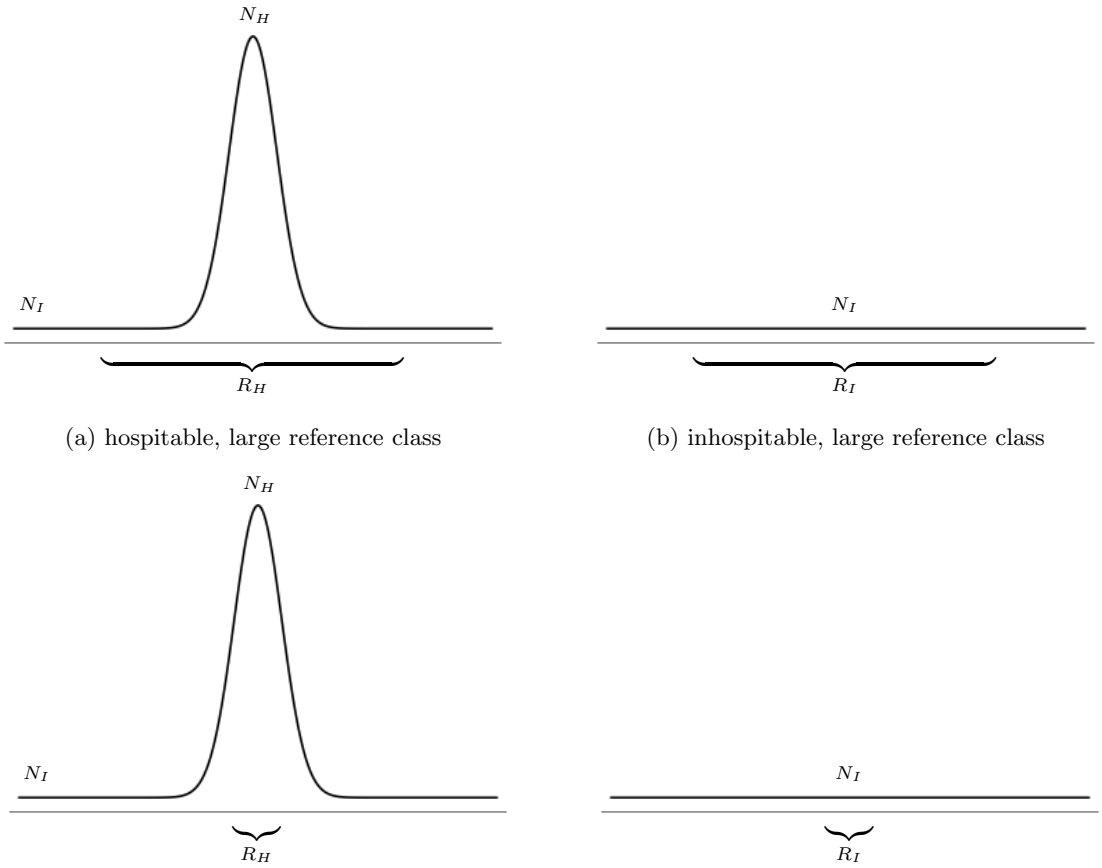
So the self-locating evidence α changes nothing. Intuitively, this is because, in choosing our reference class to be our observer-type α, we are taking the *a priori* probability that our observer-type is α to be 1. So, conditioning on our observer-type being α, as a proposition with probability 1, does nothing. We even get the same result if we just have that $\frac{R_H}{R_I} = \frac{N_H}{N_I}$, i.e. that the ratio of non-Boltzmanian to Boltzmanian observers is the same in both our reference class and our observer-type. But this result is unacceptable. The hypothesis H was just that the world is hospitable to life. If the world is not hospitable to life, then we must be Boltzmann brains. So, this result says that our observation that we do not seem to be Boltzmanian—that we seem to live in a very structured world instead of an ephemeral bubble of almost-order—should not in any way change our assessment of the probability that we are Boltzmanian. This is absurd.

To avoid this—i.e. in order to have $P(H|\alpha) > P(H)$—we need $\frac{R_H}{R_I} < \frac{N_H}{N_I}$. This holds in cases like the example Bostrom gives in *Anthropic Bias* for gaining evidence for the temperature of the CMB from measurement. This case is illustrated in the first pair of plots in figure 1. Bostrom says that, in this case, the reference class contains a sharp spike (with small width w_1) of observer-moments measuring a temperature close to the true temperature (which is also, we assume, the temperature that I measure), and a wide base (with width limited by the range of temperatures I take as belonging to my reference class, which we take to have large width w_2) of Boltzmanian observers with uniformly distributed measured temperatures. This is a fairly good approximation for a large reference class, containing many Boltzmanian observers. Then $R_H \approx w_1 N_H + w_2 N_I$: in a hospitable world, the number of observer-moments in my reference class is the width of the spike times its height plus the width of the base times its height.[18] Also, $R_I \approx w_2 N_I$: in an inhospitable world, the number of observer-moments in my reference class is just the width of the base times its height. Then

$$\frac{R_H}{R_I} = \frac{w_1 N_H + w_2 N_I}{w_2 N_I} = \frac{\frac{w_1}{w_2} N_H + N_I}{N_I} \ll \frac{N_H}{N_I},$$

i.e. the proportion of my reference class that measures the right temperature is much less than the proportion of the set of observers exactly like me that measures the right temperature. Let c be the small constant such

[18] This neglects the quantity $w_1 N_I$, corresponding to the region of the base which falls below the spike and is double-counted, but this is very small.



(a) hospitable, large reference class

(b) inhospitable, large reference class

(c) hospitable, small reference class

(d) inhospitable, small reference class

Figure 1: The first pair of distributions are for a large reference class, involving almost all non-Boltzmanian observers. This is the example that Bostrom uses to demonstrate the utility of SSSA. The second pair of distributions are for a small reference class, omitting many of the non-Boltzmanian observers. SSSA does not work well in this case.

that $\frac{R_H}{R_I} = c\frac{N_H}{N_I}$. Then we can plug this into the equation above to find that

$$P_{SSSA}(H|\alpha) \approx \frac{1}{1 + \frac{c\frac{N_H}{N_I}+1}{\frac{N_H}{N_I}+1}\left(\frac{1}{P(H)} - 1\right)} \gg \frac{1}{1 + \frac{\frac{N_H}{N_I}+1}{\frac{N_H}{N_I}+1}\left(\frac{1}{P(H)} - 1\right)} = P_{SSSA}(H).$$

So here we do indeed find that our seeming non-Boltzmanianness does reflect a lower chance of actually being Boltzmanian.

However, this assumption about our reference class is not necessarily satisfactory. In particular, if our reference class is of type (2) (as illustrated in the second pair of plots in figure 1)—it does not contain essentially all the experiences of measuring the correct temperature, or equivalently it excludes a significant number of non-Boltzmanian observers—then R_H will instead be proportional to N_H, while R_I is still proportional to N_I (with both having the same proportionality factor of the short width of temperatures belonging to the reference class), so that $\frac{R_H}{R_I} = \frac{N_H}{N_I}$; as we saw above, this means that we learn nothing from our evidence.

So, on reasonable assumptions about the distribution of observers, the reference classes of type (2) still do not let us learn from our evidence. In particular, reference classes which are a strict subset of the set of awake, adult human observers seem unlikely to provide much useful information; reference classes which are a strict subset of intelligent life also seem unlikely to provide much useful information, since it seems plausible that intelligent life clusters together in many salient dimensions (for example, their observations are a fairly reliable guide to some aspects of their environments, and are conducive to their continued survival),

whereas Boltzmann brains are fully random. So, if—as Bostrom suggests as a possibility—intelligent aliens do not belong to our reference class, then we may learn little from anthropic reasoning, and be left unable to reach basic conclusions such as the conclusion that our impression that we are not Boltzmanian actually suggests that we are not Boltzmanian.

Note that one natural choice of reference class—the maximal reference class, consisting of all experiences—was excluded from the analysis above. This is because this reference class does allow us to make predictions. However, adopting this reference class makes SSSA nearly equivalent to SIA! Within this reference class, all experiences are weighted equally. But experiences of some sort are nearly universal. As I argued above, there is no specific point in the development from a fertilized egg to a human where experience or observerhood suddenly appears where it did not previously exist at all. So, fertilized eggs must have experiences—and since the reference class does not allow us to weight these experiences less, they are equally counted as observers. Then, we can imagine removing tiny pieces from the egg one-by-one, until we are left with just separate electrons and quarks. Although at the very small end there could be a cutoff where experience disappears, it would be incredibly arbitrary to choose this cutoff to occur at scales larger than a few particles. Therefore, the overall result is that the number of experiences in a universe is a function of the number of very small objects in the universe.[19] So, the μ given by SSSA with this reference class is one over some function of the mass. When comparing universes with the same amount of mass, this is just SIA. For universes with different amounts of mass, it differs, in that it weighs experiences in larger universes less. But this is just density! As I will discuss below, we have good reasons in the infinite case to move to density of experiences instead of total number of experiences. Since, on the one non-arbitrary choice of reference class for SSSA, SSSA becomes identical to SIA with density instead of total number, this provides compelling evidence that SIA is the better formalization.

4 Density instead of total number?

The equations I have proposed rely heavily on the assumption that the total measure of experience $P_\alpha(w_\sigma)\mu_\alpha(\sigma)$ is finite when summed over all possible experiences (in the AA) or at least over all experiences in a given observer-type (in the GOE). However, this is a questionable assumption in general. Presumably my current observer-moment has some finite, nonzero measure. However, if the universe is spatially infinite, then there are infinitely many observers exactly like me. So, the total measure across all of these observers is infinite; *a fortiori*, even the total measure of all actual experiences of a given observer-type is infinite. However, we want to apply anthropic reasoning in the real world; so a good theory of anthropic reasoning should be able to handle these cases. I discuss moving from the *number* of observers in an infinite universe to the *density* of observers in an infinite universe. This might allow us to generalize the principles above to infinite universes: instead of SIA prioritizing more populated universes, it might prioritize denser ones. This will yield the same answer in finite cases, since it is equivalent to dividing μ by a constant representing the size of the universe, which we can see will then cancel out in both the AA equation and the GOE. Unfortunately, we cannot naively make the same move in the infinite case, since the size of the universe and the number of observers within it will both be infinite, so the result is undefined.[20] However, we can still use density in some infinite cases, and these answers will roughly correspond to intuition as long as we believe that all infinite universes we are comparing have the same "size", to the extent that size makes sense.[21]

[19]There may seem to be a hint of paradox here—after all, the human body and brain is made up of cells and fundamental particles, so this system should, to be consistent, count each of these as observers, as well as the overall brain. The question of how experiences arise from matter is far beyond the scope of this work. (For a discussion of the problem, see [10].) However, the logic works if we let experiences correspond to the most fundamental particles—in effect making the measure of experience proportional to mass—or if we have a hierarchical system where fundamental particles have experiences, but so do larger structures involving them (as proposed in [9]), noting that here the total number of experiences is also a function of mass—just, instead of being linear, it is exponential (from the exponentially many subsets of particles).

[20]Surreal numbers may allow us to concisely formalize this, since a sum of infinitely many infinitesimals can be well-defined. So, a surreal version of anthropic reasoning could have μ be as in the finite case but divided by the infinite "size" α of a universe—hence, infinitesimal—and the sum over all of the infinitely many observer-moments in this universe of μ times the probability could then come out to be finite. However, this may not be desirable, since it requires a very sensitive notion of different infinite sizes of universes. As I show below, there are other ways we might want to measure density that do not require this strange sensitivity to different sizes of infinities. So, I do not pursue this formalization further.

[21]I give an example of a case where we might intuitively feel that two infinite universes are not the same size. Consider a sequence of universes of size 2^n, and a sequence of universes of size 2^{n+1}. In both of these sequences, the $(n+1)$st universe

However, there are many problems with defining density in a useful way. For one, as discussed by Arntzenius and Dorr, if the universe has an unusual structure—either built into its physical laws or just because of the way observers happen to be arranged within it—then there may not be a clear way to define the global density of observers.[4] (I show an example of this below.) However, in many cases, a global observer-density can be defined, at least up to a choice of a volume metric (a way of assigning volumes to any finite region of space that respects the expected rules such as the volume of the union of two disjoint regions equalling the sum of the volumes of the original regions, etc.). This is because, if we have such a volume metric, we can use it to define a local observer-density metric which assigns an observer-density to any region of space: we can just sum the measures of all the observers within a given region R (or find the integral $\int_R \mu(\beta)d\beta$) and divide the result by the volume of R, as long as the total observer measure is not infinite on any finite region of space. Then, we can use this local observer-density metric to find the global observer density in a connected, n-dimensional universe, with a *bounded* observer-density (i.e. an observer density that is not just everywhere finite, but does not exceed a certain finite value everywhere in space[22]): pick any point in the universe, and find the observer-densities of the n-dimensional balls centered at this point. The limit of the density as the radius grows arbitrarily large is then, if it exists, the global density. Note that in a connected universe, this is (if it is well-defined) equal regardless of the selected initial point, since, given any two points within the universe, the radii of balls centered at these points can be increased until the balls overlap on almost all of their area. Then, since the local observer-density is bounded, the observer-measure in the non-overlapping regions become arbitrarily small relative to the observer-measure in the overlapping region, and both densities become identical in the limit. This allows us to define density in some cases where Arntzenius and Dorr cannot. For example, my figure 2 copies their figure 7.

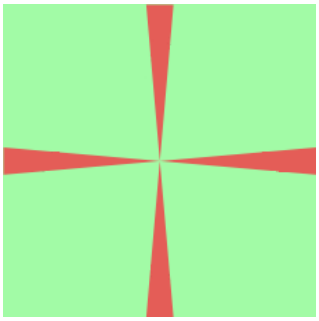


Figure 2: The red regions are within five degrees in either direction of one of the cardinal directions measured from the origin. I consider the local observer-density within green regions to be 0 and within red regions to be 1.[4]

Since the observer-density is clearly bounded here, we are free to consider balls centered at the origin. Then it is clear that the observer-density is $\frac{(360-10\cdot4)\cdot0+10\cdot4\cdot1}{360} = \frac{1}{9}$. Arntzenius and Dorr point out that a sequence of nested squares does not converge to the same result. But, in fact, unless the orientation of the squares remains consistent, a sequence of nested squares doesn't converge to a density *at all*. And, without a reason based on the physics of the system—for example, if the physics of the system was not isotropic, instead privileging some directions—then there is no reason to think that such a quantity has any real significance. (Furthermore, the most obvious way of turning this into a quantity that does have significance—averaging the results for consistently-oriented squares over all the possible square orientations—yields the same quantity

contains the nth universe as a subset—perhaps the universes are balls, and going to the next step in the sequence is adding a shell around the original ball, like an onion. The nth universe in the first sequence is always half the size of the universe in the second sequence, so if the universes have the same observer-density, then there will be twice as many observers in the second universe, so we should take it to be twice as likely that we belong to the second universe than the first for any n. But what about the limits of these sequences, i.e. the universes we get by taking the union over the whole infinite sequence? Our intuition might say that, since each pair in the sequence differs in size by a factor of 2, then so should the limits. My arguments here require us to deny this. If there is a way in which infinite universes can differ in size like this, then we would have to include it in our priors, without referring to the "size" of these infinite universes.

[22]This is not an excessively strict restriction; it seems plausible that it holds in our universe, since, at high enough mass densities, singularities form, and it seems unlikely that there can be arbitrarily large observer-densities with bounded mass densities, and it seems unlikely that there could be observers within singularities.

we found above using balls.)

However, there are still cases where the observer-density remains ill-defined. Consider a universe consisting of alternating populated and unpopulated rings centered at some point, with a small unpopulated disk forming "ring 0" at the center (so that there isn't an infinite number of rings within a finite area near the origin). Each ring has an area equal to the sum of the areas of the rings contained within it. Let $D(n)$ indicate the observer-density of the ball whose radius coincides with the outside of ring n. Then $D(0) = 0$, $D(2n) = \frac{D(2n-1)}{2}$, and $D(2n+1) = \frac{D(2n)+1}{2}$. The even terms of this sequence converge to $\frac{1}{3}$ as $n \to \infty$, while the odd terms converge to $\frac{2}{3}$. Thus the limit does not exist, and this method does not yield a global density. However, our universe likely does not suffer from a similar issue, since it is very unlikely that it has the sort of large-scale structure that would be required for the density limit to not converge.

So, if we have a volume metric, we can in most non-pathological cases define a global observer-density. But how do we choose a volume metric? Our choice should depend in a substantive way on the structure of the universe, since otherwise there is no reason to believe that the densities of different universes are comparable—for example, we certainly don't want to have universes that are identical except that in one, all physical structures (like electrons, protons, etc.) are 50% smaller than in the other according to the volume metric. We have no reason to believe that there is some objective fact about length scales that could make this plausible. We also might like to have universes with different numbers of dimensions be comparable. If we use the same method of measuring volume for n and $n+1$ dimensions, then the volume of any finite region in n dimensions will be infinitely smaller than the volume of any finite region in $n+1$ dimensions (i.e. the former is finite and the latter is infinite, or the former is zero and the latter is finite). It would be strange—although not impossible—for this to be an appropriate method of measuring density. It would overwhelmingly privilege smaller dimensions, since each observer occupies infinitely more space in $n+1$ dimensions, so $n+1$ dimensional spaces have infinitely lower observer-density than all n-dimensional spaces which have nonzero observer-density. So, on this view, it may be surprising that we live in three spatial and one time dimension, when it seems that life may be possible in two spatial dimensions. As I will discuss later, we indeed have reason to believe that life is possible in two spatial dimensions, which suggests that this is not a good choice of density metric.

One obvious avenue for resolving the issue of defining volume in a non-arbitrary way is to use *information*. This neatly sidesteps issues of dimensionality, so it is especially promising as a foundation for a volume metric. We consider only universes where we have a volume metric that is internally consistent and well-motivated; we only want to figure out how to compare density metrics across different universes. Consider universes where there is a finite upper bound to the information density of space. (Something like this seems plausible in our universe; the objects with the greatest entropy are black holes, and the entropy of a black hole is proportional to the surface area of its event horizon; so, there is a finite entropy density.[23]) Suppose that we have a pair of universes with volume metrics v and v'. Choose c and c' so that the maximally complex possible region of space R in the first universe with $v(R) = c$ and the maximally complex possible region of space R' in the second universe with $v'(R') = c'$ both require n bits of information to represent. (Of course, the exact number of bits required depends on the exact representation scheme used. However, in the complex-system limit, different representations of information can be converted between each other; although one might be half as parsimonious as another (for example if every other bit is always 0), the systems considered complicated by one will be considered complicated by the other as well.[24]) Then we can define a new metric for the second universe, $\tilde{v} = v' \frac{c}{c'}$; then $\tilde{v}(R') = c' \frac{c}{c'} = c = v(R)$. So, v and $\tilde{v}$ both have the same relation to the number of bits required to represent a region of space; so, we have some reason to believe that they might allow cross-universe comparisons for anthropic reasoning.

Let's see how this works in an example. Suppose that there is a universe U consisting of a 1-dimensional tape of cells which can be in 2 states: 0 or 1. Suppose also that there is a universe U' consisting of a 2-dimensional grid of cells which also have the 2 states 0 and 1. Let the volume metric v for the 1-dimensional tape assign volume 2 to each cell, and let the volume metric v' for the 2-dimensional grid assign volume 1 to each cell. Now we need to consider how to represent the information contained within a region. Although there are

[23]Note that there is some strangeness with dimensionality here—the black hole entropy is *not*, as one might expect, proportional to the volume contained within the event horizon.

[24]I'm pretty sure I remember reading about this, but couldn't find a source.

many ways to do this, it is clear that, in general, we need one bit per cell (so, each bit represents the state of the cell). So, the maximally complex possible spatial region of m cells must, according to any method of measuring information, require at least 2^m bits to represent (although it may require more); it cannot require fewer since that would render us incapable of distinguishing all 2^m possible states on this region. Let's arbitrarily take $c = 4$ (noting that only the ratio of c to c' matters.). Then we have $v(R) = c$; the maximally complex region consists of $\frac{4}{2} = 2$ cells (since we assign volume 2 to each cell), and requires 2 bits to represent. So, in the 2-dimensional grid, we want the maximally complex region to consist of 2 cells, so (since we assign volume 1 to each cell) we need $v(R') = 2$, so $c' = 2$. Then $\tilde{v}(r) = v'(r)\frac{c}{c'} = 2v'(r)$; we now assign volume 2 to each cell in the 2-dimensional case. But this is exactly what we might intuitively want, since both of the metrics now assign the same volume to each cell!

However, this does not fully solve the problem of defining density. If the universe is continuous, then the amount of information required to represent a region of space isn't finite. For example, consider a universe consisting of a meter-long line and a single point particle. The position of the point particle then is a real number in the interval $[0, 1]$. But almost all real numbers require an infinite number of bits to represent. Hence, such a system does not have a bounded information density, and the above scheme does not work. There may be ways to define information density in such systems regardless. At the very least, it seems like there may be a way to define information density in systems which can be approximated well by fewer bits. It seems plausible that there is somehow less information in the point-particle-along-rod system above than in the example proposed by Arntzenius and Dorr in their figure 8 where the size of structures (and hence observers) decreases exponentially in space, so that there is a region (the "cliff edge") near which infinite information density is achieved relative to other areas of the system ([4], p.51). This could be because the first system—at least if it obeyed physics somewhat like ours—would have dynamics that are approximated well by the system where the position is rounded to the 10th decimal place. The second system, however, cannot be approximated so well; if we decrease the spatial resolution an equivalent amount, then we drastically change the experience of the infinite number of observers that live arbitrarily close to the cliff face—for any resolution, no matter how small, there are an infinite number of observers smaller than that resolution, whose information is therefore completely lost in such an approximation.

Perhaps we can formalize a notion of how well a system can be approximated, and use this to define something like the information density based on how quickly approximations with certain information densities approach the behavior of the original system. However, it is not clear exactly how to do this, since there is not an obvious good way to judge how good an approximation is. An intuitive idea is to judge how much it affects the experiences of observers, but this has the uncomfortable implication that universes without observers can be approximated perfectly well by nothing at all. It is unclear whether we should bite this bullet and accept that all the relevant information about a universe is the information belonging to observers. It is not wholly impossible to do so—in panpsychist theories, for example, physical differences correspond closely to differences in observation—but in these theories, it also seems especially unlikely that we could define the closeness of an approximation to the actual world in terms of a difference in experience, because here we would expect that the possible states that observers can be in are also continuous. So, this is a conclusion that I am not at present willing to accept; but I hope that there may be a non-arbitrary definition of density that allows comparison across continuous universes.

5 Is the low observed universe population a problem?

Since SIA weights worlds with more observers more heavily, it leads us to expect a world that is highly or densely populated. However, our world seems not to be even close to as populated as it could be. Can we reconcile these facts? To answer this, we need to disentangle this statement into two related claims that SIA might make. The first claim is that SIA implies with high probability that, in the whole of reality, there are a large number of observers. The second claim is that SIA implies with high probability that, in the part of reality visible to us, there are a large number of observers. Although Bostrom takes issue with the first claim, I don't find it to be too problematic. Current theories suggest that the universe is infinite, and thus contains an infinite number of observers. So, if SIA implies that there are infinitely many observers, this would not contradict our evidence.

The second claim seems, at least initially, more concerning. It seems obvious that there could be a very large number of observers that we could detect. If the universe is infinite, there should be a part of it where every star has an Earth-like planet teeming with life, or even areas where most of the matter consists of sentient nanobots. If these regions contain many orders of magnitude more observers than our region, it might be surprising that we don't exist in such a region. And this problem is indeed specific to SIA: although SSSA does lead us to expect that we belong to one of the more densely populated *actual* areas since those areas contain more observers, SIA leads us to expect that we belong to one of the more densely population *possible* areas. However, it's possible that sparsely populated areas are vastly more common than areas populated almost to the theoretical limit. An estimate of the distribution of observer-densities is beyond the scope of this paper. Instead, I will investigate what a prior credence for the observer-density of a section of the universe about the size of the current observable universe would have to be like so that our observations are not absurdly unlikely.

The Doomsday argument in SSSA says that we should expect that the population we see is typical of populations. More specifically, the Doomsday argument goes as follows: (1) the typical human will have a typical birth order, i.e. will, with probability $1-p$ not belong to the earliest p fraction of all humans to ever live; (2) we have no reason to believe we are not typical humans; (3) if T is the total number of humans to ever exist, and 10^{11} is the number of humans born before us, then $pT = 10^{11}$, and $(1-p)T = \frac{1-p}{p}10^{11}$; so (4) we should conclude with probability p that there will be at most $\frac{1-p}{p}10^{11}$ humans living after us. (If $p = 0.05$, and the birth rate remains above 100 million (10^9) births per year until human extinction (the current birth rate per year is 131 million), then this means that humanity will, with probability 0.05, continue for no more than 15000 more years. So, if we accept the Doomsday argument, we should believe that humanity will suffer a massive population collapse sometime in the next few thousand years.)

We can "run this argument backwards" to find a prior that leads to the population size we observer being typical. We can see what posterior probabilities for human population sizes the Doomsday argument yields; then, we can use this to construct a prior such that our observations are not implausible. Then we can assess whether this is a plausible prior.

Below let O be the total number of observers and R be the number of observers I am aware of. Let's start with a uniform prior: $P(O = N) = \frac{1}{\Omega}$. Here we assume that the greatest possible value of O is Ω; this should be considered a high upper bound. It might be determined by factors like the size of the observable universe and the minimal size of an observer. As derived in the appendix in section 9,

$$P(O = N|R = M) = \frac{\chi_{M\leq N\leq\Omega}}{N\left(H_\Omega - H_M\right)} \approx \frac{\chi_{M\leq N\leq\Omega}}{N\ln\frac{\Omega}{M}}.$$

Note that falling off like $\frac{1}{N}$ is not unreasonable as a prior. (Also note that it's not normalized without Ω.) Then, also as derived in the appendix in section 9,

$$P(O \geq N|R = M) = \frac{H_\Omega - \gamma - \psi(\max(M,N))}{H_\Omega - H_M} \approx 1 + \frac{\ln\frac{M}{\max(M,N)}}{\ln\Omega}$$

Inspired by $P(O = N|R = M)$ above, let's pick the prior

$$P(O = N) = \frac{1}{N\ln\Omega}.$$

This is setting $M = 1$ there, which captures the general idea that our rank is a lot smaller than the total number of possible observers; all that really matters is that $\Omega \gg M$, since the exact choice of Ω is unclear.[25] Note that this is normalized since the antiderivative is $\frac{\ln N}{\ln\Omega}$, and $\frac{\ln N}{\ln\Omega}\Big|_{N=1}^{\Omega} = \frac{\ln\Omega}{\ln\Omega} - \frac{\ln 1}{\ln\Omega} = 1$. Using this prior,

$$P(O \leq N) = \frac{\ln n}{\ln\Omega}\Big|_{n=1}^{N} = \frac{\ln N}{\ln\Omega}.$$

[25]This is not well-defined for $N = 0$, but we are free to add this case, for example by replacing N with $N + 1$. I do not do this here for simplicity.

If $N = e^{23}$ (about 3.5 billion, the closest power of e below the current population of the earth) and $\Omega = e^{189}$ (an upper bound on the number of atoms in the observable universe, and hence a very rough upper bound for the number of possible observers that we could detect) then

$$P(O \leq N) = \frac{23}{189} \approx 0.1.$$

So we find that despite our very loose upper bound, if we take our spacetime geometry and approximate time in the universe as given, it's not so unlikely that we see so few observers!

This argument can be extended in a slightly less elegant form to the case with infinite universe using surreal probabilities.[26] (It may be possible to extend it just as elegantly, but the math in this case is more difficult.) Surreal numbers include the real numbers, but also a very large number of infinite and infinitesimal numbers. Shockingly, a great deal of standard mathematical operations on the real numbers—like addition, multiplication, and exponentiation—can be uniquely defined on the surreals. This allows us to work with infinities and infinitesimals almost as though they were real numbers.

The infinity of the naturals is denoted as ω. Suppose that we start with a prior f that we belong to a finite universe, and that, if our universe is infinite, it is at most the surreal number Ω, greater than the surreal number η and it is every intervening surreal number with equal probability. Then our credence that we belonging to a universe containing a surreal number α of observers is:

$$P^{SIA}(\alpha) = \begin{cases} \frac{f}{\alpha \ln \omega} & \alpha \in \mathbb{N}^+ \\ \frac{1-f}{\Omega - \eta} & \eta < \alpha \leq \Omega \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This is indeed a valid credence function; we see that the probability summed over all surreals α is

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{f}{n \ln \omega} + \frac{1-f}{\Omega - \eta} \cdot (\Omega - \eta) &= \frac{f}{\ln \omega} \sum_{n=1}^{\infty} \frac{1}{n} + \frac{1-f}{\Omega - \eta} \cdot (\Omega - \eta) \\ &= f + \frac{1}{\ln \omega}(\gamma + \frac{1}{2\omega}) + 1 - f \\ &= 1 + \frac{1}{\ln \omega}(\gamma + \frac{1}{2\omega}) \\ &\approx 1. \end{aligned}$$

where above we use the fact that $\sum_{n=1}^{N} \frac{1}{n} = \ln N + \gamma + \epsilon_k$, where γ is a constant and $\epsilon_k \sim \frac{1}{2k}$, and then neglect infinitesimal terms, since we can correct these by subtracting an equal infinitesimal amount from each case in our prior (but doing so is inelegant, so I have ommitted this step). To get the prior credences over worlds that yield this posterior, we just need to divide by the total measure of observers in each universe and renormalize (with renormalization factor γ equal to the sum of the part of the equation within brackets below over all surreal α):

$$P(\alpha) = \frac{1}{\gamma} \begin{cases} \frac{f}{\alpha^2 \ln \omega} & \alpha \in \mathbb{N}^+ \\ \frac{1-f}{\alpha(\Omega - \eta)} & \eta < \alpha \leq \Omega \\ 0 & \text{otherwise.} \end{cases}$$

So, this prior does hold that finite worlds are much more likely—$\ln \omega$ is generally much smaller than $\alpha(\Omega - \eta)$—but it does not seem excessively unreasonable (although we do have some justified reason to worry about our arbitrary upper and lower bounds on the possible infinite numbers of observers). Thus it is an example of how we need not conclude with probability 1 that our universe is infinite if we reason using SIA.

[26] My understanding of surreal probabilities is largely informed by Chen and Rubio[5] and Gonshor[6].

6 Presumptuous philosophers in SSSA

Although SSSA does not imply that the universe has infinite spatial extent, its issues with reference classes open it up to its own presumptuous philosopher arguments. Since presumptuous philosopher arguments are the main reason proposed for rejecting SIA, this means that SSSA, since it is also vulnerable to these arguments, has no clear advantage over SIA. Manley ([3], p.32) describes a thought experiment:[27] suppose that there are two competing physical hypotheses: (1) our universe will have an extremely long period where there are no complex structures but Boltzmann brains, so there will be trillions of times more Boltzmann brains than there ever were humans; and (2) our universe will collapse into a lower-energy state in which Boltzmann brains are not possible within the next 100 billion years, so that humans still vastly outnumber Boltzmann brains. Both of these are physically plausible given our current evidence. However, SSSA leads us to conclude that the latter is true with overwhelming probability, as long as our reference class includes a reasonable number of Boltzmann brains that are perceptibly different from natural life—which it must if it is to be useful at all, as I showed above. This is because, on hypothesis (1), only a tiny fraction of actual observers are non-Boltzmanian, and thus our reference class contains a reasonable number of Boltzmann brains which do not think they are not Boltzmann brains. Thus, in this case, only a tiny fraction of our reference class thinks that it is not a Boltzmann brain. Therefore, on SSSA, our current observation—that we do not seem to be Boltzmann brains—is extremely unlikely. However, on hypothesis (2), a large fraction of actual observers are non-Boltzmanian, so our reference class consists mostly of non-Boltzmanian observers. On this hypothesis, SSSA indicates that our observation is fairly typical. Thus, a philosopher ascribing to SSSA accepts hypothesis (2) with high probability! This seems to me to be at least as implausible of a conclusion as the conclusion that the universe is infinite.

SIA does not have this conclusion. It seems plausible that there is a time T until the universe ends (i.e. replacing 100 billion years above) such that, in case (2), the number of Boltzmanian observers that don't think they're non-Boltzmanian is large while the number of Boltzmanian observers that do think they're non-Boltzmanian is much smaller than the total number of humans. In this case, SIA leaves us neutral between the two possibilities: the existence of Boltzmann brains that think they are Boltzmanian does not decrease the probability that I am a non-Boltzmanian brain that thinks it is non-Boltzmanian. This is in contrast to SSSA, where $\mu(\sigma) = \frac{1}{|\Omega_\sigma \cap \Omega(w_\sigma)|}$, so the probability that I have any experience σ where I believe that I am non-Boltzmanian decreases as $\Omega_\sigma \cap \Omega(w_\sigma)$, which includes some Boltzmanian brains that believe they are Boltzmanian, increases. It may be possible that there is no such timescale, and the number of Boltzmanian brains within our reference class which think they are not Boltzmanian must outweigh the number of humans if the number of Boltzmanian brains within our reference class which do not think they are Boltzmanian outweighs the number of humans; in this case, the total number of observers matching our evidence is much higher in (1), so SIA leads us to accept (1). But this seems implausible since it is doubtful that an appreciable number of Boltzmanian observers in our reference class would believe that they were non-Boltzmanian. So here, SSSA leads us to strictly more physically unjustified conclusions than SIA.

7 Application: dimensionality of spacetime

7.1 Introduction

Now that we have developed an epistemology of anthropic reasoning, let's see how it can be applied. One case where we might hope to apply anthropic reasoning is to determine why the universe we observe has three spatial dimensions and one time dimension. Since this is just a fact about the structure of the universe, it seems unlikely that a causal explanation is possible. The dimensionality of spacetime might just be a brute fact. Although this is not impossible, it is undesirable; there are very many possible combinations of space and time dimensions, and the fact that the numbers of dimensions are so small is very strange if we think there is nothing determining those numbers. A simplicity prior might let us be unsurprised that these

[27] Manley's exact form is perhaps clearer, but less physically motivated. His example involves two possible universes which differ only in their number of green observers, and shows that SSSA leads us to believe that there are fewer green observers, even though we know that we are not green. My example shows how a situation like the one Manley describes arises from natural physical laws.

numbers are so small, but we are still then left in some confusion: why not zero spatial and time dimensions? Anthropic reasoning may provide an explanation for the dimensionality of spacetime. Tegmark [2] claims to have found just such an explanation. He claims that 3 spatial and 1 time dimension (or, isomorphically, 1 spatial and 3 time dimensions) is the only combination of dimensions that can give rise to universes containing life—or at least universes containing a great deal of life, instead of just a few Boltzmann brains. I find Tegmark's anthropic explanation for why there is exactly one time dimension—assuming that physical laws are second order linear PDEs—to be convincing. Tegmark's argumentation for the number of spatial dimensions, however, is less careful; I show below that there are some good reasons to doubt it.

Before discussing the particulars of Tegmark's argumentation, we need to establish epistemological fundamentals. Tegmark himself does not endorse a particular methodology for anthropic reasoning. It is interesting to note, however, that demonstrating that universes with a certain dimensionality can only contain sparse life—as Tegmark does—does not provide any reason to believe that we exist in a less sparsely populated universe without a more careful treatment of epistemology. Firstly, since universes of any dimensionality may contain infinitely many observers[28], if we treat all infinities of observers as identical, then proving that life is sparse in some universe is not enough to show that we should expect not to live in that universe. But if we move to density rather than overall number of observers, this problem disappears; we are then obviously less likely to belong to sparse universes than more populated ones. So, in this application, density provides the intuitively correct answer (and the answer that Tegmark assumes is correct without proof).

Secondly, if (1) we accept SSSA as our epistemology and (2) we believe that there is just one universe, with certain dimensions, instead of a multiverse, then we learn nothing from Tegmark's arguments. This is because, if there is no multiverse, then each possible world is normalized to have the same total observer-measure, as mentioned in section 2. But then, by construction, I am equally likely to exist in each possible universe which can support life at all—even if life is incredibly sparse. SIA does not have this problem, since in SIA I am drawn evenly from the set of observers without normalization by universe, I am more likely to belong to universes with more observers. Note that, if assumption (2) does not hold—worlds with all the dimensionalities under consideration are actual—then SSSA and SIA grow closer in their predictions, since now the observers belonging to less populated actual universes are crowded out by observers belonging to more populated actual universes. If modal realism is true—all possible worlds are actual[29]—then SSSA and SIA make the same predictions:

$$P_\alpha^{SSSA}(\sigma) = \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|} = \frac{1}{|\Omega_\sigma|}$$

because there is just one possible world, the modal realist multiverse, which occurs with probability 1 and contains all observers; and

$$P_\alpha^{SIA}(\sigma) = \frac{P_\alpha(w_\sigma)}{E} = \frac{1}{|\Omega_\sigma|}$$

because the expected measure of observers is just the measure of actual observers. So, we assume SIA reasoning (or a multiverse) from now on.

Tegmark makes a few assumptions about the nature of the universe. Firstly, Tegmark assumes that spacetime has a geometry with certain numbers of space and time dimensions. There are a vast number of mathematical structures which cannot be described (or cannot be described well) in this way—for example, discrete universes, universes without distance metrics, and universes which have different dimensionality in different areas; these are all ignored as possibilities, since it is almost impossible to show anything about such a wide class of structures. Secondly, Tegmark assumes that physical laws are given by second order linear PDEs. Again, there are very many mathematical structures that do not follow this pattern; and again they are ignored to make the problem tractable.

[28] Zero dimensions is included here, as we will see below; but we can also see it in a much weaker form by noting that there could be a different property of the single point corresponding to the state of each point in some higher-dimensional space.

[29] Or at least, if the ratios of actual observers in each of the hypotheses under consideration (here the different numbers of space and time dimensions) are the same as the ratios among possible observers

7.2 Turing machines

All the processes of physics—and hence also of life—are believed to be able to be modeled to arbitrary accuracy by some Turing machine.[30] Thus, Turing machines are sufficient—although perhaps not necessary—for life.[31] So, if a system is complex enough to allow Turing machines, we cannot take it to be inhospitable to life—either without life entirely, or containing only Boltzmann brains—without some specific reason why: all such systems can contain a model of a human brain which is, to arbitrary accuracy, computationally identical to a real one, so we cannot dismiss the possibility of life in such a universe until we show that Turing machines (or Turing machines simulating life) are rare. So, to show that these systems are in fact barren, we need to show that complex, life-like processes cannot arise through processes like natural selection. For example, we could show—as Tegmark [2] does for systems governed by second order linear PDEs with more or less than one time dimension—that the environment of a region is unpredictable given the contents of that region. This means that life is very unlikely to evolve, since, if the environment is relevant to the persistence of the organism, the organism must be able to interact with the environment in order to achieve the outcomes necessary for its persistence.[32] Tegmark does not, however, demonstrate such an explicit result for the cases where the number of spatial dimensions is not 3. He cites work showing that, in more than three spatial dimensions, matter is unstable—*given generalizations of our actual physical laws*, a very strong assumption that we cannot take for granted when considering the possible worlds we could have inhabited. He also suggests that, in fewer than three spatial dimensions, systems may be too simple for complex life. I argue that this latter case cannot be dismissed so quickly—stable Turing machines can be built in these systems, so they are capable of the complexity required for life, and this complexity can evolve without exceptional fine-tuning.

7.3 Zero dimensions

Bournez et al. [8] show that systems governed by differential equations of the form $y'(t) = p(y(t))$, where p is a vector of polynomials, are Turing complete, meaning that they can emulate any Turing machine. These systems are zero-dimensional—as ODEs, all the derivatives are in one variable: time. So, since polynomial vectors have finite length, the spatial structure of the systems modeled by these equations consists of a finite number of points.[33] For life to exist in such a system, the physical laws would need to be complicated. The vector of polynomials p governing the dynamics of this system might have a rather large length and contain polynomials with rather high degree.[34] Linearity, however, will remain unattainable, so the cases that Tegmark considers—restricted to be first-order, i.e. only containing first-order time derivatives, to match the first-order equation $y'(t) = p(y(t))$—indeed cannot support life in the zero-dimensional case. (Perhaps it will be shown that a Turing machine can be simulated in zero dimensions using second-order linear ODEs, but investigating this is beyond the scope of this paper.) This is because, if the equations are to be linear,

[30] This is a version of the Church-Turing thesis.

[31] There is good reason to believe that they are necessary—after all, humans can simulate Turing machines, so, in the absence of Turing machines, there are fairly severe limits on the complexity of life.

[32] If the environment is irrelevant to the persistence of an organism, then presumably there are little or no causal connections between the environment and some subset of the processes going on within the region in question. Then we should restrict our concern to the region in question and the processes within it that are not affected by the environment, and consider whether life could exist within this subsystem. Tegmark does not explicitly make this move, but it seems plausible that his arguments also mean that a brain cannot form without exceptional levels of fine-tuning: for a brain to work, its different regions need to interact in predictable ways in order to communicate information. In systems with zero time dimensions, adjacent regions do not interact at all, so evolution of a brain is clearly impossible. In systems with more than one time dimension, the behavior of the system is chaotic, in that arbitrarily small differences in inputs result in large differences in outputs. This means that a brain would need to start out in a very precise configuration in order to effectively communicate information internally for any appreciable period of time. But evolution is not precise: for a feature to evolve, there must be a wide region in configuration space around the feature where increasing the feature somewhat increases fitness. So, because in more than one time dimension the region in configuration space around functioning brains is very narrow, they cannot evolve. The exception is brains that exist at a single point. I will consider this zero-dimensional case below.

[33] Most naturally, they are just a single point. However, there can be additional geometry, but adjacent points will not interact in any way, since the laws only encode interactions between finite sets of points: for each point p, there is a neighborhood of p such that that none of the points within this neighborhood interact with p. So, this geometry is not relevant to the processes that can arise.

[34] Bournez et al. do not calculate upper bounds for these values for Turing completeness. However, there is no reason to believe that they would be particularly small, given the nature of their scheme for expressing Turing machines. It is possible—if, in my uninformed opinion, unlikely—that future developments will show these numbers to be small.

each polynomial must have degree at most 1. Then the solution to the differential equation above for each y_i takes the form $y_i(t) = \sum_i c_i e^{\lambda_i t}$. The result of a computation is read off by looking at the state of the system after a certain amount of time has elapsed. We don't require the state at a specific time; this is just a time after which the system is guaranteed to be within the required error bounds of the result, so the answer we determine is guaranteed to be correct. It is also acceptable to measure the system at any point after this. So, since the only input is t, there cannot be any dependence on the input, and so no computations can be performed.[35] So, while intelligent life in 0 spatial dimensions is not impossible and might be evolvable, we might take the laws required for its existence to be *a priori* unlikely.

7.4 One and two dimensions

The situation in one dimension is very different—not only are Turing machines possible here, but they can arise even given laws that we might take to be *a priori* plausible. The rule 110 cellular automaton is a discrete one-dimensional system which can simulate any Turing machine. The rule 110 cellular automaton consists of a discrete tape of cells which contain either a 0 or a 1 value. Time is also discrete; at every timestep, the value at each cell is replaced by a value given by a function of its previous value and the previous values of the two adjacent cells.[36] So, life is possible in some one-dimensional worlds.

It is still an open question whether second-order linear PDEs in one (and two) dimensions are Turing complete. However, it seems plausible that the rule 110 cellular automaton could be simulated in a continuous one-dimensional system governed by second-order linear PDEs. If so, then life may be possible in one dimension, even given elegant, non-arbitrary laws. It seems plausible that life can evolve naturally here, too; after all, rule 110 exhibits self-replicating structures—"gliders"—given many starting configurations.[37] So, it exhibits stable, interesting structures even in very simple systems—no fine-tuning necessary. In two dimensions, the outlook is even better for life, since all the dynamics possible in one dimension are still possible. If our starting system is translation-invariant in one of the dimensions, then the dynamics will be isomorphic to the dynamics of the one-dimensional system: we can just take one of the identical slices with a fixed value of the coordinate of the dimension with the symmetry. Of course, most systems in the higher dimension will look nothing like this; but this is because they are massively increased, not decreased, in complexity. This additional complexity may render life even more evolvable.

7.5 Four or more dimensions

Tegmark argues that life is not possible in four or more dimensions because the equations of our natural laws, generalized to four dimensions, do not allow for stable orbits or hydrogen atoms. Tegmark broadens this to say that such a world "cannot contain any objects that are stable over time", which seems much more difficult to justify. But, even granting this conclusion for these specific laws, we do not need to conclude that four-dimensional life is impossible. Tegmark's argument relies on the assumption that the natural laws for our universe hold, appropriately generalized to four dimensions. This is a much stronger assumption than Tegmark generally makes; Tegmark has suggested that all mathematical structures—or perhaps all computable mathematical strucutures—may be actual. It makes sense to limit the cases under consideration because it is very difficult to make progress in the most general setting; but this assumption is also much stronger than the assumption Tegmark makes when proving that life is not evolvable in more than one time dimension—there, Tegmark just assumes that the equations are given by second-order linear PDEs. There is no requirement that space be isotropic—indeed, it is completely possible for the coefficients of all of the derivatives in one of the dimensions to be 0, in which case each slice behaves as a space in one fewer dimension, without any interactions between these spaces.[38] Life is clearly possible in four spatial dimensions

[35] We can also work this out more explicitly: terms with $\lambda_i < 0$ quickly vanish; terms with $\lambda_i > 0$ grow to infinity, and thus never stabilize such that a consistent result can be determined, so any y_i with such a term can be ignored; and terms with $\lambda_i = 0$ remain c_i throughout the computation. So, the end result of all such computations is given by the vector of final y_i's with $c_{ij} = 0$ whenever $\lambda_j > 0$: $\sum_j c_{ij} \chi_{(\lambda_j = 0)}$, where χ is the characteristic function.

[36] There are other cellular automata with the same overall structure but different transition functions; "rule 110" refers to the specific function that results in complex behavior.

[37] The proof that the rule 110 cellular automaton is Turing complete builds a Turing machine extensively using these gliders.

[38] This is similar to the case discussed above for displaying the dynamics of $n-1$ dimensions in n dimensions.

if this is the case. However, the inhabitants of such a universe would perceive their world as being three-dimensional, so this does not show that four-dimensional life is possible. (This is completely satisfying if all we want to do is explain our observation that we seem to not exist in a universe with four spatial dimensions. However, if we are wondering about the actual truth about the dimensionality of our universe, it is not good enough. But we should generally believe, by Occam's razor, that the dimensionality we percieve matches the true dimensionality, unless we have a good reason to believe otherwise.) But, even if we do require that space be isotropic, we find that stable objects can exist. The n-dimensional wave equation allows for stable objects; a Gaussian wave packet is a stable object. The n-dimensional wave equation is also hyperbolic, so the environment of a region is predictable.[39] Thus, if we broaden the laws we consider to be possible even slightly—the laws we are considering are still *a priori* plausible—we have little reason to believe that life is not possible in four or more dimensions. Indeed, the additional complexity available may make it even easier for life to evolve and take on complex forms. So, Tegmark has not shown that life is only abundant in worlds with 3 spatial and 1 time dimension, given the assumption that the physical laws are second-order linear PDEs.

However, Tegmark's goal—using anthropic reasoning to explain the observed dimensionality of the universe—may be achievable. It may even be that a careful treatment of epistemology helps us in this goal. For example, if we make the (questionable) decision mentioned above to use the same volume metric regardless of dimensionality, so that lower numbers of spatial dimensions are overwhelmingly privileged compared to greater numbers of spatial dimensions, then we would only need to show that life is (extremely) sparse in fewer than three spatial dimensions—even if life is abundant in more than three spatial dimensions, the density penalty would be enough to make the third dimension overwhelmingly privileged, explaining our presence within it. Or, following the route of using information to define density, we might find that in lower dimensions, the same processes require much more volume to be simulated. (Although I don't know whether this is true at all, is conceivable that there is an exponential blowup in the volume required as the dimensionality decreases from three.) This might explain why our universe has three spatial dimensions rather than fewer.

8 Conclusion

There are ways in which we can make sense of anthropic reasoning in an infinite world. SIA does not seem to have a worse issue with presumptuous philosopher arguments than SSSA; but even if we do require that it not be an implication of SIA that our universe is infinite, there is a (surreal) prior we can hold about the number of observers per universe that is both not utterly unreasonable and does not require us to assign probability 1 to living in an infinite universe. Furthermore, there are ways in which we can extend normal anthropic reasoning to infinite universes. Most of these methods are restricted to infinite worlds with nice properties, like lack of a complicated global structure, bounded observer-density, or even discreteness. Others, like using surreal numbers for probabilities, are promising but difficult to work with. More work is needed, but I believe that there is still much progress to be made. And even this progress can bear fruit in the case of dimensionality: without a careful treatment of epistemology, no conclusions can be drawn, but with this epistemology, we can rigorously arrive at the intuitively correct conclusions. It may even be that a careful treatment of epistemology removes part of the burden on physics to show that certain dimensions cannot support life.

[39] To see that it is hyperbolic, consider the equation for an arbitrary second order linear PDE in n dimensions:

$$\left(\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} + \sum_{i=1}^{n} b_i \frac{\partial}{\partial x_i} + c\right) u = 0.$$

As Tegmark points out, the signs of the eigenvalues of the matrix A for the wave equation match the metric signature. So, for metric signature $(-,+,\ldots,+)$, the signature for $n-1$ spatial dimensions and one time dimension, A will have one negative eigenvalue and $n-1$ positive eigenvalues; and thus, by the definition Tegmark provides, the wave equation is hyperbolic.

9 Appendix: Deriving priors for presumptuous philosopher

Recall that R is the number of observers I am aware of, and O is the total number of observers, with upper bound Ω. Let H_n denote the nth harmonic number. (This is drawn partially from the Wikipedia page on the German Tank Problem, and Mathematica.)

$$\begin{aligned}
P(O = N|R = M) &= \frac{P(R = M|O = N)P(O = N)}{P(R = M)} \\
&= \frac{\chi_{M\leq N} \cdot \frac{1}{N} \cdot \chi_{N\leq\Omega} \cdot \frac{1}{\Omega}}{\sum_{n=1}^{\Omega} P(R = M|O = n)P(O = n)} \\
&= \frac{\chi_{M\leq N} \cdot \frac{1}{N} \cdot \chi_{N\leq\Omega} \cdot \frac{1}{\Omega}}{\sum_{n=1}^{\Omega} \chi_{M\leq n} \cdot \frac{1}{n} \cdot \frac{1}{\Omega}} \\
&= \frac{\chi_{M\leq N} \cdot \frac{1}{N} \cdot \chi_{N\leq\Omega}}{\sum_{n=1}^{\Omega} \chi_{M\leq n} \cdot \frac{1}{n}} \\
&= \frac{\chi_{M\leq N} \cdot \frac{1}{N} \cdot \chi_{N\leq\Omega}}{\sum_{n=1}^{\Omega} \frac{1}{n} - \sum_{n=1}^{M} \frac{1}{n}} \\
&= \frac{\chi_{M\leq N\leq\Omega}}{N\left(H_\Omega - H_M\right)} \\
&\approx \frac{\chi_{M\leq N\leq\Omega}}{N\left(\ln\Omega + \gamma - \ln M + \gamma\right)} \\
&\approx \frac{\chi_{M\leq N\leq\Omega}}{N\left(\ln\Omega - \ln M\right)} \\
&\approx \frac{\chi_{M\leq N\leq\Omega}}{N \ln\frac{\Omega}{M}}
\end{aligned}$$

Note that falling off like $\frac{1}{N}$ is not unreasonable as a prior. (Also note that it's not normalized without Ω!) Below, let ψ denote the digamma function.

$$\begin{aligned}
P(O \geq N|R = M) &= \sum_{n=N}^{\Omega} P(O = n|R = M) \\
&= \sum_{n=N}^{\Omega} \frac{\chi_{M\leq n\leq\Omega}}{n\left(H_\Omega - H_M\right)} \\
&= \frac{1}{H_\Omega - H_M} \sum_{n=\max(M,N)}^{\Omega} \frac{1}{n} \\
&= \frac{H_\Omega - \gamma - \psi(\max(M, N))}{H_\Omega - H_M} \\
&\approx \frac{\ln\Omega + \gamma - \gamma - \ln(\max(M, N))}{\ln\frac{\Omega}{M}} \\
&\approx \frac{\ln\frac{\Omega}{\max(M,N)}}{\ln\frac{\Omega}{M}} \\
&\approx \frac{\ln\Omega - \ln\max(M, N)}{\ln\Omega - \ln M} \\
&\approx \frac{\ln\Omega}{\ln\Omega - \ln M} - \frac{\ln\max(M, N)}{\ln\Omega - \ln M} \\
&\approx \sum_{n=0}^{\infty} \left(\frac{\ln M}{\ln\Omega}\right)^n - \frac{\ln^n M \cdot \ln(\max(M, N))}{\ln^{n+1}\Omega} \\
&\approx 1 + \frac{\ln M}{\ln\Omega} - \frac{\ln\max(M, N)}{\ln\Omega}
\end{aligned}$$

$$\approx 1 + \frac{\ln \frac{M}{\max(M,N)}}{\ln \Omega}$$

References

[1] Bostrom, Nick. *Anthropic bias: Observation selection effects in science and philosophy*. Routledge, 2013.

[2] Tegmark, Max. "On the dimensionality of spacetime." *Classical and Quantum Gravity* 14.4 (1997): L69.

[3] Manley, David. *On being a random sample.* Unpublished Manuscript, 2015.

[4] Arntzenius, Frank, and Dorr, Cian. *What to expect in an Infinite World*. Unpublished Manuscript, 2019.

[5] Chen, Eddy Keming, and Daniel Rubio. "Surreal decisions." *Philosophy and Phenomenological Research*, 2018.

[6] Gonshor, Harry. *An introduction to the theory of surreal numbers*. No. 110. Cambridge University Press, 1986.

[7] Carroll, Sean M. "Why is the universe accelerating?." *AIP Conference Proceedings*. Vol. 743. No. 1. AIP, 2004.

[8] Bournez, Olivier, Daniel S. Graça, and Amaury Pouly. "Polynomial time corresponds to solutions of polynomial ordinary differential equations of polynomial length." *Journal of the ACM (JACM)* 64.6 (2017): 38.

[9] Rosenberg, Gregg. "Causality and the combination problem." *Consciousness in the Physical World: Perspectives on Russellian Monism* (2015): 2016-224.

[10] Lockwood, Michael. "The grain problem." (1993): 271.