

Probabilistic inference in very large universes

Feraz Azhar,^{1,*} Alan H. Guth,^{2,†} and Mohammad Hossein Namjoo^{3,‡}

¹*Department of Philosophy, University of Notre Dame, Notre Dame, Indiana 46556, USA*

²*Department of Physics, Laboratory for Nuclear Science, and Center for Theoretical Physics—A
Leinweber Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

³*School of Astronomy, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*



(Received 16 December 2025; accepted 12 February 2026; published 10 June 2026)

Our current favored cosmological theories allow for the striking and controversial possibility that the observable universe is just a small part of a much larger universe in which parameters that describe the effective, low-energy laws of physics vary from one region to another. The controversy is largely driven by the fact that such a “very large universe” is mostly observationally inaccessible to us, so the issue arises of how we can reasonably assess a theory that describes such a universe. In this paper, we propose a Bayesian method for theory assessment based on theory-generated probability distributions for our observations. We focus on the principles that define this method, leaving aside concerns about how, in practice, one would carry out the required calculations. (One important issue that we set aside is the measure problem.) We argue that cosmological theories can be tested by the standard method of Bayesian updating, but we need to use theoretical predictions for “first-person” probabilities—that is, probabilities that *we* should use for *our* observations, taking into account all relevant selection effects. These selection effects can vary from one observer to another and can vary with time, so, in principle, first-person probabilities are defined for each *observer instant*—an observer at a specific instant of time. Calculations of first-person probabilities should take into account everything that the observer believes about herself and her surroundings, which we refer to as her *subjective state*. If the universe is very large, a theory might predict that there are many observer instants in the same subjective state; we argue that first-person probabilities should be calculated using a *principle of self-locating indifference* (PSLI), the assumption that any real observer should make predictions for her future as if she were chosen randomly and uniformly from the theoretically predicted observer instants that share her subjective state. We believe the PSLI is intuitively very reasonable, but we also argue that, if the theory is correct, the use of this principle maximizes the expected fraction of observers who will make correct predictions. A further complication is that cosmological theories are not expected to fully predict the detailed properties of the universe, but rather will predict a set of possible universes, each with a probability. Different possible universes will generically have different numbers of observers. We argue that, in the calculation of first-person probabilities, the probability for each possible universe should be weighted by the number of observer instants in the specified subjective state that it contains. These issues have been controversial in the literature, so we also provide a rebuttal to the claim that principles like the PSLI involve a “selection fallacy”; a rebuttal to what we dub the *principle of required certainty*; an argument rejecting theories that predict a preponderance of Boltzmann brains; a rebuttal to a parable about humans and Jovians used by Hartle and Srednicki to argue that assumptions of typicality can lead to absurd consequences; and, finally, a discussion about how the use of “old evidence” can be fit into a Bayesian mold.

DOI: [10.1103/n3lg-b1x6](https://doi.org/10.1103/n3lg-b1x6)

I. INTRODUCTION

A striking possibility suggested by the theories that underlie our current understanding of the observable universe—such as cosmic inflation [1–4]—is that the observable universe is just one “domain” in a vast (mostly unobservable) multidomain structure (henceforth, a “very large universe,” also known as a “multiverse”) [5–7]. In such theories, the parameters that appear in the effective, low-energy laws may vary from one domain to the next.

*Contact author: fazhar@nd.edu

†Contact author: guth@ctp.mit.edu

‡Contact author: mh.namjoo@ipm.ir

A central concern that arises in this context is how do we test theories of a very large universe?

There are, of course, many practical problems that arise in trying to test theories of a very large universe, but there are also a number of questions of principle that have remained controversial in the literature. In this paper, we will focus on addressing some of these questions of principle. The first step, in our view, is to specify the idealized methods of evaluating theories of this type—i.e., methods that would be adopted if there were no constraints on our ability to calculate the consequences of the theory. Then the development of practical methods should be based on these idealized methods.

We assume that the theories in which we are interested are sufficiently complete so that they can be used, in principle, to calculate a probability distribution over the possible states of the (very large) universe. A crucial feature of the theories that we have in mind is that they predict probabilities for the state of the universe; they do not directly predict probabilities for what any particular observer will measure. In the context of the theory, observers are simply objects that can arise as the universe evolves. Following Srednicki and Hartle [8], we will refer to these theory-generated probabilities for the state of the universe as “third-person” probabilities. If we were outside the universe, and could see the entire universe, then we would be able to test theories by comparing these third-person probabilities with the universe or universes that we observe. The problem, however, is that we are not outside the universe, but instead are among the objects that are produced within the universe. Furthermore, if the theory predicts (with some probability) the existence of observers who could be us, then it may very well predict the existence of many such observers, and the theory will provide no information about which copy actually corresponds to us. So, to be able to test such a theory, we need to be able to use the third-person probabilities predicted by the theory to infer “first-person” probabilities—probabilities for *our* observations.¹

One problem that arises in defining first- and third-person probabilities is known as the “measure problem.” It involves issues with defining a probability measure over sample spaces in which events of interest occur an infinite number of times. We will not attempt to address the measure problem in this paper. (See Refs. [14–18] for a discussion of this problem.) Thus, our paper will be directly relevant to universes that are very large, but nonetheless have finite spacetime volumes, where there generally are no measure problems. We expect, however, that our claims

will also be applicable for infinite universes once a solution to the measure problem has been adopted.

We adopt a standard Bayesian framework for testing theories, in which the probabilities assigned to theories are updated according to the results of experiments. To calculate the updated probabilities, however, we need to know the probabilities that the theory being tested predicts for the possible outcomes of the experiment. Since these are probabilities for experiments to be performed by us, they are unambiguously first-person probabilities.

To understand the importance of the first-person perspective in predicting the outcomes of physics experiments, we can consider a model of a very large universe in the context of string theory, in which different regions in this very large universe are built upon different types of vacuum. We can imagine that many vacua are consistent with our current data, but each of these vacua might lead to different probabilities for the outcomes of future experiments. So the predictions of any theory for experiments that we will perform in the future (which are first-person probabilities) will depend on what the theory tells us about the probabilities for the different types of vacua in which we might reside. The bottom line is that we cannot test theories unless we have a methodology for understanding the predictions of theories for first-person probabilities, taking into account all relevant selection effects.²

A key problem that arises in determining first-person probabilities is the “typicality problem,” which concerns the issue of whether, given a solution to the measure problem, we should assume that we (and thus our future observations) are *typical* [8,20–24]. By typical, we mean that probabilities for our future observations can be obtained by treating ourselves as if we were randomly and uniformly chosen from some specified set of observers. There are two distinct positions that have been argued for in the literature. The first position, dubbed the “principle of mediocrity” by Vilenkin [25], asserts that we should reason as though we are indeed typical [11,25–28]—this can be understood as a version of the “principle of indifference” [29]. The second position asserts that typicality is just one assumption among a “spectrum” of possible assumptions enforcing various degrees of (a)typicality, with the principle of indifference corresponding to a single such assumption [8,20,22–24,30].

In this paper, we argue for the first position—that we should proceed under the assumption of typicality. Typicality has traditionally been the assumption underlying probabilistic reasoning, and it has so far been very successful. When theories make only probabilistic predictions, there is no method of analysis that can guarantee that correct

¹Third-person and first-person perspectives have a rich history (see, for examples from various disciplines, Refs. [9–13]). The first authors to explicitly highlight these terms in the context of probabilistic reasoning in a very large universe, we believe, were Srednicki and Hartle in Ref. [8]. In earlier papers, starting in, e.g., Hawking and Hertog [12], such third-person and first-person probabilities have also been referred to as “bottom-up” and “top-down” probabilities, respectively.

²According to McMullin [19], “A selection effect is a feature found in what is observed, due to the circumstances of the observation or to the means used, rather than being intrinsic to the object observed.” Note that the “features” mentioned in this definition do not have to be qualitatively new characteristics, but might simply be shifts in statistical distributions.

conclusions will result. However, we will argue that the assumption of typicality will always lead to the largest expectation value for the fraction of observers who reach correct conclusions. We will also argue that, without an assumption of typicality, it would be essentially impossible to exclude *any* theory on the basis of observation.

Once we decide that we are “typical,” we face another important question: What is the “reference class” within which we can assume that we are typical? This is also known as the “conditionalization problem.” That is, probabilities for our future measurements are often assigned under the assumption that we are typical of some reference class of observers, where the probabilities are based on the fractions of the reference class that are predicted to obtain any particular experimental result. One approach to choosing an appropriate reference class is to employ anthropic constraints, namely, conditions that are viewed as being necessary for the existence of any observers (in the style of Carter’s weak anthropic principle [31]), but such constraints are difficult to make precise. Another option is to conditionalize on all previous data we have accumulated, but it remains necessary to define exactly what is meant by “data.” (See Refs. [28,30,32–34] for further discussion.)

In this paper, we will adopt a particular idealized method of conditionalization, based on the goal of avoiding arbitrary choices whenever possible. So, to avoid having to decide if Neanderthals are enough like us to be counted in the same reference class, we advocate that the reference class should be taken as narrowly as possible. (In Sec. IV E, however, we will have further comments about the possible relevance of larger reference classes.) We would expect that such a reference class would lead to the most accurate possible predictions, for it makes use of all the information available to us—thereby precluding the charge of biasing the predictions by ignoring relevant facts, or even possibly cherry-picking the data to find a desired result. This choice means roughly that we should conditionalize our predictions on observers who are *exactly* like us.

But we are not exactly like each other, nor are we even exactly like ourselves at different times. So we are led to the proposition that predictions should, in principle, be made separately for each observer and for each point in time.³

³In Ref. [28], p. 3, Garriga and Vilenkin give a toy example of a situation in which a prediction depends strongly on the observer. A universe is populated by civilizations that are each one of two sizes, large or small, where the number of large and small civilizations are known to be equal. For some reason, nobody has been able to count, or even estimate, the population of her civilization. It is also known that each civilization, large or small, has exactly one astronomer. In this situation, the astronomers would assign a probability of 1/2 for being in a large or small civilization, since half of the astronomers live in each type of civilization. But each ordinary citizen would consider herself more likely to be in a large civilization, since most ordinary citizens live in large civilizations. While these two probabilities differ, each correctly describes the predictions that can be made for the observers for whom the probabilities were calculated.

The reader might at this point be concerned that if we draw the reference class so narrowly, then maybe the only member of the class would be the observer for whom the predictions are sought, and nothing could be learned. However, we must keep in mind that we are discussing theory-generated probabilities, and not probabilities based on the frequency of actual observed events. As long as the theory predicts a nonzero probability for an observer meeting some particular description $\mathcal{D}$, it is, in principle, possible to infer *conditional* probabilities $p(\mathcal{E}|\mathcal{D}) = p(\mathcal{E}, \mathcal{D})/p(\mathcal{D})$ for the events $\mathcal{E}$ that an observer meeting this description will experience. [$p(\mathcal{E}, \mathcal{D})$ denotes the probability of $\mathcal{E}$ and $\mathcal{D}$.]

We use the symbol $\mathcal{O}$ to denote an “observer,” a term that we use to describe any physical entity that has the capacity to record and utilize information (similar to an IGUS—an *information gathering and utilizing system*—discussed by Gell-Mann and Hartle [35]; see also Ref. [8]).⁴ Much of our discussion will revolve around *observer instants* $\mathcal{I}$, where $\mathcal{I}$ denotes the doublet $(\mathcal{O}, t)$, specifying the observer $\mathcal{O}$ at a *particular* time t .⁵ We will use the notation $\mathcal{O}(\mathcal{I})$ to denote the observer associated with observer instant $\mathcal{I}$, and $t(\mathcal{I})$ to denote the time associated with $\mathcal{I}$.

To make predictions for an observer instant $\mathcal{I}$, we want to conditionalize on observer instants exactly like it. But there is a further complication, in that nobody knows, or can know, how to identify observer instants *exactly* like some particular observer instant $\mathcal{I}$. An observer $\mathcal{O}$ can examine her own memories and thought processes, but she cannot determine her precise physical state. She cannot really be sure of her own thought processes, either, since she can examine her thought processes only by relying on them. If she were insane, and maybe even if she were not, she may have a radically deluded image of herself. One could even imagine that the observer $\mathcal{O}$ is actually a disembodied brain kept alive in a vat of nutrients, with all sensory nerves connected to a computer that presents the brain with a totally artificial world.

Since the observer $\mathcal{O}$ cannot know her precise physical state, we use the term *subjective state* to describe what she actually believes about herself and her surroundings. We therefore define

Subjective state. The subjective state of an observer instant $\mathcal{I} \equiv (\mathcal{O}, t)$ is the set of all properties that the observer $\mathcal{O}$ believes, at time t , about herself and her surroundings. These

⁴We are assuming that all properties of observers (or at least all relevant properties) can be described by the laws of physics.

⁵This concept is similar to Bostrom’s “observer-moment” [11]. An observer-moment, however, refers to a brief time segment of an observer, while we use observer instant to refer to a particular time t . While nothing *happens* at an instant of time, we focus on the state of the observer, which is well-defined at an instant t .

beliefs may or may not be true, and include her memories, her self-conception of her thought processes, and anything else that she is aware of about herself or her surroundings. We denote this set of properties by $\mathcal{S}(\mathcal{I})$, or simply $\mathcal{S}$ when $\mathcal{I}$ need not be specified.

We will assume that the best predictions that an observer $\mathcal{O}$ can make will be those that are conditionalized on her subjective state at the current time, i.e., the subjective state $\mathcal{S}$ of her current observer instant $\mathcal{I}$. This point of view agrees with Garriga and Vilenkin (see Sec. II in [28]), who advocated that “ideally” one should use a reference class of “observers with identical information content,” which should be “characterized by the full information content at her disposal.”

We will later discuss the use of an *ensemble* of universes constructed from a theory, which is a list of a very large number (approaching infinity) of fully detailed universe descriptions, generated with the probability distribution implied by the theory. By “fully detailed,” we mean that the description completely specifies not only the cosmological features of the universe, but also the complete histories of everything that occurs in the entire spacetime of the universe. To determine first-person probabilities for a given real observer, it will be necessary to identify observer instants in the ensemble that are in the same *subjective state* as the real observer’s current observer instant. Although we do not believe that such an ensemble of fully detailed universe descriptions can actually be constructed, we believe that thinking about such an ensemble is a good way to describe how probabilities should in principle be extracted from theories. In standard classical statistical mechanics, after all, we discuss ensembles of 10^{23} particles without worrying that the description of one element of this ensemble would, as of 2025, exceed the total storage capacity of all the computers on planet Earth.

We also introduce a concept which we refer to as *LIKEness*, defined in terms of the concept of *subjective state*:

LIKE. Two observer instants $\mathcal{I}$ and $\mathcal{I}'$ are *LIKE* each other if they share the same *subjective state* $\mathcal{S}$. (“*LIKE*,” when used in this way, will always be capitalized.)

Since $\mathcal{S}$ includes everything that each observer believes about herself or her surroundings, if an observer instant $\mathcal{I}'$ is *LIKE* an observer instant $\mathcal{I}$, then the observer $\mathcal{O}(\mathcal{I})$ cannot, using only information available at time $t(\mathcal{I})$, distinguish her current observer instant $\mathcal{I}$ from $\mathcal{I}'$, no matter how much she learns about $\mathcal{I}'$. If $\mathcal{O}(\mathcal{I})$ is a real observer and $\mathcal{I}'$ is an observer instant in the theoretically generated ensemble, then $\mathcal{O}(\mathcal{I})$ might know more about $\mathcal{I}'$ than its subjective state. For example, $\mathcal{O}(\mathcal{I})$ would be able to tell if $\mathcal{I}'$ is a brain in a vat. This additional information, however, cannot affect her judgment about whether $\mathcal{I}'$ is *LIKE* her, since she does not know anything about her own

observer instant beyond her subjective state; i.e., she does not know whether she herself is a brain in a vat.⁶

Finally, we point out that the *LIKEness* of two observer instants, according to our definitions, is an objective property—a third person with full access to the memories, thought processes, and other aspects of the subjective state of two observer instants $\mathcal{I}$ and $\mathcal{I}'$ can in principle decide unambiguously if $\mathcal{I}'$ is *LIKE* $\mathcal{I}$.

In summary, we propose that an observer making predictions for her current observer instant $\mathcal{I}$ should conditionalize on observer instants $\mathcal{I}'$ that are *LIKE* $\mathcal{I}$. Other observers can also make predictions for $\mathcal{I}$, and should also make use of everything they believe about $\mathcal{I}$. But they might know more or less about $\mathcal{I}$ than $\mathcal{O}(\mathcal{I})$ does, so they would in general make different predictions. The methods that we will describe apply to any observer making predictions for any other, but we will focus on predictions made by an observer for herself. Our main goal is to describe how to test theories in very large universes, and an experimenter will do this by comparing her experimental results with the predictions implied by the theory for *her* experiments.⁷

If the universe is sufficiently large, then it would be expected to contain many observer instants *LIKE* any particular $\mathcal{I}$. But even if the universe is small, it is still very reasonable to condition on observer instants *LIKE* $\mathcal{I}$. As discussed above, if we are talking about theory-generated probabilities, then as long as the theory predicts a nonzero probability for such an observer instant, the conditional probabilities for any observation, given that the observer instant is *LIKE* $\mathcal{I}$, are well-defined.

Another important feature of the theories that we have in mind is that the number of observer instants that are *LIKE* any particular $\mathcal{I}$ can vary from one possible outcome (viz., one possible very large universe in the ensemble) to

⁶We understand that some philosophers argue that we *can* tell that we are *not* brains in a vat [36,37]; or—at least—we can tell that the external world, as we perceive it, exists [38]. We, however, will be interested in theories of very large universes that predict the existence of Boltzmann brains. In the context of these theories, it seems clearly rational to take seriously the idea that a conscious being might have no reliable information about the external world.

⁷Our description here has been entirely classical. We are assuming that a classical description of observers can effectively capture their relevant features, though we recognize that (i) a complete physical description would require a specification of their quantum state and (ii) we do not have a complete understanding of how the classical picture emerges from the quantum description. Phrased another way, we accept the experience of our everyday lives, in which the world that we observe, despite its underlying quantum-mechanical nature, is described well by a classical picture. For example, the evolution of very large universes that arise through the nucleation of “bubbles” can mostly be described by classical general relativity, despite the fact that the nucleation event is intrinsically quantum mechanical and must be described probabilistically. On a much smaller scale, in an experiment involving the radioactive decay of an atom, the local “click” of a Geiger counter can be described in terms of classical probabilities, despite this event being intrinsically quantum mechanical.

another. The question arises: How should we weight each such outcome in our computations of the first-person probabilities for the future of an observer instant $\mathcal{I}$? One suggestion asserts that we should weight each outcome in proportion to its (third-person) probability of occurrence, while restricting to outcomes that contain at least one instance of an observer instant that is LIKE $\mathcal{I}$.

As will be discussed more completely in Sec. III B, we argue that predictions for a given observer instant $\mathcal{I}$ should be calculated according to what we dub the *principle of self-locating indifference* (PSLI)—i.e., they should be calculated as if the observer instant were chosen randomly and uniformly from the set of all observer instants in the ensemble that are in the same subjective state.⁸ This principle leads to weighting each model universe in proportion to the number of observer instants LIKE $\mathcal{I}$ that it contains.⁹

⁸This statement has to be worded carefully. Note that our statement of the PSLI is *not* equivalent to

Statement A. Each of the observer instants $\mathcal{I}_i$ in the ensemble that are LIKE $\mathcal{I}$ is equally likely to have a future that matches that of $\mathcal{I}$.

Contrary to Statement A, the PSLI implies that a particular observer instant $\mathcal{I}_i$ is more likely to have a future that matches that of $\mathcal{I}$ if there are many different $\mathcal{I}_i$'s that share the same future. Thus, we contend that Statement A is false. Similarly, the PSLI is *not* equivalent to

Statement B. The observer instant $\mathcal{I}$ is equally likely to have a future that matches that of any of the observer instants $\mathcal{I}_i$ in the ensemble that are LIKE $\mathcal{I}$.

Statement B is equivalent to Statement A, since the statement that $\mathcal{I}_i$ has a future that matches that of $\mathcal{I}$ is equivalent to the statement that $\mathcal{I}$ has a future that matches that of $\mathcal{I}_i$. Thus, Statement B is also false. Our statement of the PSLI is not phrased in terms of the equality of probabilities, but can be summarized by the somewhat vague statement that all of the observer instants $\mathcal{I}_i$ that are in the same subjective state as $\mathcal{I}$ contribute equally to the predictions for $\mathcal{I}$.

⁹Note that this weighting by the number of observers does not commit us to what is known as the “self-indication assumption.” Bostrom (see p. 66 in [11]) defines the self-indication assumption as follows: “Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist.” Such an assumption increases prior probabilities for hypotheses in which there are a larger number of observers and leads to the patent absurdity known as the “presumptuous philosopher” problem (see p. 124 in [11])—the problem that any hypothesis can claim a high prior probability simply by presumptuously appending an assumption that it gives rise to a colossal number of observers. We endorse weighting by the number of observers in the calculation of the first-person probabilities predicted by a given theory, but this does not lead to this absurdity.

The PSLI can be implemented in operational terms by what we call the *method of many repetitions*. In this context, we imagine some proposition pertaining to an observer instant, such as the proposition that the card that was just drawn by the observer $\mathcal{O}(\mathcal{I})$ was a jack of spades, or the proposition that $\mathcal{O}(\mathcal{I})$ is about to be hit by lightning. The method stipulates that the first-person probability for any such proposition is equal to the fraction of observer instants in the same subjective state in the ensemble for which the proposition is true.

The plan for this paper is as follows. In Sec. II, to be explicit about the general probabilistic setting in which we will be working, we describe the standard Bayesian framework for theory assessment. In Sec. III, we develop and motivate the principle of self-locating indifference for computing first-person probabilities for observables, as well as the method of many repetitions. In Sec. IV, we provide further thoughts on probabilistic inference in very large universes, including a rebuttal to the claim that principles like our PSLI involve a “selection fallacy”; a rebuttal to what we dub the *principle of required certainty*; an argument rejecting theories that predict a preponderance of Boltzmann brains; a rebuttal to a parable about humans and Jovians used by Hartle and Srednicki [20] to argue that assumptions of typicality can lead to absurd consequences; and, finally, a discussion about how the use of “old evidence” can be fit into a Bayesian mold. Concluding remarks follow in Sec. V. In Appendixes A and B, we provide further technical details that motivate the principle of self-locating indifference. And, in Appendix C, we provide an explicit comparison of first-person probabilities calculated via (a) an application of the principle of self-locating indifference and (b) methods used by Hartle and collaborators in Refs. [8,13,20,39]—highlighting differences in the results obtained.

To summarize our overall position, we contend that, despite the novelty inherent in theories that describe a very large universe, these novelties can be treated pragmatically by the same Bayesian methods that are commonly used in empirical physics. We assume the principle of self-locating indifference, but we regard this assumption as a natural extension of the typicality assumptions that have always been used in statistical data analysis. This is an optimistic stance, considering how little we know about the possibility of a multiverse.

II. BAYESIAN THEORY ASSESSMENT

We will use a Bayesian framework for theory assessment, which we describe in this section—mainly to establish notation. When Bayes’ theorem is used to update the probabilities that we assign to our theories, we will need to use first-person probabilities, because they are the relevant probabilities for our measurements. First-person probabilities were described briefly in the Introduction, and a precise proposal for calculating them will be described in the following section.

In general terms, Bayes' theorem can be derived from the definition of conditional probability. For two propositions, A and B , provided that $P(B) \neq 0$, the probability of A given B is defined as $P(A|B) \equiv P(A, B)/P(B)$, where $P(A, B)$ is the probability of A and B . Similarly, provided that $P(A) \neq 0$, $P(B|A) \equiv P(B, A)/P(A)$. Since $P(A, B) = P(B, A)$,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1)$$

provided that $P(B)$ and $P(A)$ are both nonzero.¹⁰

One important application of Eq. (1) describes how new data, D_{new} , impacts the probabilities that we assign to theories $\mathcal{T}_i$. If we set, in Eq. (1), $A = \mathcal{T}_i$ and $B = D_{\text{new}}$, we obtain

$$P(\mathcal{T}_i|D_{\text{new}}) = \frac{P(D_{\text{new}}|\mathcal{T}_i)P(\mathcal{T}_i)}{P(D_{\text{new}})}, \quad (2)$$

provided that $P(D_{\text{new}})$ is nonzero.

The quantity on the left-hand side of Eq. (2), namely, $P(\mathcal{T}_i|D_{\text{new}})$, is the probability that we assign to $\mathcal{T}_i$ after D_{new} is obtained, and is known as the *posterior probability* of $\mathcal{T}_i$. The probability that we assign to $\mathcal{T}_i$ prior to our learning D_{new} , $P(\mathcal{T}_i)$, is known as the *prior probability* (or just the “prior”) of $\mathcal{T}_i$. Thus, Bayes' theorem provides a means to update one's confidence in a theory based on new data. The probability $P(D_{\text{new}}|\mathcal{T}_i)$, the theory-generated probability for D_{new} given the theory $\mathcal{T}_i$, is often called the “likelihood.” In general, it may be very difficult to calculate this quantity, and this difficulty can serve as a limitation on our ability to evaluate a theory. In this paper, however, we will consider the idealization in which all such quantities are assumed to be calculable.

Finally, if we also assume that $\mathcal{T}_i$ belongs to an exhaustive and mutually exclusive set of theories $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_N\}$ (see Refs. [40–45]), one can use the law of total probability to write Bayes' theorem as

$$P(\mathcal{T}_i|D_{\text{new}}) = \frac{P(D_{\text{new}}|\mathcal{T}_i)P(\mathcal{T}_i)}{\sum_{j=1}^N P(D_{\text{new}}|\mathcal{T}_j)P(\mathcal{T}_j)}. \quad (3)$$

III. CALCULATING FIRST-PERSON PROBABILITIES IN VERY LARGE UNIVERSES

We begin by assuming that we are exploring the predictions of a fully detailed cosmological theory. An (ideal) complete cosmological theory provides a set of possible universes $\{U_\alpha\}$, where each U_α is a *full*

description of a possible spacetime *and* everything that happens in it. Each U_α describes an entire universe, even if the universe has disconnected components. (Some theories may predict universes with disconnected components, and others may not. We are allowing for both possibilities.) Each U_α is fully detailed, describing not only the cosmological features of the universe, but also the complete histories of every being and every object in the universe, including every experiment that is ever done, every word that is ever uttered, every thought that is ever thought, and every star, planet, meteorite, and speck of dust that ever exists. As described in footnote 7, we are assuming that, for these purposes, a classical description suffices. We assume that U_α describes the universe modulo coordinate redefinitions, so two descriptions that are related by a coordinate redefinition correspond to the same U_α . (In this paper, as we described in the Introduction, we will consider explicitly only the case in which the universe has a finite spacetime volume, although we expect our conclusions to also apply to infinite universes, once a measure is adopted.)

In this language, the physical world that we observe—which we denote by U_{real} —is theorized to be the realization of one of the $\{U_\alpha\}$, meaning that U_{real} is fully described by a particular model universe U_α . The model universe U_α that is realized will be denoted by U_* .

The theory does not tell us which of the U_α will be realized, and we do not have any expectation that we will ever gather enough information to learn which U_α is realized. But the theory in principle allows us to calculate a probability function $P_{\text{cos}}(U_\alpha)$, which describes the probability that the real universe, U_{real} , is described by U_α (equivalently, that $U_\alpha = U_*$). (The subscript “cos” stands for “cosmological.”) The theoretical framework of interest relies on the assumption that it is meaningful to assign probabilities to such unverifiable outcomes, just as classical statistical mechanics relies on the assumption that it is meaningful to assign probabilities to $\mathcal{O}(10^{23})$ phase-space variables, even though we know that we could never measure all of these variables. Thus, while we can never directly test any prediction for the probability of a particular model universe U_α , as specified by $P_{\text{cos}}(U_\alpha)$, knowledge of $P_{\text{cos}}(U_\alpha)$ for all the U_α 's can be used to extract probabilistic predictions for observables that we *will* be able to test.¹¹

It may at first seem inadequate to assign probabilities only to fully described universes U_α while not explicitly assigning probabilities to events within such universes. To understand why only these full-universe probabilities are needed, imagine that the label α can be written explicitly as a huge set of real-valued parameters. Then the specification of the probabilities $P_{\text{cos}}(U_\alpha)$ would amount to a specification of the joint probability distribution for these parameters. The probability distributions for any subset of

¹⁰Note that the probabilities in Bayes' theorem are sometimes explicitly written as being conditional on some background knowledge—comprising information that can affect $P(A|B)$ [40]. For simplicity, we suppress such notation throughout—but assume that probabilities can indeed depend on such background knowledge, provided that all relevant probabilities are conditioned on the same background knowledge.

¹¹For simplicity, we will treat the label α as a discrete variable, but it would be straightforward to generalize our arguments to the case in which the possible universes are labeled by a set of variables, any of which may be discrete or continuous.

parameters can be found by marginalizing the full joint probability distribution over the other parameters.

A. Probabilities, third person and first person

The probabilities $P_{\text{cos}}(U_\alpha)$ describe the probability distribution for the universe from a *third person* point of view—the point of view of an external observer (who is outside of the universe) and can observe the entire universe. However, the theory says nothing whatever about who *I* am, so we have to make further assumptions to make predictions about what *I* (or any other particular observer) should expect to experience. The probabilities that describe what any particular observer should expect to experience are called first-person probabilities.

In more detail, even if we assume that the theory is correct and that we know which U_α describes the real universe U_{real} , it is quite possible in a very large universe that U_{real} will contain multiple observer instants that are LIKE any given observer instant $\mathcal{I}_*$ for which predictions are sought. That is, it is possible that U_{real} contains multiple observer instants $\mathcal{I}_i$ that are in the same subjective state as $\mathcal{I}_*$, $\mathcal{S}(\mathcal{I}_*)$, so that the observer $\mathcal{O}(\mathcal{I}_*)$ (who is *inside* the universe) cannot at time $t(\mathcal{I}_*)$ distinguish any of them from herself. Furthermore, since we do not expect that anyone will actually know which U_α describes the real universe U_{real} , there is the further complication that different possible universes U_α might each have different numbers of observer instants $\mathcal{I}_i$ which are in the subjective state $\mathcal{S}(\mathcal{I}_*)$. In this situation, the following question arises:

How should the observer $\mathcal{O}(\mathcal{I}_*)$ compute probabilities for what she will experience?

That is, how should one calculate the *first-person* probabilities for what the observer $\mathcal{O}(\mathcal{I}_*)$ will experience?

B. Principle of self-locating indifference

To describe the answer that we propose to this question, we first define some terms that will be useful. Recall that we are attempting to describe the consequences of a fully detailed cosmological theory, which provides a set of possible physical states (i.e., the set of possible universes $\{U_\alpha\}$) and a probability distribution $P_{\text{cos}}(U_\alpha)$ over these physical states.

Following discussions that are common in statistical mechanics textbooks, we will imagine using the theory to generate a large list of model universes $U_1^{\text{ens}}, U_2^{\text{ens}}, \dots, U_N^{\text{ens}}$ —called an *ensemble*. Each entry U_λ^{ens} on the list is one of the possible universes $\{U_\alpha\}$, chosen according to the probability distribution $P_{\text{cos}}(U_\alpha)$.¹²

¹²Gibbs [46] originally defined the notion of an ensemble in exactly this way, as a “great number of independent systems,” each in a definite state described by a point in phase space. The density of these points in phase space was used instead of any notion of probability.

Note that the ensemble is not a set, since a given model universe U_α can appear in the ensemble multiple times, while a given element can appear in a set only once. Note also that the ensemble is not uniquely determined by the theory, since each entry is chosen randomly. We will be interested, however, exclusively in the statistical properties of the ensemble in the limit $N \rightarrow \infty$, which *are* uniquely defined by the theory. To be clear, we coin the term *asymptotic ensemble* to refer to an ensemble that is described as a list of N entries, but for which the only properties that one is allowed to extract are those that arise from the statistical distribution in the limit $N \rightarrow \infty$. Thus, each theory determines a unique asymptotic ensemble.

The ensemble is not essential to our methods, but it simplifies the discussion by encoding the probability distribution $P_{\text{cos}}(U_\alpha)$ into the multiplicities of universes U_α in the ensemble. Thus, choosing an entry randomly and uniformly from the asymptotic ensemble is equivalent to choosing a model universe U_α according to the probability distribution $P_{\text{cos}}(U_\alpha)$.

If we know the asymptotic ensemble implied by a theory, and if somebody asks us the probability that U_α is the universe that is realized, we can immediately answer $P_{\text{cos}}(U_\alpha)$, without any further thought. But to compare the predictions of the theory with observations, we need to be able to answer a much wider class of questions. Since each U_α is a complete description of the universe, we should be able to answer questions about probabilities for anything that can happen within a universe.

We are interested in constructing a method to calculate probabilities for arbitrary questions that an observer might ask about her future, or even her past, or for events in regions that are spacelike separated from her current observer instant. In the classical approximation to reality that we are describing, the entire spacetime around the observer has a well-defined description, whether the observer has direct access to these regions of spacetime or not.

Our goal is to define a method to calculate the probability for any proposition of interest pertaining to an observer at a particular instant in time. We want to be able to calculate, for example, the probability that a specific event happens in the future or the past of some observer instant, or perhaps the probability that a specific set of such events happen in the future or the past of some observer instant. It could also be, as another example, the probability that a distant vacuum decay event occurs (one that was initially spacelike separated from the observer), which sets off an expanding bubble wall that will collide with the observer 5 years into the future. The probabilities of interest can be time specific (e.g., the probability that the observer will have a car accident 22 days in the future), or not time specific (e.g., the probability that the observer will have a car accident sometime in her future).

We can now state the principle that we advocate to calculate such first-person probabilities. We will first state the principle, discussing motivation and justification below.

Principle of self-locating indifference (PSLI). An observer should calculate first-person probabilities for any proposition pertaining to her current observer instant as if the observer instant were chosen randomly and uniformly from the set of all observer instants in the asymptotic ensemble (generated from the theory) that are in the same subjective state.

We can illustrate this principle via a simple example. Consider a scenario in which our fully detailed cosmological theory allows two possible universes U_1 and U_2 , with probabilities $P_{\text{cos}}(U_1) = 1/3$ and $P_{\text{cos}}(U_2) = 2/3$. Assume that we are interested in predictions for an observer instant whose current subjective state is $\mathcal{S}$. Let there be one observer instant in U_1 and two observer instants in U_2 that are in the subjective state $\mathcal{S}$. Suppose that the (lone) observer of interest in U_1 suffers a myocardial infarction within the next 24 hours; while one of the observers of interest in U_2 suffers a myocardial infarction in the next 24 hours. We may then ask:

If an observer's current subjective state is $\mathcal{S}$, what is the probability that the observer will have a myocardial infarction within the next 24 hours?

To answer this question, we construct the theory-generated ensemble—a list of a large number N of universe descriptions, as discussed above. If the list is sufficiently large, we expect U_1 to appear on the list about $N/3$ times and U_2 to appear about $2N/3$ times.¹³ There will be about $N/3$ observer instants in subjective state $\mathcal{S}$ in universes of type U_1 and about $4N/3$ such observer instants in universes of type U_2 . Since about $3/5$ of these $5N/3$ observer instants suffer a myocardial infarction within the next 24 hours, an observer instant chosen randomly and uniformly from this set will have a probability of 60% of having a myocardial infarction within the next 24 hours. Thus, according to the PSLI, the real observer of interest should predict that she has a 60% probability of having a myocardial infarction within 24 hours.

This example may be generalized to a statement which we dub the method of many repetitions. The method of many repetitions is logically equivalent to the PSLI, but its wording more directly describes how first-person probabilities are to be operationally calculated.

¹³Our style here follows Gibbs' example, discussing an ensemble with a finite number N of elements, while relying on statements that are only true in the limit $N \rightarrow \infty$. In a footnote on p. 5 of Ref. [46], Gibbs wrote, "In strictness, a finite number of systems cannot be distributed continuously in phase [phase space]. But by increasing indefinitely the number of systems, we may approximate to a continuous law of distribution, such as is here described." Gibbs goes on to say that such language, "although wanting in precision of expression," will be used "to avoid tedious circumlocution ... when the sense in which it is to be taken appears sufficiently clear."

Method of many repetitions (MMR). The first-person probability for any proposition pertaining to a particular observer instant is the fraction of the observer instants in the same subjective state, in the theory-generated asymptotic ensemble, that belong to observers for whom the proposition holds.

The MMR is an operational procedure to calculate the consequences of the PSLI. In particular, it can be used to derive an explicit equation for the first person probability $p^{(1p)}(\mathcal{P}|\mathcal{S})$ that any proposition $\mathcal{P}$ will hold for an observer-instant in subjective state $\mathcal{S}$. Once the theory of interest has been used to calculate the set of possible universes $\{U_\alpha\}$ and their probabilities $P_{\text{cos}}(U_\alpha)$, we can let $n_{\text{tot}}(U_\alpha, \mathcal{S})$ denote the total number of observer instants in subjective state $\mathcal{S}$ in the model universe U_α , and we can let $n_{\text{prop}}(U_\alpha, \mathcal{S}, \mathcal{P})$ denote the number of observer instants in subjective state $\mathcal{S}$ in the model universe U_α for which the proposition $\mathcal{P}$ is true. If N is the total number of model universes in the ensemble, then the expected total number of observer instants in subjective state $\mathcal{S}$ in the ensemble is given by

$$\mathcal{N}_{\text{tot}}(\mathcal{S}) = N \sum_{\beta} P_{\text{cos}}(U_\beta) n_{\text{tot}}(U_\beta, \mathcal{S}). \quad (4)$$

Of these observer instants, the expected number for which the proposition $\mathcal{P}$ is true is given by

$$\mathcal{N}_{\text{prop}}(\mathcal{S}, \mathcal{P}) = N \sum_{\beta} P_{\text{cos}}(U_\beta) n_{\text{prop}}(U_\beta, \mathcal{S}, \mathcal{P}). \quad (5)$$

The MMR tells us that the first-person probability that proposition $\mathcal{P}$ is true for an observer instant in subjective state $\mathcal{S}$ is given by the ratio of these two quantities:

$$p^{(1p)}(\mathcal{P}|\mathcal{S}) = \frac{\mathcal{N}_{\text{prop}}(\mathcal{S}, \mathcal{P})}{\mathcal{N}_{\text{tot}}(\mathcal{S})} \quad (6)$$

$$= \frac{\sum_{\beta} P_{\text{cos}}(U_\beta) n_{\text{prop}}(U_\beta, \mathcal{S}, \mathcal{P})}{\sum_{\beta} P_{\text{cos}}(U_\beta) n_{\text{tot}}(U_\beta, \mathcal{S})}. \quad (7)$$

This prescription for calculating first-person probabilities based on the PSLI is one of the main results of this paper. Since it fully describes how to calculate the first-person probability for an arbitrary proposition $\mathcal{P}$, it is fully equivalent to the definition of the PSLI (written out in words as above).

Finally, the above formalism allows us to address the question of the first-person probability that $\mathcal{O}(\mathcal{I})$, for a given observer instant $\mathcal{I}$ in subjective state $\mathcal{S}$, lives in a universe described by a specific U_α . If $\mathcal{P}_\alpha$ is the proposition that the specified observer instant $\mathcal{I}$ lives in a universe described by U_α , then $n_{\text{prop}}(U_\beta, \mathcal{S}, \mathcal{P}_\alpha) = n_{\text{tot}}(U_\alpha, \mathcal{S}) \delta_{\alpha\beta}$, so Eq. (7) implies that

$$p^{(1p)}(U_\alpha|\mathcal{S}) = \frac{P_{\text{cos}}(U_\alpha) n_{\text{tot}}(U_\alpha, \mathcal{S})}{\sum_{\beta} P_{\text{cos}}(U_\beta) n_{\text{tot}}(U_\beta, \mathcal{S})}. \quad (8)$$

Importantly, Eq. (8) states that the probability that I live in a particular universe U_α is not only proportional to the probability $P_{\text{cos}}(U_\alpha)$ that the universe is realized (according to the theory), but it is also proportional to the number $n_{\text{tot}}(U_\alpha, S)$ of observer instants LIKE mine in that universe. The latter factor—that is, the weight associated with the number of observer instants—is an “anthropic” effect, resulting from the PSLI.

To justify the procedure that we are proposing to define first-person probabilities, in the following section we will discuss the underlying motivation for principles of indifference. For readers inclined to accept indifference as a reasonable assumption, these arguments might be sufficient. However, in Appendixes A and B, we attempt to give more detailed arguments based on more primitive assumptions.

In Appendix A, we make the simplifying assumption that we know which model universe $U_* \in \{U_\alpha\}$ is actually realized. Even if U_* is known, however, the calculation of first-person probabilities for some real observer instant $\mathcal{I}$ is complicated by the possibility that the model universe U_* might contain any number of observer instants that are LIKE $\mathcal{I}$, and are therefore indistinguishable from $\mathcal{I}$ [to the observer $\mathcal{O}(\mathcal{I})$ at the time $t(\mathcal{I})$]. Thus, we need to specify the probability that $\mathcal{I}$ might be described by any particular one of them. Since the goal is to define a procedure for extracting first-person predictions from theories in very large universes, we will use the success of such predictions as our criterion for comparing different possible procedures. In particular, in Appendix A, we will show that the use of the PSLI to define first-person probabilities maximizes the expected fraction of observers who will deduce correct inferences from their observations.

In Appendix B, we will drop the simplifying assumption that U_* is known, and we will give an argument, based on what we consider very reasonable assumptions, that the first-person probability for any particular model universe U_α should be weighted by the number of observer instants in the model universe that are LIKE $\mathcal{I}$. These arguments in Appendix B will allow us to rederive Eq. (7), a key result of this paper which is equivalent to the PSLI.

C. Motivating indifference

Our key principle, the PSLI, mandates that observers ought to calculate probabilities for outcomes of interest as if their current observer instant were chosen randomly and uniformly from those observer instants in the same subjective state in the asymptotic ensemble generated from the theory. This method of making predictions for a real observer implements the assumption that her observer instant is typical of the observer instants in the ensemble that are in the same subjective state. So we must address the question: When is it appropriate to assume that predictions should be made as if the case of interest is typical of some reference class? Some have argued that we should consider ourselves typical of any reference class for which there is no evidence to

the contrary. Others have argued that there is no evidence that we are typical of any group, and that we should therefore allow for the possibility that we are simply not typical.

The question of our typicality is an example of the more general question of deciding under what circumstances the probabilities of alternative outcomes should be taken to be equal. The attempt to answer this more general question has led to a variety of points of view, including various formulations of the *principle of indifference*. The principle of indifference has its origins in work by Bernoulli [47] and Laplace [48]. The term “principle of indifference” was coined by Keynes [29], who also proposed a more precise description of the qualifications that are needed before a set of outcomes can be assigned equal probabilities. For further discussion, see Ref. [49].

For the present discussion, we need only a principle of indifference that is adapted to the self-location problem discussed in the previous section. Here we adopt the PSLI.

Note that principles of indifference adapted to the self-location problem in cosmology are apparently not (yet) universally accepted among cosmologists. They have been challenged in at least two ways.

First, a possible position is that if an observer does not know which observer she is (in some appropriate reference class), then she should just accept the fact that she does not know the probabilities. We agree that it is always possible to adopt the position that we do not know the probabilities. Similarly, if somebody asks us what is the capital of Kazakhstan, we can honestly reply that we do not know. “We do not know” is a possible answer to most any question, but it does not exclude the possibility of other answers that are also correct and possibly more informative. While the observer may not know which of the observers she corresponds to, the PSLI is a way of quantifying her uncertainty, based on the available information. We can sometimes be forced to make such judgments. If a tornado is approaching, we may have to decide in which direction to run, which would presumably involve some estimate of the probabilities of different paths that the tornado might follow. We would make the best estimate we could, based on available information, and there would be no value in maintaining the position that we do not know the probabilities.

Second, principles of indifference adapted to the self-location problem have been challenged, most emphatically in Ref. [8], wherein it is claimed that typicality is “no more motivated by observation than atypicality.”

We agree that there are no observational tests of our typicality in the universe, but we believe that there is a solid argument for typicality, based on counting.

The counting argument that we have in mind can be illustrated in a standard textbook probability problem involving coin flipping. If a million people each flip a fair coin 20 times, there are better than even odds that at least one will flip 20 heads in a row. But neither probability theory, nor anything else that we know about the laws of physics, tells us anything about which of the million people is likely to

be the one who flips 20 heads in a row. In evaluating the implications of the probability distribution, however, we always assume that *we* are very unlikely to be the one-in-a-million coin flipper who flips 20 heads in a row. More generally, if we consider 1.5 million coin flippers, probability theory tells us that most likely there will be at least one flipper who flips exactly n heads, for every n from 0 to 20. If we did not make the assumption that we are typical, then all we would know is that any n is possible, and we would lose all the probabilistic predictions of coin flipping.

The counting argument, in the context of betting, was presented very clearly by Vilenkin in Ref. [50]. Vilenkin imagines a conference in which all attendees have been given hats, 80% white and 20% black, but nobody has been able to see the color of his or her own hat. To register for the conference, everyone is required to bet £100 on the color of her hat. Everyone can look around and notice that the hats are about 80% white. Thus, if the attendees all assume that they are typical and therefore bet on white, 80% will win their bets. By contrast, if the attendees all assume that they do not know the probabilities, they might place a random bet and on average only half would win. Clearly, the assumption of typicality is the better betting strategy. While the 80% vs 50% in Vilenkin's example may not sound compelling, the advantage of assuming typicality can be made more dramatic by considering repetitions of this scenario. If the same group attends 20 conferences with the same protocol and assume that they are typical, 99.7% will have net winnings. By contrast, if they bet randomly, again on average only half would win.

In Appendix A, as we summarized above, we will show explicitly that Vilenkin's white hat/black hat parable generalizes to the case under discussion here—the evaluation of theories. We will show that analyzing experimental data using the PSLI (or, equivalently, the method of many repetitions) maximizes the expectation value of the fraction of experimenters who will reach correct conclusions from their data. This is our take on the issue of typicality.

Note that as emphasized by Garriga and Vilenkin (see p. 3 in [28]), such assumptions of typicality are routine in standard laboratory settings. Say, for example, an experimenter measures the value of some physical quantity, obtaining the result m . Suppose further that we are interested in comparing two theories, T_1 and T_2 , in light of this experimental result. The standard Bayesian approach is to compare theory-generated probabilities for m (viz., likelihoods), given a description of the experiment, where this description is denoted by E . That is, one compares $P(m|T_1, E)$ with $P(m|T_2, E)$. If $P(m|T_2, E) \ll P(m|T_1, E)$ (that is, if the likelihoods strongly favor T_1), then the standard conclusion would be that T_2 is strongly disfavored. We would not expect it to be claimed that T_2 is perfectly acceptable, since we can freely assume that we are among the extremely atypical observers who obtain the result m , despite its low probability. In other words, in standard laboratory settings, assumptions of atypicality are never introduced to avoid disfavoring a theory with a very small likelihood.

In Ref. [8], Srednicki and Hartle propose that it is not appropriate to imagine that we can test theories directly, but rather we should always think of testing *frameworks*, where a framework consists of a pair (T, ξ) , where T is a theory and ξ is a *xerographic distribution*. (They assume that the full description of the actual universe is known, to focus on the question of our place in the universe.) If our data D_0 are replicated at spacetime locations $\{x_1, x_2, \dots, x_N\}$, then the xerographic distribution is an assumed (i.e., not derived from a cosmological theory) set of probabilities $\{\xi_1, \xi_2, \dots, \xi_N\}$ that we are located at each of the spacetime locations $\{x_1, x_2, \dots, x_N\}$, respectively. We would argue that if one allows xerographic distributions in which we are highly atypical, then empirical science as we know it would become essentially impossible. To see this, consider a theory T for which $P(D_0|T)$, the probability of our data D_0 given the theory T , is extremely small. If $P(D_0|T)$ is nonetheless large enough so that D_0 is likely to occur at least once in the very large universe, we can always arrange for $P(D_0|T, \xi)$ to be nearly one, by choosing a xerographic distribution ξ which implies that we are among the very rare observers who observe D_0 . (This is analogous to assuming that we are likely to be the 20-head coin flipper mentioned earlier in this section.) In a very large (or possibly infinite) universe, a huge range of theories are likely to be found acceptable by this criterion.

While some may conclude that we are forced to accept this huge range of theories, we do not see any justification for this conclusion. Recall (i) the expectation value of the number of observers who make correct deductions is maximized by using the assumption of typicality, which then allows us to drastically narrow the range of acceptable theories; (ii) such assumptions of typicality are routine in standard laboratory settings.

IV. OTHER THOUGHTS ON PROBABILITIES IN VERY LARGE UNIVERSES

In this section, we discuss five issues related to the general topic of probabilistic reasoning in the context of very large universes.

A. The *selection fallacy* fallacy

In Ref. [20], Hartle and Srednicki wrote:

“To compute likelihoods as though we had been randomly selected by some physical process, when there is no evidence for such a process, commits what might be called the *selection fallacy*. We are not a disembodied entity that was randomly selected to have a particular physical description; instead, we are the meaning of a transcription of ‘we’ into the language of physical theory.”

Here, we argue that this claim of a fallacy is, in fact, false. To assume that we are typical of some reference class does not require us to assume that we are a disembodied entity that was randomly selected by some process to match some particular member of that reference class. All that is necessary, we argue, is to know that we are a member of the reference class, and to have no reason to believe that we are any more likely to be one particular member rather than another.

To make our point, let us consider a thought experiment involving the hypothetical cloning of humans. Suppose a subject is put to sleep, and while asleep five clones of the subject are made, and for simplicity (recognizing that this is only a thought experiment), we assume that the original subject is killed. We then have five clones, all identical, with nothing to distinguish them other than their *creation number*, from 1 to 5, which indicates the order of their creation. They are placed into identical rooms, numbered 1–5, according to their creation number. They are all awakened, but are not told their creation number and cannot see the numbers on their rooms.

In addition, each clone is also assigned a randomized number, also from 1 to 5. There are 120 permutations of the numbers from 1 to 5, so the randomized numbers are assigned by flipping a fair 120-sided coin, with each outcome linked to a permutation. The permutation is then used to assign the randomized numbers. For example, if the randomly chosen permutation is (4,2,3,5,1), then the clones in rooms 1–5 will be assigned the randomized numbers 4, 2, 3, 5, and 1, respectively.

Now we consider the probabilities that each clone should assign to their creation numbers and their randomized numbers. Hartle and Srednicki's *selection fallacy* argument would seem to imply that the creation numbers cannot be treated as random, since there was no random process involved in determining them. (They are closely analogous to the observer-instant numbers that we use in Appendix B.) They are simply an indication of the room in which the clone is sitting, so the creation number is purely a question of self-location.

On the other hand, the randomized numbers were indeed chosen randomly, so surely probability theory should apply here. Each clone should regard his randomized number as having a probability of 1/5 to be any number from 1 to 5.

We argue, however, that there is no relevant difference between these two cases (i.e., the two number assignments). If someone offers all the clones an opportunity to bet on their creation number or randomized number with better than 5:1 odds, it would pay on average for them to take the bet. If they all bet on "5" for either their creation number or randomized number, exactly one will win and four will lose, but the winnings of the winner would exceed the total of the losses. Most importantly, there would be no difference between the two cases.

As is usual in probabilistic predictions, the predictions can be improved significantly with repetition. The cloning

experiment can be repeated by making five clones of each of the clones, and then making five clones of each of those clones, etc. After each cloning, the original subject is destroyed, but each of the five clones considers himself to be the continuation of the original subject—each clone acquires all the memories of the original subject, and interprets these memories as memories of his own past.

As a concrete example, let us suppose that the experiment is repeated 100 times. At the end, there will be 5^{100} clones, which are a lot of mouths to feed, but since this is a thought experiment we will not worry about this hypothetical hunger emergency. At each repetition, each of the clones is offered the opportunity to place separate bets on his creation number and his randomized number. If he guesses wrong, he loses \$1.00; if he guesses right, he wins \$10.00, so the odds are in his favor. At these odds, clones who guess right at least 10 of the 100 times will come out ahead. Suppose that on each repetition they all bet on the same choice as each other, on both the creation number and randomized number. Since 1 in 5 will guess right on each trial, the fraction of clones who guess right exactly n times out of N trials is given by the binomial formula as

$$F(n) = \frac{N!}{n!(N-n)!} \left(\frac{1}{5}\right)^n \left(\frac{4}{5}\right)^{N-n}. \quad (9)$$

Using this formula to sum the fractions from $n = 10$ to $n = 100$, with $N = 100$, we find that under these assumptions 99.77% of the clones will come out ahead on their bets, for either the creation number bets or the randomized number bets. Again, there is no difference between the two cases. If we discard the assumption that all clones bet on the same choice as each other in each repetition, and instead assume that they bet randomly, then we conclude that the expectation value for the fraction of clones that come out ahead is 99.77%. Again, all predictions are the same for the two cases.

At this point, the selection fallacy advocates might say that of course there is no difference between self-location and random choice if one only looks at global averages, but if I were part of this experiment, only one of these clones would be *me*. For the randomized numbers, they would say, probability theory convinces me that I am very unlikely to be part of the 0.23% who lose on their bets, while for the creation numbers I do not know this. But if pressed to describe the physical difference between these two cases, we are not aware of any sensible answer that the selection fallacy advocate might give.

In fact, we will argue that the clones would have absolutely no way to distinguish between their creation numbers and their randomized numbers. To be concrete, let us assume that, after each bet is concluded, each clone is told the correct value of his creation number and randomized number. At the end of the experiment, there are then 5^{100} clones, each of which has recorded a sequence of 100 creation numbers, each from 1 to 5, and a sequence of 100

randomized numbers, also from 1 to 5. Suppose we imagine a variant of the experiment in which the clones are told that they will each be assigned a creation number and a randomized number, as in the original version, but in this case the numbers will be presented as X and Y. The clones are told only that the symbols will be used consistently—X will always be the creation number and Y the randomized number, or maybe the other way around. In this case, we see no way that the clones could determine which sequence was which—even if they had the opportunity to pool all of the data from the 5^{100} clones. We believe that, in either case, we can rely on the belief that if only 0.23% of the group will lose money, and I have no way of knowing where in the group I will fall, then I can assume that it is very unlikely that I will lose money. Thus, for either sequence, it pays for the clones to place their bets.

We would further argue that the physical description of the system after 100 clonings allows no place for *me* to be special. Each possible creation number sequence occurs once and only once, and the same can be said for the randomized number sequences. They were assigned by different methods, but their description at the end of the experiments seems to be the same in all relevant aspects.

The physics of the system knows nothing about who *I* am. Rather, the physics of the system includes 5^{100} different *I*'s, where each "*I*" refers to the thoughts happening in one of the 5^{100} heads of the clones. The physical situation associates one of the 5^{100} *I*'s with each possible creation number sequence, and similarly one of the 5^{100} *I*'s with each possible randomized number sequence. Physics, as far as we know it, is always formulated in terms of a description of possible physical states and a prescription for how such states evolve with time. Within that framework, we cannot see any way that any one of the clones can be treated as special, not subject to the same probability distribution as the rest.

Finally, we point out that in practice there is no doubt that the world accepts the use of probabilistic reasoning in situations where there is no "random process" giving rise to the uncertainty. In Sec. IV D, especially in footnote 16, we will emphasize the medical application. Here, we highlight the use of Monte Carlo calculations, which have become ubiquitous in physics and in many other fields. The "random numbers" used in these calculations are actually pseudorandom—they are generated by a purely deterministic algorithm, but the algorithm is chosen to be sufficiently complex so that there is no practical way to predict the next bit that will be generated with better than even odds. Thus, although there is no random process, there is also no practical way to conclude that the next bit is more likely to be a 0 or 1, even after observing a large number of output digits and applying a suite of pattern-searching algorithms. In our view this is all that is needed, and all that we often have, to justify the use of probabilistic reasoning based on indifference.

B. When we evaluate theories, should we use only what we know for certain?

When we perform a measurement obtaining data D_0 , the only fact about the universe that we know *for certain* (assuming that we can trust our analysis) is that there exists at least one instance of D_0 in the universe. In Ref. [20], Hartle and Srednicki argued that, in evaluating theories, we should use only this certain information. As described in Sec. III, we (strongly) disagree with this assertion, but for purposes of discussion we give it a name—the principle of required certainty.

Principle of required certainty (PRC). When we perform a measurement obtaining some data D_0 , only the fact that we know for certain—the fact that at least one instance of D_0 exists in the universe—should enter into our evaluation of theories of the universe.

By contrast, a more standard Bayesian analysis would take into account the probabilistic inferences stemming from the fact that we have done one experiment of this type, and found the result D_0 .

In the authors' own words (Ref. [20], p. 2):

All we know is that there exists at least one such region [region with exactly our data] containing our data.

In evaluating the implications of a measurement, these authors base all of their reasoning *solely* on this point of certainty.

After stating that the existence of at least one instance of D_0 is all we know for certain, the authors of Ref. [20] proceed to reason on the basis of this fact as if it were standard in science to rely only on what we know for certain. In our view, however, this is an extremely inaccurate description of how science proceeds. Scientific reasoning has never been limited to what we know for certain, and indeed science as we know it would come to a halt if such a restriction were adopted.

If we look at the experimental literature, we often see 1σ error bars with their approximately 68% confidence level, and sometimes 2σ error bars with their approximately 95% confidence level. But we have *never* seen error bars of certainty. Indeed, if experimenters were required to use error bars of certainty, experimental papers would essentially disappear. Similarly, we have seen papers that exclude certain hypotheses at the 95% confidence level, and sometimes the 99% confidence level. But we have *never* seen a scientific paper that discusses the exclusion of a hypothesis at the 100% confidence level.

To see how the requirement of certainty is at odds with conventional Bayesian reasoning, we can consider the following textbook problem, from the textbook by Skorokhod [51]:

There are two urns of which the first contains two white and eight black balls and the second eight white and two black balls. An urn is selected at random and a ball is drawn from it. It is white. What is the probability that the first urn was chosen?

The standard answer, as given in the textbook, is that our selection of a white ball favors the second urn, in which the fraction of white balls is larger, so the probability of the first urn is found to be only $1/5$. On the other hand, if we approached this question by first asking what we know *for certain* about the urns, the answer would be that the draw of a white ball tells us only that the urn contains at least one white ball. Since we already knew that both urns contain at least one white ball, we have learned nothing. The PRC would therefore imply that the probability that we chose the first urn would be unchanged from its prior, which was taken to be $1/2$. Thus, the PRC is in contradiction with conventional Bayesian reasoning. (Note that we are not saying that Hartle and Srednicki [20] have incorrectly applied Bayes' theorem. Rather, we are saying that they departed from conventional Bayesian reasoning by using only the fact that the universe contains at least one instance of our data.)

To see how the PRC affects the analysis of an experiment, consider an effort to measure the decay rate of a particle X . Suppose that we put N of them in a bottle, and in time Δt we observe that k of them decay. We assume the conventional model of particle decays: each particle that has not already decayed is assumed to have a probability λdt of decaying in any time interval dt , independent of the environment or prior history. Provided that $k \ll N$, we would estimate the decay rate to be $\lambda = k/(N\Delta t)$, and an estimate of the 1σ uncertainty would be $\Delta\lambda \approx \sqrt{k}/(N\Delta t)$. But suppose, instead, we asked what do we learn *for certain* about the decay rate? Is it possible, for example, that the actual decay rate is smaller by a factor of a billion, or larger by a factor of a billion? Yes, although the result that we found would have been very unlikely in that case. But it is not impossible. So the only thing that we learn *for certain* is that the decay rate is nonzero. This situation is not changed if we imagine repeating the experiment many (but only a finite number) of times. If we pool all the data from all the repetitions, we can ask the same questions, and they will have the same answers.

In the standard Bayesian analysis, by contrast, we never learn for certain what the decay rate is, but if the experiment involves a high number of counts, it will strongly favor the theory in which the decay rate is near $k/(N\Delta t)$. Physics as we know it depends crucially on such probabilistic inferences. If only certainties were allowed in our scientific reasoning, we would know essentially nothing.

One important consequence of the PRC is to dramatically weaken our ability to use empirical data to compare theories. As Hartle and Srednicki [20] put it,

Cosmological models that predict that at least one instance of our data exists (with probability one)

somewhere in spacetime are indistinguishable no matter how many other exact copies of these data exist.

To illustrate their methods, Hartle and Srednicki introduce a toy model in which the universe consists of N cycles in time, each of which is either “red” or “blue,” and each of which contains an “observing system” with probability p_E . They consider two alternative theories: an “all red” (AR) theory in which all cycles are red, and a “some red” (SR) theory in which $N_R < N$ cycles are red and the rest are blue. Among other limits, they consider the limit in which N and N_R approach infinity, writing:

If the number of red cycles is infinite in both models, then the two theories are not distinguished by our data even if the number of [*sic*] fraction of red cycles is small in SR. That is because our data are that there is at least one red cycle. As the number of red cycles approaches infinity, that becomes a certainty in both theories. Even though the typical observing system in the SR theory is observing blue, our data provide no evidence that we are typical.

Thus, we conclude that acceptance of the PRC would make science as we know it impossible, at least under the assumption that the universe might be extremely large or infinite.

Since the PRC seems to be a radical and unmotivated departure from scientific practice, and since—in the case of a very large universe—it can render us almost completely powerless to discover observational reasons to prefer one theory over another, we see absolutely no reason to adopt it.

C. Rejection of theories in which Boltzmann brains dominate

Some of the cosmological theories currently under discussion lead to very large spacetime regions of exact or approximate de Sitter space, which leads to a phenomenon known as “Boltzmann brains.” The issue arises from the fact that quantum fluctuations in the ground state of de Sitter space mimic thermal fluctuations at the Gibbons-Hawking temperature $T_{\text{GH}} = H/(2\pi)$, where H is the Hubble expansion rate of the de Sitter phase [52]. In principle, any kind of thermal fluctuation is possible, with a probability proportional to $e^{-E/T_{\text{GH}}}$, where E is the energy of the fluctuation (and the Boltzmann constant $k_B \equiv 1$). With some nonzero probability, such thermal fluctuations can give rise to “Boltzmann brains” (BBs), which are objects like our own brains that are capable of thinking. Unlike what we believe for our own brains, BBs are formed as entirely random thermal or quantum fluctuations, so any memories that they possess are merely reflections of the random configuration of atoms. Nonetheless, with some very small but nonzero probability, a BB might have exactly the same

memories and current thoughts as I have at this moment. More generally, there is some nonzero probability that a BB might be in the same subjective state as any observer instant $\mathcal{T}$ allowed by the cosmological theory. A BB might be surrounded by a whole Boltzmann planet arising from thermal fluctuations, or even a whole region resembling our visible universe. But for any specified subjective state $\mathcal{S}$, the overwhelmingly most likely BB would be an isolated brain surrounded by a random thermal bath [53]. Such fluctuations would be expected to quickly dissolve into the surrounding thermal equilibrium material, though even rarer fluctuations could persist for longer, or even much longer.

We use the term “ordinary observers,” or OOs, to describe observers like we believe we are, who evolved over a period of $\mathcal{O}(10^{14})$ years from a low-entropy hot big bang past. In some cosmological theories, but by no means all, BBs can outnumber OOs by an outrageously large or possibly infinite factor. We will refer to such theories as “BB dominated.” In theories of infinite multiverses, the question of whether BBs dominate over OOs can often depend on the choice of measure that is adopted to regulate the infinities. In the scale factor cutoff measure [54,55], for example, the ratio of BBs to OOs is finite and depends on parameters, so it can be either large or small. The proper time cutoff measure does not lead to BB domination, but it is ruled out in any case by the “youthfulness” problem [56–59].

As a simple example of a cosmological theory that leads to BB domination, we can consider the currently standard Λ CDM model¹⁴ that is used to describe our universe. Assuming that the vacuum is absolutely stable, the universe in this model approaches at late times a future-eternal de Sitter phase, as the matter density thins to zero, leaving only the cosmological constant to control the evolution. The Gibbons-Hawking temperature for this model is about $T_{\text{GH}} \sim 2 \times 10^{-34}$ eV, so the probability of spontaneously nucleating anything like a human brain is absurdly small. However, since the de Sitter phase would under these assumptions continue forever, the ratio of BBs to OOs would be infinite. To keep the discussion on the topic of very large but finite universes—to avoid the need to introduce a measure—we will imagine modifying the description so that the universe ends at some time, but we will consider an ending time that is arbitrarily long. In particular, we will assume that for any possible subjective state $\mathcal{S}$, the BBs in that state vastly outnumber the OOs in that state.

According to the PSLI, we should calculate first-person probabilities as if our current observer instant were chosen randomly and uniformly from all observer instants in the same subjective state in the asymptotic ensemble generated from the theory. Recall that a real observer in subjective state $\mathcal{S}$ would have absolutely no way of distinguishing herself

from any of the OOs or BBs in the ensemble in subjective state $\mathcal{S}$. (She could in principle determine from the model universes U_α which of these observers were OOs and which were BBs, but since she would not know her own status, there would be no logic to using this distinction to influence her beliefs about which of these observers in the model universe describes her own situation.) Thus, we see no reasonable alternative to her making predictions as if her current observer instant were chosen randomly from this group.

Using this method, the first-person probabilities would imply that any real observer is vastly more likely to be a BB than an OO. However, even if we focus on BBs who are selected to exactly match the current subjective state of a real observer, those BBs would still, with overwhelming probability, see the world around them rapidly dissolve into a thermal equilibrium mix. Thus, this model seems to clearly be in conflict with our observations. This consequence of the Λ CDM model with an absolutely stable vacuum was noticed by Page [60] in 2008, who suggested that it might show that our vacuum must decay on a timescale of gigayears—so that BBs cannot dominate.

But the situation is more subtle than it seems, because if the true theory really does predict BB domination, then with overwhelming likelihood we would be BBs. However, since BBs have not evolved in coordination with their surroundings over the history of the universe, everything they think they know is really just the result of the random configurations of atoms in their mental circuitry. There is, therefore, no reason for a BB to trust either his memory or thought processes. So, if BBs really vastly outnumber OOs, how can we rule out the possibility that we are BBs who merely think that we can describe and to some extent understand the real world? It might be all delusion.

Our understanding is that there is no logically sound way to rule out the possibility that a BB-dominated theory might be correct. In that case, however, we would almost certainly be BBs, and we would have no reason to trust any claims that we can “do science.” That is, there would be no reason to expect any validity to the theories that we construct, or imagine that we are constructing. The conclusion is that there is no point in pursuing any BB-dominated scientific theory, not because we have any good reason to believe that they are false, but rather because they are useless. If they are valid, then science is simply not possible.

Is there perhaps still some possibility that the correct theory is BB dominated, but we are nonetheless in the situation in which we happen not to be BBs, and therefore can (we hope) do science? Such a possibility, we will see, is very strongly disfavored by the kind of probabilistic reasoning that we have been advocating. Suppose, for example, that we consider two theories, $\mathcal{T}_1$ and $\mathcal{T}_2$. Theory $\mathcal{T}_1$ predicts that all observers in my subjective state $\mathcal{S}$ are OOs, while theory $\mathcal{T}_2$ predicts that only a tiny fraction q of such observers are OOs, with the rest being BBs. Suppose that I assign first-person prior probabilities to the two theories as $p^{(1p)}(\mathcal{T}_1|\mathcal{S}) = p_1$ and $p^{(1p)}(\mathcal{T}_2|\mathcal{S}) = p_2 = 1 - p_1$. Now I want to consider the probabilities that

¹⁴That is, the cosmological model which assumes that the dominant contributions to the energy density of the universe are a cosmological constant (Λ) and cold dark matter (CDM).

“I can do science” and its alternative, “I cannot do science.” We take these statements to be synonymous with “I am an OO” and “I am a BB,” respectively, and abbreviate them as “OO” and “BB,” respectively. Using the properties of the theories as stated, and evaluating first-person probabilities assuming the PSLI, we find

$$\begin{aligned} p^{(1p)}(\mathcal{T}_1 \text{ and OO}|\mathcal{S}) &= p^{(1p)}(\text{OO}|\mathcal{T}_1, \mathcal{S})p_1 = p_1, \\ p^{(1p)}(\mathcal{T}_2 \text{ and OO}|\mathcal{S}) &= p^{(1p)}(\text{OO}|\mathcal{T}_2, \mathcal{S})p_2 = qp_2, \\ p^{(1p)}(\mathcal{T}_1 \text{ and BB}|\mathcal{S}) &= p^{(1p)}(\text{BB}|\mathcal{T}_1, \mathcal{S})p_1 = 0, \\ p^{(1p)}(\mathcal{T}_2 \text{ and BB}|\mathcal{S}) &= p^{(1p)}(\text{BB}|\mathcal{T}_2, \mathcal{S})p_2 = (1 - q)p_2. \end{aligned} \quad (10)$$

Assuming that $q \ll p_1/p_2$, which we take as the quantitative criterion for BB domination, then only two of these four options have non-negligible probabilities: $\mathcal{T}_1$ and OO, or $\mathcal{T}_2$ and BB. Only the first of these two options, however, allows the possibility for us to do science. There is no option in which $\mathcal{T}_2$ is true *and* we can do science. Thus, there is no option (with non-negligible probability) in which $\mathcal{T}_2$ can be a useful scientific theory.

The subject of Boltzmann brains and their implications has been controversial in the literature. Carroll [61] has expressed views very similar to ours. In the abstract, Carroll wrote “The issue is not that the existence of such observers [a preponderance of Boltzmann brains] is ruled out by data, but that the theories that predict them are cognitively unstable: they cannot simultaneously be true and justifiably believed.” We agree with this conclusion completely, but we have minor quibbles with some statements that Carroll made along the way.¹⁵

¹⁵In particular, Carroll wrote (Sec. 5.3) that “While we know (or at least seem to perceive) that we are not disembodied brains floating in an otherwise empty void, we do not know that we and our local environments have not fluctuated into existence from a higher-entropy equilibrium state—in a long-lived, randomly fluctuating universe, it is overwhelmingly likely that we have indeed done so.” Although we “seem to perceive” that we are not disembodied brains, we, of course, cannot trust these perceptions if we are likely to be Boltzmann brains. We agree with Albrecht and Sorbo (in Sec. III C of Ref. [62]) that “The most likely fluctuation consistent with everything you know is simply your brain (complete with “memories” of the Hubble deep fields, microwave background data, etc.) fluctuating briefly out of chaos and then immediately equilibrating back into chaos again.” Carroll also wrote (also Sec. 5.3) that “The best we can do is to decline to entertain the possibility that the universe is described by a cognitively unstable theory, by setting our prior for such a possibility to zero (or at least very close to it). That is what priors are all about: setting credences for models on the basis of how simple and reasonable they seem to be before we have collected any relevant data.” To us, this sentence seems to contradict the key point that Carroll and we agree on: BB-dominated theories should be rejected *not* because they are difficult to believe, but because they cannot be used to do science.

In Ref. [63], Page took issue with Carroll’s Ref. [61], while also, like us, agreeing with Carroll’s main point about cognitive instability. But Page argued that BB-dominated theories can be ruled out without “appeal to this argumentation” by “conventional Bayesian reasoning.” In our view, Page’s argument appeared to rely on the ability of the reader to understand and verify the Bayesian argument even if he is a BB, which we find dubious.

Srednicki and Hartle (in Sec. VII of Ref. [8]) take a very different point of view, arguing that BB-dominated theories are perfectly acceptable, as one can prevent the prediction that we should be BBs by adopting a xerographic distribution that is “nonzero only for locations sufficiently close to a big bang to make it much more likely that we are OOs rather than BBs.” They argue that “scientists favor frameworks that are predictive,” and that a uniform xerographic distribution would be unpredictable, since then BBs would dominate, and the observations of any particular BB would be “overwhelmingly likely to be disordered, and inconsistent with its apparent history.”

We agree that predictivity is a very desirable attribute for a scientific theory, but our view of nonuniform xerographic distributions is exactly the opposite. It is, of course, very difficult to predict how any living creature will interpret the stimuli that it receives, and it may be more difficult to do this for BBs. But, under our assumption (see footnote 4) that observers can be described by the laws of physics, the observations of BBs are in principle predictable, like the observations of any other kind of creature.

We argued at the end of Sec. III C that if one takes the liberty to append to a theory an arbitrary xerographic distribution, the predictivity of the theory essentially disappears. If one wishes the theory to predict that the cosmic microwave background temperature should be 5 degrees, with everything else matching what we see, one can argue that if the universe is infinite or sufficiently large, this is likely to happen somewhere. We can then choose a xerographic distribution which requires us to be where it happens, and 5 degrees becomes the prediction of the theory. A BB-dominated theory is another example of this. Such a theory is unacceptable by our standards, but like almost any conceivable theory it is acceptable to those who think it makes scientific sense to imagine a xerographic distribution that makes us highly atypical.

Gott [64] dismissed Hartle and Srednicki’s claims that we might be atypical, pointing out that

All the other authors considering this question have implicitly assumed the Copernican principle—or the principle of mediocrity as Vilenkin has sometimes called it. As laid out by Gott [26], this runs as follows: the Copernican principle states that it is unlikely for your location in the Universe to be special among intelligent observers. Why? Because out of all the places for intelligent observers

to be there are by definition only a few special places and many nonspecial places, so you are simply likely to be at one of the many nonspecial places.

This is essentially a version of the counting argument that we described in Sec. III C, although Gott can be credited with having said it in 1993. Gott went on, however, to give two arguments why BB-dominated theories are nonetheless acceptable. First, he asserted that BBs are not intelligent observers, because they cannot pass a Turing test. While a very small fraction of BBs can perhaps pass for a human in answering any specified number of questions, the tendency of BBs to be very short-lived would make it highly unlikely that the next question could be successfully answered. This argument does not make sense to us, however, since a Turing test cannot be infinite. For any finite number of questions, in a very large universe the number of BBs that pass the test can vastly exceed the number of OOs who do. Gott also argues that BBs are not real, because they are “observer dependent.” This is a subtle question that involves the emergence of the classical world of our perceptions from an underlying quantum description, and is outside the scope of this paper. We stated in footnote 7 that we do not claim to have a full understanding of this issue, so we will not comment further on Gott’s argument.

D. Humans and Jovians

In this section, we discuss a parable, introduced by Hartle and Srednicki [20], that illustrates a situation which, they claim, shows that reasoning based on the assumption that we are typical can lead to absurd conclusions. Here, we argue that their example is indeed absurd, but is by no means evidence against the PSLI that we defined in Sec. III B. We will show that the parable can be modified so that the PSLI applies, and then conclusions are manifestly sensible.

The parable as described by Hartle and Srednicki goes as follows.

HJ parable. Consider two theories of the development of planet-based intelligent life, based on the appropriate physics, chemistry, biology, and ecology. Theory *A* predicts that there are likely to be intelligent beings living in the atmosphere of Jupiter; theory *B* predicts that there are no such beings. Because Jupiter is much larger than Earth, theory *A* predicts that there are today many more Jovians than humans.

Would we reject theory *A* solely because humans would not then be typical of intelligent beings in our Solar System? Would we use this theory to predict that there are no Jovians, because that is the only way we could be typical? Such a conclusion seems absurd (p. 2 in [20]).

The absurdity, referred to by the authors, trades on their surprise about reaching a strong conclusion about the population on Jupiter by simply demanding that we are typical intelligent beings (i.e., without making any observations). Absurd conclusions are easy to reach if one starts with the blind assumption that almost anything should be typical of most any reference group. As another example, if we assumed that we should be typical of life on Earth, we would conclude that we are overwhelmingly likely to be single-cell organisms. It similarly sounds absurd to us to assume that life that evolves in the atmosphere of Jupiter should be similar to intelligent life on Earth.

To show that there *are* situations in which the assumption of our typicality can be used rationally to strongly disfavor the existence of a hypothetical population, we will change two important characteristics of the HJ parable. First, Bayesian logic requires that we are discussing an update based on new information, and not merely the comparison of two theories *ab initio*. (To compare two theories *ab initio* is to choose priors, and there are no firm rules about how to do that.) Second, to apply the PSLI, we must confine the assumption of typicality to observers in the asymptotic ensemble who are in the same subjective state. (Only this very limited typicality assumption is needed to justify the methods proposed in this paper.)

For the situation to involve an update, we must be initially ignorant of whether we are “humans” or “jovians,” and to apply the PSLI it must be possible for us to be in the same subjective state in either case. For these reasons, we will consider a parable that involves only observers on Earth, but the “jovians” will be distinguished from the “humans” by their genetics—“jovians” each have a jovian gene, which is not linked to any observable traits, and the probability of carrying this gene is independent of a person’s subjective state. But the jovian gene can affect the outcome of treatments for one rare disease, jovian flu. Like the HJ parable, we will consider two theories about these populations. Theory *A* predicts that the population of Earth consists of 100 million people with no jovian gene, and 100 billion people with the jovian gene. Theory *B* predicts that the total population is 100 million, and nobody has the jovian gene. In the fanciful world of our parable, the population of the world has not been determined.

At the start of the parable, nobody knows if they are “human” or “jovian”, and nobody knows whether theory *A* or theory *B* is correct. The update occurs when I come down with jovian flu. I see my doctor and learn more about the disease. When untreated, the patient will usually have a very painful rash for two weeks but then recover. Fortunately, however, the hospital with which my doctor is affiliated has just developed the first Jovian flu antibiotic. It is completely untested, but reliable computer simulations show that if the antibiotic is given to an infected Jovian, the patient will be cured in three days, with no severe symptoms. However, if

the antibiotic is given to a human, it will make the disease worse, leading to a painful rash for a full month.

For reasons that he does not explain, my doctor assigns a very strong prior to theory A , so he plans my treatment accordingly. There is, unfortunately, no available test to distinguish “humans” from “jovians”, so the doctor relies on the methods of statistical medicine. Since there are 1000 times as many “jovians” as “humans” in the general population (according to his preferred theory, theory A), and since he cannot find any factors in my medical history to suggest that I am different from the averages, he concludes that the odds of my being a “jovian” are 1000:1.¹⁶ He therefore gives me the jovian flu antibiotic.

Unfortunately, the antibiotic intensifies my disease, and I have a painful rash for a month before I recover. Since theory A predicted the probability of this outcome as $1/1001$, while theory B predicted the probability as 1 , my doctor used Bayes’ theorem to update his degree of belief in the two theories. If we denote his prior estimates as p_A and p_B , his posterior probabilities are found from Bayes’ theorem to be

$$\begin{aligned} p_{\text{post},A} &= \frac{\frac{1}{1001} p_A}{\frac{1}{1001} p_A + p_B}, \\ p_{\text{post},B} &= \frac{p_B}{\frac{1}{1001} p_A + p_B}. \end{aligned} \quad (11)$$

That is, theory A has been strongly disfavored, because if theory A were true, I would have to be significantly atypical. Because of this typicality argument, we have strong evidence that the “jovian” population is nonexistent.

The reader will probably have noticed that, in our version of the human-Jovian parable, the logic is identical to that used in the urn problem discussed in Sec. IV B. If we draw a white ball from an urn, we discover evidence that the urn is more likely to be the one for which white balls are typical.

E. Use of “old evidence”

This paper has been mainly aimed at developing a method of assessing theories of a very large universe, in

¹⁶Clinicians are trained to look for signs that their patients might not match the characteristics of a study group, but, in the absence of such signs, it is standard practice to take the percentages of a study group as predicted probabilities for the patient. For example, in the *Clinicians’ Guide to Statistics for Medical Practice and Research: Part I* [65], an example is given of a study of heart transplant patients who received hearts from donors over the age of 35. 27% of the study patients experienced cardiac death within 5 years, so the textbook advises that the data can be used “to predict that the next patient seen in clinic ... will have a probability of 0.27 of experiencing cardiac death within 5 years.” In other words, in the absence of evidence to the contrary, it is standard medical practice to assume that any given patient is typical of the study group.

which the effective, low-energy approximations to the laws of physics might vary from one region to another. We have focused on the definition of first-person probabilities—probabilities predicted by a given theory for what *we* should expect to see, taking into account all selection effects. Such first-person probabilities would then be used to update the probabilities assigned to theories on the basis of new experiments, following the usual Bayesian method. When assessing theories, however, one cannot avoid the question of how previously acquired data should affect the probability that we might assign to a new theory. Philosophers refer to this as the “old evidence problem” [66]. A commonly discussed example is the role of the precession of the perihelion of Mercury’s orbit as evidence for general relativity, as the precession was known many years before the theory was formulated. The precession of the perihelion of Mercury is widely viewed as an important piece of evidence for general relativity, but it does not fit into the mold of a Bayesian update. We do not claim to have anything new to say about this issue, but for completeness we give a brief summary of the situation as we see it.

In practice, old evidence tends to be evidence that is known to very good precision, so the assessment of new theories in the light of old evidence is usually carried out without any sophisticated statistical analysis. In looking for alternatives to general relativity, we limit attention to theories that give the right prediction for the perihelion precession of Mercury, within reasonable errors. Nonetheless, for those of us who have adopted the Bayesian approach to theory assessment, it seems important to ask how such old evidence can fit into a Bayesian analysis.

Many authors [66] (including ourselves) agree that such old evidence $\mathcal{E}$ can be used to update the probability assigned to a theory $\mathcal{T}$ by imagining that $\mathcal{E}$ is not yet known, and then carrying out a standard Bayesian update. There are, however, some controversial issues associated with what exactly is meant by imagining that $\mathcal{E}$ is not yet known, since $\mathcal{E}$ might be logically intertwined with other pieces of information in our knowledge base.¹⁷

To resolve this issue, we would suggest that the situation in which $\mathcal{E}$ is not yet known can be defined historically. We can imagine turning back the clock to the time before $\mathcal{E}$ was discovered. Note that, by turning back the clock, we are presumably enlarging the reference class, since it now includes all observer instants that share all memories up to some point in the past, but which might diverge for later times. To carry out this turning back of the clock, we would have to ask what would have been reasonable priors to

¹⁷Reference [66] gives an example in which our knowledge base contains the pair of propositions $\{\mathcal{E}, \mathcal{E} \rightarrow \mathcal{B}\}$. But the knowledge base might alternatively be described by including the pair of propositions $\{\mathcal{E}, \mathcal{B}\}$, which leads to the same consequences. The two descriptions are equivalent, but if $\mathcal{E}$ is removed, then $\mathcal{B}$ would still hold in the second case, but not the first.

assign to the relevant theories at that time, and then we could carry out the Bayesian update. In choosing priors, we would be using modern theories, with a modern understanding of their consequences, but we would need to rely exclusively on experimental information available at the earlier time. Such a process would produce probability assignments that would have been relevant just after $\mathcal{E}$ was discovered, so to obtain currently relevant probabilities we might have to continue the process to include all other relevant experiments between then and now. A crucial difficulty with this procedure is the assignment of prior probabilities at the time just before $\mathcal{E}$ was discovered, because it is hard to know how much our choices might be biased by knowledge of data obtained later. For that reason, the implications of old evidence are generally more controversial—more subject to the differing opinions of various researchers—than the implications of new experiments.

V. CONCLUSION

Probabilistic reasoning in very large universes is a thorny issue, and analyses of these novel physical settings have led to confusion and controversy in the literature. We are interested in theories that describe universes in which the effective, low-energy approximations to the laws of physics can vary from one region to another, so the measurements of any observer will strongly depend on where she lives. Furthermore, the local low-energy laws of physics in any region can have a strong influence on whether observers exist in the region. If observers do exist, their number and their characteristics will also be strongly influenced by the local low-energy laws. In this complicated situation, there is as yet no generally accepted procedure to calculate probabilities for what a given observer might experience, or for how she should assess the credibility of competing theories.

In this paper, we have argued that these questions are not as novel as some authors have claimed, and that the best method of analysis is closely analogous to standard methods of physics. We distinguish between third-person probabilities, that would be measured by a hypothetical external observer who has access to the entire universe, and first-person probabilities, that would be measured by any particular real observer within the universe, taking into account all selection effects associated with the observer. We have made a specific proposal for how any particular observer should calculate the first-person probabilities for what she will experience. We proposed that such first-person probabilities should be used to update probabilities assigned to competing theories, using the completely standard Bayesian procedure. We focused on the principles that define this method, ignoring any practical issues associated with implementing it. We believe that it is important to settle these issues of principle first, and later one can seek practical approximations.

We described a generic theory $\mathcal{T}$ as a system that provides a set of possible universes $\{U_\alpha\}$, and a probability $P_{\text{cos}}(U_\alpha)$ that any one of them describes the real universe

U_{real} . Each U_α is a complete description of the entire spacetime and everything in it. We then imagined constructing a theory-generated ensemble of possible universes, with each possible universe contributing to the ensemble in proportion to $P_{\text{cos}}(U_\alpha)$.

To make predictions for any particular observer $\mathcal{O}$, one must take into account all selection effects. We argued that in principle each observer is subject to her own selection effects, and that these selection effects can in general depend on time. We concluded that predictions should be made separately for each observer instant $\mathcal{I}$, referring to a particular observer $\mathcal{O}$ at a particular time t . Predictions can be made by associating a given observer instant $\mathcal{I}$ in the real universe with a reference class of observer instants $\mathcal{I}_i$ in the theory-generated ensemble. We believe that the best method of prediction should take into account all available information, so we defined the *subjective state* $\mathcal{S}$ of an observer instant as the set of all properties that the observer believes, at time t , about herself and her surroundings. The reference class then consists of all observer instants $\mathcal{I}_i$ in the same subjective state as $\mathcal{I}$.

There will, in general, be many different $\mathcal{I}_i$'s in different model universes U_α in the ensemble, so we need a procedure to extract predictions from the set. By these definitions the real observer $\mathcal{O}$ cannot, at time t , distinguish any of the observer instants $\mathcal{I}_i$ from her own observer instant, so she has no evidence that any one of them is more likely to describe her future than any other. These are precisely the conditions under which the principle of indifference is usually taken to apply. In this context, we propose a PSLI, which holds that an observer should predict first-person probabilities for her future as if her current observer instant were chosen randomly and uniformly from the observer instants in the ensemble that are in the same subjective state. We pointed out that this is analogous what physicists assume in standard laboratory settings, where it is always assumed that the results obtained in the experiment are typical, within expected statistical fluctuations, of the results that any experiment of the same type would yield.

We believe that the PSLI is a very reasonable principle, but we also argue that using this method optimizes the accuracy of predictions. In a probabilistic system, there is never any way to guarantee that a specific prediction will turn out to be true, but we argued (in Appendix A) that the use of the PSLI by real observers maximizes the fraction of them who will make correct predictions if the theory is correct. Whether one is discussing the outcomes for the flipping of idealized fair coins or for experiments in a very large universe, one can predict what most observers will find, but one cannot guarantee what any particular observer will find. In either case, we believe that it is sensible to assume that what most observers will find is most likely what I will find. Some authors have argued that there is a “random process” that motivates this assumption for the case of coin flipping, but not for the self-locating

uncertainties involved in predicting experimental outcomes in a very large universe. In Sec. IV A, we argued that there is no relevant difference between these two situations.

A possibly controversial consequence of the PSLI is the “weighting by the number of observers”—that is, in determining first-person probabilities, the weight associated with any possible model universe U_α is proportional to both its third-person probability $P_{\text{cos}}(U_\alpha)$ and to the number of observer instants it contains that are in the same subjective state as the observer instant for which predictions are sought. In contrast, Hartle and collaborators (in Refs. [8,13,20,39,67], for example) advocate weighting each possible model universe only by its third-person probability, restricted to universes that have at least one instance of our data D .¹⁸ In Appendix B, we show how the weighting by observers can be derived from a set of simpler, reasonable assumptions.

One issue for which probabilistic reasoning in very large universes is relevant is the question of Boltzmann brains. In Sec. IV C, we argued that theories in which Boltzmann brains dominate over ordinary observers should be rejected. They should be rejected not because we see a rational reason why they cannot be valid, but instead because, if they are valid, then we would most likely be Boltzmann brains and would not be capable of doing science. Thus, we see no possibility that such theories can be useful for scientific investigations.

One possibly significant issue that we did not consider in this paper, but leave as a topic for future investigation, is the time evolution of subjective states and its possible relevance to the calculation of first-person probabilities. One possibility is that subjective states are well-modeled as being discrete, so an observer evolves for a short period of time in one subjective state, and then transitions instantaneously to a subsequent state. In that case, it might be appropriate to calculate first-person probabilities by weighting these intervals in proportion to their duration, rather than assuming that all observer instants count equally. Alternatively, it might be better to treat the time evolution of subjective states as a continuous evolution, in which case one might consider whether the weight given to an observer instant might depend on how quickly the observer passes through the specified subjective state.

The methods developed in this paper are relevant to several interesting questions which we have not discussed

here. One is the question of whether evidence that the universe is fine-tuned to allow for the possibility of life should be interpreted as evidence in favor of a multiverse. Another is the so-called “doomsday argument”—the assertion that human civilization is not likely to survive much longer than its current age, since a much longer period of survival would place us atypically early in human history. We leave these questions as topics for future research.

Note that this paper is directly relevant to very large universes where the total spacetime volume is finite. However, many cosmological theories under consideration today describe infinite “multiverses,” where the definition of probabilities requires a solution to what is known as the measure problem. Of course, there remains much work to be done. We are very far from being able to produce cosmological theories for which we can calculate consequences at the level assumed in this paper. It is not clear how the measure problem will be resolved, but perhaps theoretical reasoning can lead to a unique solution, or perhaps it will be necessary to consider a range of solutions to be tested observationally. In any case, we expect that the methods developed in this paper, based on the principle of self-locating indifference, will continue to be applicable as these advances unfold.

ACKNOWLEDGMENTS

We thank Jeremy Butterfield, Don Page, and Scott Sheffield for helpful conversations. F. A. acknowledges support from (i) the Black Hole Initiative at Harvard University, where this work began, which is funded through a grant from the John Templeton Foundation and the Gordon and Betty Moore Foundation; and (ii) the Faculty Research Support Program (FY20FRSPE) at the University of Notre Dame. A. H. G.’s work was supported in part by the U.S. Department of Energy, Office of Science, Office of High Energy Physics under Grant Contract No. DE-SC0012567.

DATA AVAILABILITY

No data were created or analyzed in this study.

APPENDIX A: MOTIVATING THE PRINCIPLE OF SELF-LOCATING INDIFFERENCE ASSUMING THAT THE MODEL U_* FOR THE REAL UNIVERSE IS KNOWN

As part of our justification of the principle of self-locating indifference, here we adopt the simplifying assumption that we know which model universe $U_* \in \{U_\alpha\}$ is actually realized. Since the goal is to define a procedure for extracting first-person predictions from theories in very large universes, we will use the success of such predictions as our criterion for comparing different possible procedures. Thus, we will motivate the PSLI by considering a toy model of an experiment, and we will show that the expected number of

¹⁸The definition of “data” used by Hartle and collaborators seems very similar to our definition of “subjective state.” In Ref. [8], the authors describe “the data D_0 that the HSI [human scientific IGUS, where IGUS = information gathering and utilizing system] has: every scrap of information that the HSI possesses about the physical universe, including every record of every experiment, every astronomical observation of distant galaxies, every available description of every leaf, etc., and necessarily every piece of information about the HSI itself, its members, and its history.”

observers who will make correct inferences from their experiments is maximized if every observer interprets his data using the PSLI. This is an application of the same “counting argument” that we discussed in Sec. III C: If we can devise a procedure for which observers who make incorrect inferences are very rare, then we can assume that we are very unlikely to fall into this class.

For our toy model experiment, we will imagine that some observable quantity θ is measured by many different observer instants $\mathcal{I}$ in the same subjective state, all in the universe described by U_* . The quantity θ might be, for example, the value of the Higgs mass, or the fine-structure constant, or the cosmological constant. While any observable in reality could be correlated with any number of other observables, we will assume for our toy model that θ has no correlations with any other observable. Furthermore, we will assume that θ has the same value throughout the universe.

Consider two possible assumptions about typicality that each of the observers could adopt: “typicality” and “atypicality.” Define typicality as the assumption of the PSLI—the assumption that each observer should calculate first-person probabilities as if their current observer instant were chosen randomly and uniformly from among all the observer instants $\mathcal{I}_i$. Observers who adopt typicality will be referred to as “typicality accepting.” We also consider an atypicality assumption—the assumption that each observer should calculate first-person probabilities as if their current observer instant is expected to be atypical in some specific way. Observers who adopt atypicality will be referred to as “typicality denying.” For our toy model, typicality-denying observers will assume that their first-person expectation value for the measurement of θ is higher than the theoretical value by 10%. (Note that the assumption that the first-person probability distribution for θ is affected by the atypicality assumption is generic. If no observables are affected, then the atypicality assumption is completely irrelevant. As long as at least one observable is affected, we can choose θ to be that observable. We assume here for simplicity that the expectation value is affected, but similar conclusions could be justified if the standard deviation or any other property of the probability distribution is affected.)

The observers use their assumptions about their own typicality in interpreting the results of their measurements. We will show that the assumption of typicality leads to the largest expectation value for the fraction of observers reaching accurate inferences. In the absence of any way to know where we will lie in the distribution of outcomes, the best way to ensure that our inferences will be correct is to maximize the expectation value of the number of observers who will make accurate inferences.

The observers have (the same) theories of the measurement of this observable at their disposal. Each theory yields a probability density function for the measurement of θ , which includes any experimental errors that arise in the measurement process, as well as any quantum uncertainties

intrinsic to the measurement. We denote a given theory by $\mathcal{T}(\theta_t)$, where we have labeled the theory via the mean θ_t of the probability density function that the theory yields for θ . The probability density function for measuring θ , according to the theory $\mathcal{T}(\theta_t)$, will then be denoted by $p(\theta|\theta_t)$.

We will proceed under the simplifying assumption that this distribution is a normal distribution:

$$p(\theta|\theta_t) = \mathcal{N}_\theta(\theta_t, \sigma), \quad (\text{A1})$$

where $\mathcal{N}_\theta(\theta_t, \sigma)$ represents a normal distribution in θ with mean θ_t and standard deviation σ . Denote the true value of θ in the model universe U_* by θ_{true} , which we assume to be nonzero. For our toy model experiment, we will take $\sigma = \theta_{\text{true}}/100$. We will denote the measured value of θ by θ_D (where “D” is for “data”).

For our example, suppose that all the observers adopt a prior probability distribution that is flat and ranges from $-10\theta_{\text{true}}$ to $10\theta_{\text{true}}$; that is,

$$p_{\text{prior}}(\theta_t) = U_{\theta_t}(-10\theta_{\text{true}}, 10\theta_{\text{true}}), \quad (\text{A2})$$

where $U_\theta(a, b)$ denotes the uniform distribution for the random variable θ , between $\theta = a$ and $\theta = b$. [The observers, of course, do not know the value of θ_{true} , but we will assume for our example that their choice of prior corresponds to Eq. (A2).]

According to the assumptions stated above about typicality-accepting ($a = \text{typ}$) and typicality-denying ($a = \text{atyp}$) observers, the first-person probabilities that these observers will adopt can be written as

$$p_a^{(1p)}(\theta|\theta_t) \equiv \mathcal{N}_\theta(\kappa_a \theta_t, \sigma), \quad (\text{A3})$$

where

$$\kappa_a = \begin{cases} 1 & \text{if } a = \text{typ}, \\ 1.1 & \text{if } a = \text{atyp}. \end{cases} \quad (\text{A4})$$

Recall that we are working in a Bayesian setting, where the primary distribution of interest, as the observer completes further trials, is the first-person posterior probability $p_{\text{post};a}^{(1p)}(\theta_t|\theta_D)$. This is given, in general, by Bayes’ theorem:

$$p_{\text{post};a}^{(1p)}(\theta_t|\theta_D) = \frac{p_a^{(1p)}(\theta_D|\theta_t)p_{\text{prior}}(\theta_t)}{p(\theta_D)}, \quad (\text{A5})$$

where

$$p(\theta_D) = \int d\theta_t p_a^{(1p)}(\theta_D|\theta_t)p_{\text{prior}}(\theta_t). \quad (\text{A6})$$

Our two classes of observers can then compute two different first-person posterior probability distributions for θ_t by substituting Eqs. (A2) and (A3) into Eq. (A5).

The posterior probability density for either kind of observer peaks at $\theta_t = \theta_D/\kappa_a$. One may ask, what is the probability that, according to the true theory, the peak is within a range around the true value, say in the range $[\theta_{\text{true}}(1 - \epsilon), \theta_{\text{true}}(1 + \epsilon)]$, where ϵ is a small number? This probability is given by

$$\begin{aligned} P\left(|(\theta_D/\kappa_a) - \theta_{\text{true}}| < \epsilon\theta_{\text{true}} \middle| \theta_{\text{true}}\right) \\ &= \int_{\kappa_a\theta_{\text{true}}(1-\epsilon)}^{\kappa_a\theta_{\text{true}}(1+\epsilon)} p(\theta_D|\theta_{\text{true}})d\theta_D \\ &= \frac{1}{2}\text{erf}\left(\frac{\theta_{\text{true}}}{\sqrt{2}\sigma}[\kappa(1+\epsilon) - 1]\right) \\ &\quad - \frac{1}{2}\text{erf}\left(\frac{\theta_{\text{true}}}{\sqrt{2}\sigma}[\kappa(1-\epsilon) - 1]\right). \end{aligned} \quad (\text{A7})$$

For typicality-accepting observers, where $\kappa_{\text{typ}} = 1$, this probability is significantly higher than for typicality-denying observers. For example, for $\epsilon = 2\%$, this probability is 0.95 for typicality-accepting observers and 3.1×10^{-15} for typicality-denying observers. Thus the expected number of observers who will find the peak in the specified range is vastly larger if the observers are typicality accepting.

The toy model example described here is very simplified, and uses numbers chosen to be fairly extreme, but the qualitative result is universal. Ideal experiments are designed to be “unbiased,” which means that the expectation value of the result should be equal to the true value of the quantity. The expectation value is what, in our language, is the expected average value that would be obtained by typical observers. If one were to apply a correction to the result, motivated by the assumption that we are atypical, then the protocol would become biased. The toy model example in this appendix is really just a step-by-step illustration of why unbiased measurements are considered to be obviously preferable to biased measurements (provided that all other aspects are equal).

APPENDIX B: MOTIVATING THE PRINCIPLE OF SELF-LOCATING INDIFFERENCE: DROPPING THE ASSUMPTION THAT U_* IS KNOWN

In this appendix, we continue our justification of the principle of self-locating indifference, now dropping the assumption that we know which U_α is the model universe U_* that is realized—that is, which model universe U_α describes the real universe U_{real} . This requires us to show that, for any given observer instant $\mathcal{I}$ in the real world, the probability that any proposition $\mathcal{P}$ is true can be calculated as if $\mathcal{I}$ were chosen randomly and uniformly from the set of all observer instants in the asymptotic ensemble that are in the same subjective state $\mathcal{S}$. In Sec. III B, we showed that this statement is equivalent to evaluating $p^{(1p)}(\mathcal{P}|\mathcal{S})$, first-person probability for a proposition $\mathcal{P}$, given $\mathcal{S}$, using

Eq. (7). In this appendix, we will derive Eq. (7) starting from assumptions that we consider very reasonable, and somewhat more primitive than the assumption of the PSLI itself.

One of the most striking features of the PSLI, which will be justified in this appendix, is the weighting by the number of observers. That is, the PSLI implies that $p^{(1p)}(U_\alpha|\mathcal{S})$, the first-person probability that a given observer-instant $\mathcal{I}$ will find itself in a universe described by U_α , is weighted by the number of observer instants found in U_α that are in the same subjective state $\mathcal{S}$ that $\mathcal{I}$ is.

Throughout this appendix, we will be concerned with defining first-person probabilities for a particular observer instant $\mathcal{I}$ in the real universe. The subjective state of $\mathcal{I}$ will be denoted by $\mathcal{S}$.

For each model universe U_α allowed by the theory, there might be any number of observer instants $\mathcal{I}_i$ that are in the subjective state $\mathcal{S}$. Although they are all in the same subjective state, they all have different spacetime locations, so each can be labeled with a distinct index i , as we assumed when we first wrote “ $\mathcal{I}_i$ ”. We will refer to i as the *observer-instant number*, which ranges from $i = 1$ up to the number of such observer instants in U_α , which we denote by $n_{\text{tot}}(U_\alpha, \mathcal{S})$. It does not matter how the observer-instant numbers are assigned, so we can imagine that they are assigned randomly by whoever is analyzing the theory to extract its predictions.

Our discussion will be based on the asymptotic ensemble derived from the theory under investigation, as described in detail in Sec. III B. Each entry U_λ^{ens} is one of the possible universes U_α , so each observer instant in U_λ^{ens} can be assigned the observer-instant number that it inherits from U_α . Thus, each observer instant in the ensemble can be labeled by the triplet $(\lambda, \mathcal{S}, i)$, where U_λ^{ens} is the model universe in which it occurs, $\mathcal{S}$ is its subjective state, and i is its observer-instant number.

To proceed, we will argue that, in the context of extracting first-person probabilities from a given theory, we can imagine that each observer instant in the real universe can also be assigned a unique observer-instant number. To see this, first note that the first-person probability of any proposition $\mathcal{P}$ implied by a given theory is simply the answer to the question “if the theory is correct, what is the first-person probability of $\mathcal{P}$?” That is, when extracting a predicted probability from a theory, one is always conditionalizing on the assumption that the theory is correct. Under that assumption, the real universe U_{real} must be perfectly described by some $U_* \in \{U_\alpha\}$, where $\{U_\alpha\}$ is the set of possible universes allowed by the theory. While the subjective state $\mathcal{S}$ of the observer instant $\mathcal{I}$ in the real universe may occur many times in U_* , the spacetime location of $\mathcal{I}$ occurs only once, so $\mathcal{I}$ can, in principle, be identified with a particular $\mathcal{I}_i$ in U_* and can therefore be assigned the observer-instant number i . The observer-instant number for $\mathcal{I}$ can be determined only by an external

observer who has access to the entire real universe, but we see no barrier to considering it in probabilistic calculations.

Suppose that we know that i is the observer-instant number for $\mathcal{I}$, but we do not know which possible universe is the one that is realized. In that case, the mere existence of observer-instant number i (for the subjective state $\mathcal{S}$) implies that

$$n_{\text{tot}}(U_\alpha, \mathcal{S}) \geq i. \quad (\text{B1})$$

The criterion of Eq. (B1) divides the entries of the ensemble into two classes: (i) *allowed universes*, for which Eq. (B1) holds, and which therefore contain exactly one observer instant compatible with $\mathcal{I}$, and (ii) *forbidden universes*, for which the equation fails, and which therefore cannot be the realized universe.

Finally, to determine the first-person probability $p^{(1p)}(U_\alpha, i|\mathcal{S})$, we will make two assumptions:

- $\mathcal{A}_1$: Suppose that we know which model universe $U_* \in \{U_\alpha\}$ describes the real universe U_{real} , as we assumed in the previous appendix. Then first-person predictions for $\mathcal{I}$ should be made as if $\mathcal{I}$ were chosen randomly and uniformly from the set of observer instants $\mathcal{I}_i$ in U_* in the same subjective state, which implies that i should be chosen randomly and uniformly from the set $\{1, 2, \dots, n_{\text{tot}}(U_*, \mathcal{S})\}$. (This statement is just the PSLI restricted to the case in which U_* is known, as discussed in the previous appendix.)
- $\mathcal{A}_2$: Suppose that we know the observer-instant number i of observer instant $\mathcal{I}$, but we do not know which model universe $U_* \in \{U_\alpha\}$ describes the real universe. Then first-person predictions for U_* should be made as if U_* were chosen randomly and uniformly from the *allowed* universes in the asymptotic ensemble.

To proceed, we need to consider the realistic case in which we do not know either U_* or i . Thus, we focus on $p^{(1p)}(U_\alpha, i|\mathcal{S})$, the first-person conditional probability that U_α is the universe that is realized and that i is the observer-instant number of $\mathcal{I}$, given the subjective state $\mathcal{S}$. We will find that $\mathcal{A}_1$ determines the dependence of this quantity on i , and that $\mathcal{A}_2$ determines its dependence on U_α , so it is determined completely.

Before giving the general proof, we will give a simple example to illustrate the logic. Suppose that a theory predicts only two possible universes, U_1 and U_2 , with equal probabilities $P_{\text{cos}}(U_\alpha)$, where U_1 contains a million observer instants in my subjective state $\mathcal{S}$, while U_2 contains only one. Now suppose that I am told by a hypothetical external observer that my observer-instant number is 1. In that case, there is exactly one observer instant in U_1

compatible with my subjective state and observer-instant number, and similarly one compatible observer instant in U_2 . Both universes are therefore allowed, so by assumption $\mathcal{A}_2$, I am equally likely to live in either universe: $p^{(1p)}(U_2, i=1|\mathcal{S}) = p^{(1p)}(U_1, i=1|\mathcal{S})$. But according to assumption $\mathcal{A}_1$, if I live in universe U_1 , I should be equally likely to have any observer-instant number from 1 to 1 000 000. So $p^{(1p)}(U_1, i|\mathcal{S}) = p^{(1p)}(U_1, 1|\mathcal{S})$, for $i \in [1, 10^6]$. It then follows that if I do not know my observer-instant number, the first-person probability that I live in U_1 is given by $p^{(1p)}(U_1|\mathcal{S}) = \sum_i p^{(1p)}(U_1, i|\mathcal{S}) = 10^6 p^{(1p)}(U_1, i=1|\mathcal{S}) = 10^6 p^{(1p)}(U_2, i=1|\mathcal{S}) = 10^6 p^{(1p)}(U_2|\mathcal{S})$, demonstrating the weighting by the number of observers.

To construct the general proof, we begin by expressing assumption $\mathcal{A}_1$ by the equation

$$p^{(1p)}(i|U_\alpha, \mathcal{S}) = \frac{\Theta(U_\alpha, \mathcal{S}, i)}{n_{\text{tot}}(U_\alpha, \mathcal{S})}, \quad (\text{B2})$$

where

$$\Theta(U_\alpha, \mathcal{S}, i) = \begin{cases} 1 & \text{if } n_{\text{tot}}(U_\alpha, \mathcal{S}) \geq i, \\ 0 & \text{otherwise} \end{cases} \quad (\text{B3})$$

and where we are assuming that $n_{\text{tot}}(U_\alpha, \mathcal{S}) > 0$. If $n_{\text{tot}}(U_\alpha, \mathcal{S}) = 0$ and observer instant $\mathcal{I}$ exists, then the combination $(U_\alpha, \mathcal{S})$ is not possible. Using the definition of conditional probability, $p^{(1p)}(i|U_\alpha, \mathcal{S}) = p^{(1p)}(U_\alpha, i|\mathcal{S}) / p^{(1p)}(U_\alpha|\mathcal{S})$, Eq. (B2) can be rewritten as

$$p^{(1p)}(U_\alpha, i|\mathcal{S}) = \frac{p^{(1p)}(U_\alpha|\mathcal{S})\Theta(U_\alpha, \mathcal{S}, i)}{n_{\text{tot}}(U_\alpha, \mathcal{S})}. \quad (\text{B4})$$

To extract the consequences of $\mathcal{A}_2$, recall that choosing a universe randomly and uniformly from the asymptotic ensemble is equivalent to choosing it according to the theory-generated probability $P_{\text{cos}}(U_\alpha)$. Thus, $\mathcal{A}_2$ can be expressed as the equation

$$p^{(1p)}(U_\alpha|\mathcal{S}, i) = p(U_\alpha|U_\alpha \text{ is allowed for } (\mathcal{S}, i)) \quad (\text{B5a})$$

$$= p(U_\alpha|\Theta(U_\alpha, \mathcal{S}, i) = 1) \quad (\text{B5b})$$

$$= \frac{1}{Z_1(\mathcal{S}, i)} p(U_\alpha \text{ and } \Theta(U_\alpha, \mathcal{S}, i) = 1) \quad (\text{B5c})$$

$$= \frac{1}{Z_1(\mathcal{S}, i)} P_{\text{cos}}(U_\alpha)\Theta(U_\alpha, \mathcal{S}, i), \quad (\text{B5d})$$

where

$$Z_1(\mathcal{S}, i) = \sum_\beta P_{\text{cos}}(U_\beta)\Theta(U_\beta, \mathcal{S}, i). \quad (\text{B6})$$

To understand why the expression (B5d) is equal to the

expression (B5c), note that $\Theta(U_\alpha, \mathcal{S}, i)$ is equal to either 1 or 0, and that the equality holds in both cases.

By the definition of conditional probability, $p^{(1p)}(U_\alpha, i|\mathcal{S}) = p^{(1p)}(U_\alpha|\mathcal{S}, i)Z_2(\mathcal{S}, i)$, where $Z_2(\mathcal{S}, i) \equiv p^{(1p)}(i|\mathcal{S})$, so with Eq. (B5d) we have

$$p^{(1p)}(U_\alpha, i|\mathcal{S}) = P_{\cos}(U_\alpha)\Theta(U_\alpha, \mathcal{S}, i)\frac{Z_2(\mathcal{S}, i)}{Z_1(\mathcal{S}, i)}. \quad (\text{B7})$$

To combine the information in Eqs. (B4) and (B7), note that, when $\Theta(U_\alpha, \mathcal{S}, i) = 0$, both equations imply that $p^{(1p)}(U_\alpha|\mathcal{S}, i) = 0$. When $\Theta(U_\alpha, \mathcal{S}, i) = 1$, their equality implies that

$$\frac{p^{(1p)}(U_\alpha|\mathcal{S})}{n_{\text{tot}}(U_\alpha, \mathcal{S})} = P_{\cos}(U_\alpha)\frac{Z_2(\mathcal{S}, i)}{Z_1(\mathcal{S}, i)}. \quad (\text{B8})$$

Since the left-hand side is independent of i , the right-hand side must also be, so we can write

$$\frac{Z_2(\mathcal{S}, i)}{Z_1(\mathcal{S}, i)} = \frac{1}{Z_3(\mathcal{S})} \quad (\text{B9})$$

for some function $Z_3(\mathcal{S})$.

One can also insert Eqs. (B8) and (B9) into Eq. (B4), showing that

$$p^{(1p)}(U_\alpha, i|\mathcal{S}) = \frac{1}{Z_3(\mathcal{S})}P_{\cos}(U_\alpha)\Theta(U_\alpha, \mathcal{S}, i). \quad (\text{B10})$$

The above equation was derived under the assumption that $\Theta(U_\alpha, \mathcal{S}, i) = 1$, but it is clearly also true in the alternative case, where $\Theta(U_\alpha, \mathcal{S}, i) = 0$ and both sides are zero.

$Z_3(\mathcal{S})$ can be determined by normalization:

$$\begin{aligned} Z_3(\mathcal{S}) &= \sum_{\alpha, i} P_{\cos}(U_\alpha)\Theta(U_\alpha, \mathcal{S}, i) \\ &= \sum_{\alpha} P_{\cos}(U_\alpha)n_{\text{tot}}(U_\alpha, \mathcal{S}). \end{aligned} \quad (\text{B11})$$

To rederive Eq. (7), we can use Eq. (B10) to calculate the first-person probability for any proposition $\mathcal{P}$, given the subjective state $\mathcal{S}$ of the observer instant $\mathcal{I}$:

$$p^{(1p)}(\mathcal{P}|\mathcal{S}) = \sum_{\alpha, i} p^{(1p)}(\mathcal{P}|U_\alpha, \mathcal{S}, i)p^{(1p)}(U_\alpha, i|\mathcal{S}), \quad (\text{B12})$$

where $p^{(1p)}(\mathcal{P}|U_\alpha, \mathcal{S}, i)$ is the probability that the proposition $\mathcal{P}$ is true for a specified observer instant, given the model universe U_α , the subjective state $\mathcal{S}$, and observer-instant number i of the observer instant. Since the observer instant is fully described by these parameters, $\mathcal{P}$ is either true or false for this observer instant, so

$$p^{(1p)}(\mathcal{P}|U_\alpha, \mathcal{S}, i) = \begin{cases} 1 & \text{if } \mathcal{P} \text{ is true for observer instant } (U_\alpha, \mathcal{S}, i) \\ 0 & \text{otherwise.} \end{cases} \quad (\text{B13})$$

Substituting $p^{(1p)}(U_\alpha, i|\mathcal{S})$ from Eq. (B10) into Eq. (B12),

$$\begin{aligned} p^{(1p)}(\mathcal{P}|\mathcal{S}) &= \frac{1}{Z_3(\mathcal{S})} \sum_{\alpha, i} P_{\cos}(U_\alpha)\Theta(U_\alpha, \mathcal{S}, i) \\ &\quad \times p^{(1p)}(\mathcal{P}|U_\alpha, \mathcal{S}, i). \end{aligned} \quad (\text{B14})$$

Note that whenever $p^{(1p)}(\mathcal{P}|U_\alpha, \mathcal{S}, i) = 1$, the proposition $\mathcal{P}$ must be true for the observer instant indicated by $(U_\alpha, \mathcal{S}, i)$, and therefore the observer instant must exist, and therefore $\Theta(U_\alpha, \mathcal{S}, i) = 1$. Thus, the factor $\Theta(U_\alpha, \mathcal{S}, i)$ in the above equation is superfluous and can be dropped:

$$p^{(1p)}(\mathcal{P}|\mathcal{S}) = \frac{1}{Z_3(\mathcal{S})} \sum_{\alpha, i} P_{\cos}(U_\alpha)p^{(1p)}(\mathcal{P}|U_\alpha, \mathcal{S}, i). \quad (\text{B15})$$

Now note that

$$\sum_i p^{(1p)}(\mathcal{P}|U_\alpha, \mathcal{S}, i) = n_{\text{prop}}(U_\alpha, \mathcal{S}, \mathcal{P}), \quad (\text{B16})$$

where we recall that $n_{\text{prop}}(U_\alpha, \mathcal{S}, \mathcal{P})$ was defined in Sec. III B as the number of observer instants in subjective state $\mathcal{S}$ in the model universe U_α for which the proposition $\mathcal{P}$ is true. Using Eqs. (B15), (B16), and (B11), we have finally

$$p^{(1p)}(\mathcal{P}|\mathcal{S}) = \frac{\sum_{\alpha} P_{\cos}(U_\alpha)n_{\text{prop}}(U_\alpha, \mathcal{S}, \mathcal{P})}{\sum_{\beta} P_{\cos}(U_\beta)n_{\text{tot}}(U_\beta, \mathcal{S})}, \quad (\text{B17})$$

which is identical to Eq. (7). Thus, we have rederived the PSLL.¹⁹

¹⁹The reader may have noticed that the situation discussed here has analogies with the *Sleeping Beauty* problem, as discussed, for example, by Elga in Ref. [68]. In that scenario, the number of awakenings of Sleeping Beauty is determined by the flip of a coin, while in this problem the number of observer instants with a given subjective state is determined by the random choice of which universe U_α is realized. The conclusion that we describe here is analogous to the “thirder” position of the Sleeping Beauty problem [i.e., giving 1/3 as Beauty’s credence that the coin in the thought experiment landed heads, in a scenario where heads leads to one awakening (on Monday), and tails leads to two (on Monday and Tuesday).] The observer-instant numbers in our scenario are analogous to the days of the week in *Sleeping Beauty*. The thirder answer is in accordance with the understanding that if one outcome leads to more observations than another, then the probability of observing that outcome is enhanced. The logic that we used here to justify our conclusion is, in fact, a generalization of the argument given by Elga in Ref. [68]. Physicists whom we know generally take the thirder position, as exemplified by Polchinski’s blog post [69].

The “weighting by the number of observers” is incorporated in the factor $n_{\text{prop}}(U_\alpha, \mathcal{S}, \mathcal{P})$ in the numerator of Eq. (B17), which can be seen to arise in Eq. (B16) from the fact that each observer instant $\mathcal{I}_i$ contributes to the overall probability.

APPENDIX C: COMPARING PRESCRIPTIONS IN SIMPLE BOX MODELS

In this appendix, we consider a simple “box model” universe of the same general type that was considered by Hartle and collaborators in Refs. [8,13,20,39]. (In Refs. [8,20], the boxes were referred to as “cycles in time.”) These calculations will serve as a concrete example of how such models are treated by the method of many repetitions that we advocate, and they will also provide an opportunity to compare with several different methods discussed by Hartle and collaborators in the papers mentioned above.

The box model that we will consider, illustrated in Fig. 1, is a generalization of the box models considered in the Hartle *et al.* papers mentioned above. The model is described by the following assumptions:

- The theories $\mathcal{T}$ that we will consider will have the form described by Fig. 1. Each theory will allow model universes with either of two *box types*, labeled by $k = 1$ or $k = 2$. Each $k = 1$ universe will consist of N_1^R red boxes and N_1^B blue boxes. These boxes might be thought of as being separated from each other in time or in space (or both)—it does not matter. Similarly, each $k = 2$ universe will consist of N_2^R red boxes and N_2^B blue boxes.
- Each box of each model universe might or might not contain observers. (Time dependence is neglected in these simple box models, so we make no distinction between observers and observer instants.) For simplicity, we will assume that all the observers are identical, and in particular are in the same subjective state $\mathcal{S}$. We will also assume that they can measure the color of their box without error. The number of observers in the i th red box will be denoted by n_i^R , where i ranges from 1 to the number of red boxes, N_k^R . Similarly, the number of observers in the j th blue box

will be denoted by n_j^B , where j ranges from 1 to the number of blue boxes, N_k^B . To make contact with the notation that we used in Sec. III, a complete description of a model universe U_α is obtained by equating α with the multiplet of parameters specified above:

$$\alpha \equiv (k, n_1^R, \dots, n_{N_k^R}^R, n_1^B, \dots, n_{N_k^B}^B). \quad (\text{C1})$$

- Each theory $\mathcal{T}$ specifies the numbers N_1^R, N_1^B, N_2^R , and N_2^B of each type and color of box, and also specifies the probabilities $P_{\text{cos}}(U_\alpha)$ for any possible universe U_α . The probability that the box type is $k = 1$ will be denoted by p_1 , while the probability for $k = 2$ will be $p_2 = 1 - p_1$. We will assume that the probability that any red box contains exactly n observers is given by the function $p_{\text{obs}}^R(n)$, and similarly the probability that any blue box contains exactly n observers is given by $p_{\text{obs}}^B(n)$. Thus,

$$P_{\text{cos}}(U_\alpha) = p_k \prod_{\text{red boxes}} p_{\text{obs}}^R(n_i^R) \prod_{\text{blue boxes}} p_{\text{obs}}^B(n_j^B). \quad (\text{C2})$$

The quantities $N_1^R, N_1^B, N_2^R, N_2^B, p_1$, and p_2 , and the functions $p_{\text{obs}}^R(n)$ and $p_{\text{obs}}^B(n)$ all depend on the theory $\mathcal{T}$ under consideration.

We now imagine that I am one of the observers, and that I measure the color of my box and find that it is red. To update the probabilities that I assign to various theories, I need to calculate the first-person probability that I should see red.

Using the method of many repetitions, I would start by calculating the expectation value for the number of observers in the ensemble who will see red. If the ensemble contains N model universes, this expectation value is given by

$$\langle n_{\text{ens}}^R \rangle = N[p_1 N_1^R + p_2 N_2^R] \bar{n}_{\text{red}}, \quad (\text{C3})$$

where $\bar{n}_{\text{red}}$ is the expected number of observers in any single red box, as calculated from $p_{\text{obs}}^R(n)$. That is,

$$\bar{n}_{\text{red}} = \sum_{n=0}^{\infty} n p_{\text{obs}}^R(n). \quad (\text{C4})$$

There will be an analogous expression for $\langle n_{\text{ens}}^B \rangle$, and the expectation value for the total number of observers is just the sum of the two expressions. For the asymptotic ensemble, which corresponds to the limit as $N \rightarrow \infty$, the ratios of the actual numbers of observers who see red or blue to these expectation values should approach 1 (law of large numbers). The first-person probability that I should see red is equal to the fraction of observers in the asymptotic ensemble who see red, which is then given by

Box Type:

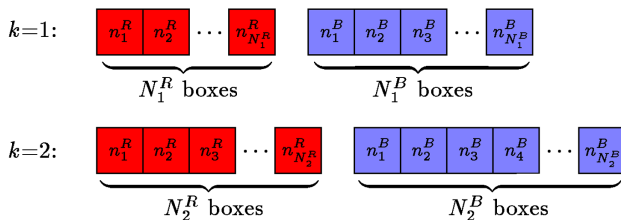


FIG. 1. A diagram to illustrate the box model theories described in Appendix C.

$$p^{(1p)}(\text{red}|T) = \frac{1}{Z} [p_1 N_1^R + p_2 N_2^R] \bar{n}_{\text{red}}, \quad (\text{C5})$$

where

$$Z \equiv [p_1 N_1^R + p_2 N_2^R] \bar{n}_{\text{red}} + [p_1 N_1^B + p_2 N_2^B] \bar{n}_{\text{blue}}. \quad (\text{C6})$$

This answer can be contrasted with several different approaches taken by Hartle and collaborators. In Ref. [20] (hereafter HS2007), Hartle and Srednicki introduced a box model which can be related to our model by taking $p_1 = 1$, $p_2 = 0$, so there is only one box type. The probability that any box contains one or more observers is denoted by p_E , so $p_E = 1 - p_{\text{obs}}^R(0) = 1 - p_{\text{obs}}^B(0)$. The authors consider two theories for this universe: one called *all red* (AR), for which $N_1^B(\text{AR}) = 0$, so all the boxes are red, and one called *some red* (SR), where both $N_1^R(\text{SR})$ and $N_1^B(\text{SR})$ are nonzero. For this special case, the answer we gave in Eq. (C5) reduces to

$$p^{(1p)}(\text{red}|T) = \frac{N_1^R \bar{n}_{\text{red}}}{N_1^R \bar{n}_{\text{red}} + N_1^B \bar{n}_{\text{blue}}}. \quad (\text{C7})$$

In analyzing this model, HS2007 adheres to what we have dubbed the principle of required certainty, as we discussed in Sec. IV B. In this approach, first-person probabilities are irrelevant. If an observer measures the color of his box and finds it to be red, he updates his belief in the relevant theories using only the third-person statement that there exists, somewhere in the universe, at least one red (R) box in which observers exist (E). This probability is given by

$$P(E, R|T) = 1 - (1 - p_E)^{N_1^R(T)}. \quad (\text{C8})$$

If $p_E > 0$ and $N_1^R(T)$ is very large, this probability approaches 1 in both the AR and SR theories, even if the boxes in the SR theory are almost all blue. By contrast, updates based on our Eq. (C5) would under these circumstances strongly favor AR. HS2007 rejects our approach, since in their view there is no reason to believe that the color that we see is in any sense representative:

In such cases it is tempting to reason as follows: In the SR theory with mostly blue cycles we are more likely to observe blue than red. Since we observe red we should reject this theory. ... This, however, is an instance of the selection fallacy and in more familiar situations leads to absurd conclusions ...

We responded to this selection fallacy argument in Sec. IV A.

Srednicki and Hartle [8] (hereafter SH2010) returned to this model a few years later, at this point recognizing that the method proposed in HS2007 is “not effective for

discriminating between theories in a very large universe,” and that therefore alternatives should be sought. So they proposed using first-person probabilities, with an assumed xerographic distribution. Although they consider it to be only one out of many possible choices, they do consider the xerographic distribution in which we are typical of all the observing systems that exist, as would be implied by the PSLI. Nonetheless, their answer does not agree with ours, because they treat the possible nonexistence of observers in a way that makes no sense to us. We all agree that first-person probabilities are the probabilities for “what *we* will measure” (SH2010, p. 1), but they nonetheless attribute a nonzero probability that our measurements will show that no observers exist. Their Eq. (5.4),

$$P^{(1p)}(E|T, \xi^{\text{typD}}) = 1 - (1 - p_E)^{N_R(T)}, \quad (\text{C9})$$

gives a probability less than one to the proposition that at least one observer exists. The first-person likelihood of seeing red, to be used in Bayesian updates, is given by

$$P^{(1p)}(E, R|\xi^{\text{typO}}, T) = \frac{N_R(T)}{N_R(T) + N_B(T)} [1 - (1 - p_E)^N]. \quad (\text{C10})$$

This agrees with our Eq. (C5) except for the factor in brackets, and the factors of $\bar{n}_{\text{red}}$ and $\bar{n}_{\text{blue}}$ that appear in our answer, due to our weighting by observers. According to the answer in SH2010, the three first-person probabilities—that we see red, that we see blue, and that we do not exist—sum to one. The comparison of the answers also shows the impact of our weighting by observers. If $N_1^R = N_1^B$, but every red box contains a million observers and every blue box contains only one observer, then our method will strongly favor our seeing red, while the method of HS2010 will assign equal probabilities.

In Ref. [13] (hereafter HH2015), written somewhat later, Hartle and Hertog introduce a slightly more complicated box model. (A similar model is discussed in Ref. [39].) In this model, p_1 and p_2 are both nonzero, with $N_1^B = 0$ and $N_2^R = 0$. Assuming a uniform xerographic distribution, HH2015 found that the first-person probability for seeing red is given by

$$p^{(1p)}(\text{red}|T) = \frac{p_1 [1 - (1 - p_E)^{N_1^R}]}{p_1 [1 - (1 - p_E)^{N_1^R}] + p_2 [1 - (1 - p_E)^{N_2^B}]}, \quad (\text{C11})$$

with a similar expression for $p^{(1p)}(\text{blue}|T)$. These two probabilities sum to one, as they should, with no allowance for a first-person probability of not existing. This answer again does not agree with our Eq. (C5), but the difference can be traced to a single difference in our method of

calculation. That is, the method of many repetitions assigns a weight to each possible universe U_α proportional to its third-person probability and to the number of observers,

while the HH2015 method weights each possible universe only by its third-person probability, restricted to universes that have at least one observer.

-
- [1] A. H. Guth, Inflationary universe: A possible solution to the horizon and flatness problems, *Phys. Rev. D* **23**, 347 (1981).
 - [2] A. D. Linde, A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems, *Phys. Lett.* **108B**, 389 (1982).
 - [3] A. Albrecht and P. J. Steinhardt, Cosmology for grand unified theories with radiatively induced symmetry breaking, *Phys. Rev. Lett.* **48**, 1220 (1982).
 - [4] A. D. Linde, Chaotic inflation, *Phys. Lett.* **129B**, 177 (1983).
 - [5] A. Vilenkin, Birth of inflationary universes, *Phys. Rev. D* **27**, 2848 (1983).
 - [6] A. D. Linde, Eternal chaotic inflation, *Mod. Phys. Lett. A* **01**, 81 (1986).
 - [7] A. D. Linde, Eternally existing self-reproducing chaotic inflationary universe, *Phys. Lett. B* **175**, 395 (1986).
 - [8] M. Srednicki and J. Hartle, Science in a very large universe, *Phys. Rev. D* **81**, 123524 (2010).
 - [9] J. Perry, The problem of the essential indexical, *Noûs* **13**, 3 (1979).
 - [10] D. Lewis, Attitudes de dicto and de se, *Philos. Rev.* **88**, 513 (1979).
 - [11] N. Bostrom, *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge, New York & London, 2002).
 - [12] S. W. Hawking and T. Hertog, Populating the landscape: A top-down approach, *Phys. Rev. D* **73**, 123527 (2006).
 - [13] J. Hartle and T. Hertog, The observer strikes back, in *The Philosophy of Cosmology*, edited by K. Chamcham, J. Silk, J. D. Barrow, and S. Saunders (Cambridge University Press, Cambridge, England, 2017).
 - [14] A. Aguirre, Making predictions in a multiverse: Conundrums, dangers, coincidences, in *Universe or Multiverse?*, edited by B. Carr (Cambridge University Press, Cambridge, England, 2007).
 - [15] A. Aguirre, S. Gratton, and M. C. Johnson, Hurdles for recent measures in eternal inflation, *Phys. Rev. D* **75**, 123501 (2007).
 - [16] A. H. Guth, Eternal inflation and its implications, *J. Phys. A* **40**, 6811 (2007).
 - [17] B. Freivogel, Making predictions in the multiverse, *Classical Quantum Gravity* **28**, 204007 (2011).
 - [18] C. Smeenk, Predictability crisis in early universe cosmology, *Stud. Hist. Phil. Mod. Phys.* **46**, 122 (2014).
 - [19] Ernan McMullin, Indifference principle and anthropic principle in cosmology, *Studies in history and philosophy of science* **24**, 359 (1993), See especially p. 373.
 - [20] J. B. Hartle and M. Srednicki, Are we typical?, *Phys. Rev. D* **75**, 123523 (2007).
 - [21] D. N. Page, Typicality derived, *Phys. Rev. D* **78**, 023514 (2008).
 - [22] F. Azhar, Prediction and typicality in multiverse cosmology, *Classical Quantum Gravity* **31**, 035005 (2014).
 - [23] F. Azhar, Testing typicality in multiverse cosmology, *Phys. Rev. D* **91**, 103534 (2015).
 - [24] F. Azhar, Spectra of conditionalization and typicality in the multiverse, *Phys. Rev. D* **93**, 043506 (2016).
 - [25] A. Vilenkin, Predictions from quantum cosmology, *Phys. Rev. Lett.* **74**, 846 (1995).
 - [26] J. R. Gott III, Implications of the Copernican principle for our future prospects, *Nature (London)* **363**, 315 (1993).
 - [27] D. N. Page, Typicality defended, [arXiv:0707.4169](https://arxiv.org/abs/0707.4169).
 - [28] J. Garriga and A. Vilenkin, Prediction and explanation in the multiverse, *Phys. Rev. D* **77**, 043526 (2008).
 - [29] J. M. Keynes, *A Treatise on Probability* (MacMillan, London, 1921), <http://www.gutenberg.lib.md.us/3/2/6/2/32625/old/2010-05-31-32625-pdf.pdf>.
 - [30] F. Azhar and J. N. Butterfield, Scientific realism and primordial cosmology, in *The Routledge Handbook on Scientific Realism*, edited by J. Saatsi (Routledge, London, 2018), pp. 304–320.
 - [31] Brandon Carter, Large number coincidences and the anthropic principle in cosmology, in *Confrontation of Cosmological Theories with Observational Data, IAU Symposium No. 63*, edited by M. S. Longair (D. Reidel Publishing Company, Dordrecht & Boston, 1974), pp. 291–298, <https://articles.adsabs.harvard.edu/full/1974IAUS...63..291C>.
 - [32] J. Hartle and T. Hertog, Anthropic bounds on Λ from the no-boundary quantum state, *Phys. Rev. D* **88**, 123516 (2013).
 - [33] A. Aguirre and M. Tegmark, Multiple universes, cosmic coincidences, and other dark matters, *J. Cosmol. Astropart. Phys.* **01** (2005) 003.
 - [34] F. Azhar and N. Linnemann, Rethinking the Anthropic Principle, *Philos. Sci.* **92**, 361 (2025).
 - [35] M. Gell-Mann and James B. Hartle, Quantum mechanics in the light of quantum cosmology, in *Complexity, Entropy, and the Physics of Information: Proceedings Volume in the Santa Fe Institute Studies in the Sciences of Complexity; v. 8*, edited by Wojciech H. Zurek (Westview Press, Boca Raton, 1990).
 - [36] G. E. Moore, *Philosophical Papers* (George Allen and Unwin Ltd, London, 1959) This volume includes the key epistemological papers by Moore: “Proof of an External World,” “Four Forms of Skepticism,” and “Certainty,” <https://ia600101.us.archive.org/15/items/in.ernet.dli.2015.188321/2015.188321.Philosophical-Papers-By-Edward-Moore.pdf>.
 - [37] Thomas Reid, Essays on the intellectual powers of man, in *Reid: Inquiry and Essays*, edited by R. E. Beanblossom and K. Lehrer (Hackett Publishing, Indianapolis, 1983), https://www.google.com/books/edition/Essays_on_the_Intellectual_Powers_of_Man/54pHAAAAIAAJ.

- [38] Robert Nozick, *Philosophical Explanations* (Harvard University Press, Cambridge, MA, 1981).
- [39] J. Hartle and T. Hertog, One bubble to rule them all, *Phys. Rev. D* **95**, 123502 (2017).
- [40] E. T. Jaynes, *Probability Theory: The Logic of Science*, edited by G. L. Bretthorst (Cambridge University Press, Cambridge, England, 2003).
- [41] H. Jeffreys, *Theory of Probability*, 3rd Ed. (Clarendon Press, Oxford, 1983).
- [42] J. Earman, *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory* (MIT Press, MA, 1992).
- [43] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory* (Wiley, Chichester, 1994).
- [44] C. Howson and P. Urbach, *Scientific Reasoning: The Bayesian Approach*, 3rd Ed. (Open Court, IL, 2006).
- [45] P. M. Lee, *Bayesian Statistics: An Introduction* (Wiley, Chichester, UK, 2012).
- [46] J. Willard Gibbs, *Elementary Principles in Statistical Mechanics* (Charles Scribner's Sons, New York, 1902), https://books.google.com/books/download/Elementary_Principles_in_Statistical_Mec.pdf?id=KRMlzl8z6H8C&output=pdf&sig=ACfU3U3RxdiBBVD2iymdT5JEuhOqnPfpTg.
- [47] J. Bernoulli, *The Art of Conjecturing, together with Letter to a Friend on Sets in Court Tennis* (The Johns Hopkins University Press, Baltimore, 2006) [translated with an introduction and notes by E. D. Sylla. Original publication year: 1713].
- [48] P. M. Laplace, *A Philosophical Essay on Probabilities* (Wiley, London, 1902) [translated from the 6th French edition by F. W. Truscott and F. L. Emory. Original publication year: 1814], https://www.google.com/books/edition/A_Philosophical_Essay_on_Probabilities/Df17AQAAAMAJ.
- [49] Bertrand Russell, Review of A Treatise on Probability by John Maynard Keynes, *Math. Gazette* **11**, 119 (1922).
- [50] A. Vilenkin, The principle of mediocrity, *Astron. Geophys.* **52**, 5.33 (2011).
- [51] A. V. Skorokhod, *Basic Principles and Applications of Probability Theory* (Springer, New York, 2005) [Original Russian edition published by Viniti, Moscow 1989].
- [52] G. W. Gibbons and S. W. Hawking, Cosmological event horizons, thermodynamics, and particle creation, *Phys. Rev. D* **15**, 2738 (1977).
- [53] M. J. Rees, *Before the Beginning: Our Universe and Others* (Perseus Books, New York, 1997).
- [54] A. de Simone, A. H. Guth, A. Linde, M. Noorbala, M. P. Salem, and A. Vilenkin, Boltzmann brains and the scale-factor cutoff measure of the multiverse, *Phys. Rev. D* **82**, 063520 (2010).
- [55] R. Bousso, B. Freivogel, and I. S. Yang, Properties of the scale factor measure, *Phys. Rev. D* **79**, 063513 (2009).
- [56] A. D. Linde, D. A. Linde, and A. Mezhlumian, Do we live in the center of the world?, *Phys. Lett. B* **345**, 203 (1995).
- [57] A. H. Guth, Inflation and eternal inflation, *Phys. Rep.* **333**, 555 (2000).
- [58] M. Tegmark, What does inflation really predict? *J. Cosmol. Astropart. Phys.* **04** (2005) 001.
- [59] R. Bousso, B. Freivogel, and I. S. Yang, Boltzmann babies in the proper time measure, *Phys. Rev. D* **77**, 103514 (2008).
- [60] D. N. Page, Is our universe decaying at an astronomical rate?, *Phys. Lett. B* **669**, 197 (2008).
- [61] S. M. Carroll, Why Boltzmann brains are bad, in *Current Controversies in Philosophy of Science*, edited by S. Dasgupta, R. Dotan, and B. Weslake (Routledge, New York, 2021), pp. 7–20, Quotations in our paper were taken from the arXiv version of Carroll's paper, [arXiv:1702.00850](https://arxiv.org/abs/1702.00850).
- [62] A. Albrecht and L. Sorbo, Can the universe afford inflation?, *Phys. Rev. D* **70**, 063528 (2004).
- [63] D. N. Page, Bayes keeps boltzmann brains at bay, *Found. Phys.* **54**, 62 (2024).
- [64] J. R. Gott III, Boltzmann brains—I'd rather see than be one, [arXiv:0802.0233](https://arxiv.org/abs/0802.0233).
- [65] M. A. Krousel-Wood, R. B. Chambers, and P. Muntner, Clinicians' guide to statistics for medical practice and research: Part I, *Ochsner J.* **6**, 68 (2006), See especially p. 76, <https://www.ochsnerjournal.org/content/ochjnl/6/2/68.full-text.pdf>.
- [66] C. Howson, The 'Old Evidence' Problem, *Br. J. Philos. Sci.* **42**, 547 (1991).
- [67] J. Hartle, S. W. Hawking, and T. Hertog, No-boundary measure in the regime of eternal inflation, *Phys. Rev. D* **82**, 063510 (2010).
- [68] A. Elga, Self-locating belief and the sleeping beauty problem, *Analysis* **60**, 143 (2000).
- [69] S. M. Carroll, Preposterous Universe [blog] (2014), comment by Joe Polchinski on July 31, 2014, <http://www.preposterousuniverse.com/blog/2014/07/28/quantum-sleeping-beauty-and-the-multiverse/comment-page-4/#comment-7295910552604295812>.