How to select observers

Robert Garisto
American Physical Society, Ridge, New York 11961, USA

(Received 27 November 2019; accepted 18 August 2020; published 22 September 2020)

A number of problems in physics, mathematics, and philosophy involve observers in given situations which lead to debates about whether observer-specific information should affect the probability for some outcome or hypothesis. Our purpose is *not* to advocate for such observer selection effects but rather to show that any such effects depend greatly on the assumptions made. We focus on the debate about the existence of a "doomsday effect"—whether observer index information should cause one to favor possibilities with fewer observers, which has been argued to have implications for models of cosmology. Our central goal is to reconcile the apparent inconsistencies in the literature by introducing a formalism to lay bare assumptions made and address a key issue that has not been clearly articulated in such problems: whether the observer is selected by *picking from* or *being in* a set of worlds. In the former there generally are observer selection effects, and in the latter there generally are not. This leads us to differentiate what we call *inclusive* from *exclusive* selection and how they relate to the concept of a multiverse. Then we relax the assumption that all observers are equally typical and consider the problem of Boltzmann brains, showing that typicality can play a role in solving the problem. We then stress the need for scale-invariant questions, which causes us to analyze J. Richard Gott's approach to the problem. This all allows us to analyze the doomsday and universal doomsday arguments. We find that there is no doomsday effect, absent a set of assumptions we find somewhat unreasonable. Then we use our formalism to resolve a debate in the philosophy community called the "Sleeping Beauty problem." Finally, we conclude with a heuristic summary, free from equations, and point to possible future directions of this line of research.

DOI: 10.1103/PhysRevResearch.2.033464

I. INTRODUCTION

Physicists usually shun observer-specific information—and for good reason. Our theories are based on invariances, such as those with respect to space and time, and should not depend on who is testing them. Emmy Noether showed that conservation laws are rooted in symmetries [1]. Yet we accept boundary conditions and symmetry breaking because of the constraints of the real world. And sometimes just being an observer can bias our viewpoint. It took millennia for humans to realize that we were not the center of the Universe and that we are atypical collections of matter in being confined to the surface of a habitable planet. Some of the apparent coincidences which seem necessary for life to have evolved may be due to generalizing this notion of us being atypical [2,3]. But our purpose here is to focus on one particular type of observer effect: that probabilities we assign to the selection of an entity may differ if the entity is an observer because the observer has the capacity to self-select. We will see that changing assumptions can completely change these effects, so, at a minimum, anyone invoking them, or decrying them, should carefully lay out all assumptions made.

The quintessential example is the "doomsday argument" [4], about which there is much debate [5–11]. Suppose you assign some prior probability p for case S, that the "world" of which you are a part (and we will define "world" in various ways) will persist only for a short time, with a relatively small number of "people" ever living in that world. The other possibility, L, is that it will persist longer, with more total "people," to which you assign probability $1-p$. But you realize that in your guess for p, you have neglected to take into account any possible observer selection effects (OSEs). The doomsday argument says that you should adjust p upward because the probability is small that you would just happen to live very, very early in the life of a world, and thus you are more likely to live in a short-lived world for which you would be more typical. Is that right? It depends on your assumptions.

Throughout most of the paper, we will be talking about probabilistic situations where there is a set P of "people" (entities capable of being observers, though not always the primary observer in the situation) from which one is selected, and we want to know the probability that the "person" belongs to a subset of P associated with some property, e.g., "born before the year 2100." A key question is whether the "person" self-selects directly from set P (which is generally embedded in enclosing sets such as worlds), which we call a "Be selection" (a Be for short), or whether they are selected in some other way, which we call a "Pick selection" (a Pick for short). In most of our scenarios, the latter entails more than one selection because in order to pick an element of set P one must generally first pick an element of one of the sets that encloses P (e.g., to pick a nut from a set of jars, one must first pick one of the jars). The posterior probabilities for Be and Pick selection differ: OSEs tend to arise in the latter but not the former.

Philosopher Nick Bostrom has written much about the doomsday argument [6,7,9]. He, too, discusses two possible ways an observer could be selected, often using problems of prisoners, which make good toy models because they entail observers confined to specific enclosing sets (cells in cellblocks in prisons). We will assume through most of the paper what he calls the self-sampling assumption (SSA), which just means that you assume you are equally likely to be any member of the set of possible observers you define in your problem, i.e., it is an assumption of *typicality*. He also considers something called the self-indication assumption (SIA), which says you should weight the probability of your existence by the number of people in the world in which you exist [5,8,12]. This is essentially a kludge factor, and why it has rightly been found to be problematic [9,11–13]. In fact, the SIA gives the wrong answer whenever there is a selection from an enclosing set, such as in the warden problem we discuss in Sec. III, or when we take theories to be mutually exclusive, as in Sec. VI. Nevertheless, we will see that the weighting factor associated with the SIA appears naturally with the SSA if we assume observers are Be selected rather than Pick selected.

So there are conflicting and problematic results and apparent misunderstandings in the literature, and much of this is due to there being no universal notation. Our goal in writing this paper is to resolve these issues. Central to doing so is our nested-set notation, which we hope will allow authors to make clear their assumptions on how observers are selected, so readers can judge for themselves whether the assumptions made, and the results they lead to, are reasonable.

The paper is structured as follows. In the next two sections, we consider the selection of observers within "worlds" (prisoners in cellblocks), first via a Be selection and then via a Pick selection, showing how OSEs arise in the latter. In the following two sections, we discuss what happens if we embed the worlds in an enclosing set E, and there is just one Be selection on P (an *inclusive selection*), or an additional Pick on set E (an *exclusive selection*), again with OSEs in the latter. If we take set E to comprise "everything," then we term the inclusive case *the inclusiverse* and the exclusive case *an exclusiverse*. The key difference between them is that in the former we assume that all hypothesized things exist, and in the latter we do not. This leads to a general principle: It is effects of the latter which lead to OSEs. Later we discuss whether it is possible to distinguish these two cases and relate them to the term "multiverse," but our purpose is to lay out how to calculate probabilities given certain assumptions, not to posit the nature of reality. Next we discuss spaces of theories, typicality, and the issue of "freak" observers in cosmology called Boltzmann brains and how our analysis can frame that problem. Then we consider an analysis by J. Richard Gott [14], which lets us phrase the doomsday argument in a scale-invariant way. We are then ready to fully address the doomsday argument and what has been called "universal doomsday." We show that while many sets of assumptions lead to no doomsday effect, it is possible to come up with a set of assumptions, however implausible, which leads to one. Then we address a related problem in philosophy called the "Sleeping Beauty problem." Finally, we summarize our results and point to future directions.

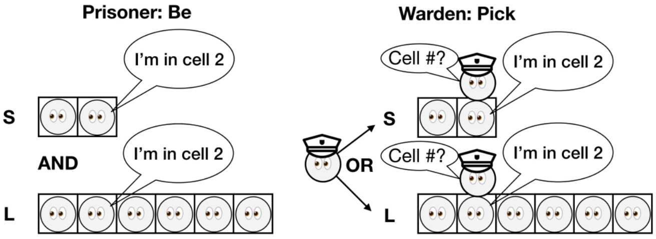

FIG. 1. Why the warden problem (with a Pick selection) leads to an OSE and the Prisoner problem (with a Be selection) does not: There are two cellblocks, S and L. Prisoners all simply ask themselves, "Which cellblock am I in?" and then observe their cell number to answer. There are more prisoners in the L cellblock to ask the question, which cancels the rank factor that a smaller faction of prisoners are in the first two cells in L than in S, so those in cell 2 are equally likely to be in either cellblock. The warden first must Pick a cellblock at random and then select a cell at random within that cellblock. If the selected prisoner is in cell 2, then it is more likely that the warden picked the S than the L cellblock because the number of prisoners per cellblock did *not* affect the odds that she picked that cellblock, and so the rank factor is not canceled as it was in the Be case.

In an effort to make the paper readable to the wider world, the summary is comprehensive of our results without equations. We have also put details of our nested-set notation and a table that summarizes our results into the Appendix. And in the body of the paper, we spell out many intermediate steps in our equations since some interested in the results here may include those less familiar with working out such steps.

II. TO BE: PRISONER PROBLEM

Imagine you are a prisoner and have the following information: The prison you are in has two types of cellblocks, small (S) and large (L), which contain $\bar{n}_S$ and $\bar{n}_L$ cells per cellblock, respectively. You want to estimate the probability that you are in an S cellblock.

Before we dive into a lot of notation, let us consider a simple numerical example, where there is one cellblock of each type, with $\bar{n}_S = 2$ and $\bar{n}_L = 6$ (see the left side of Fig. 1). You do not know your cell number at the outset, so you could be in either the S or L cellblock. Now you look at your door and learn your cell number. If it is greater than 2, then you know you are in the L cellblock. Let us assume that it is cell number 2, so you could be in either cellblock. What is the probability that you are in the S cellblock? Well, there are exactly two cells with cell number 2, one in each cellblock. And you have no reason to favor one over the other, so you should assign a probability of 1/2 for being in the S cellblock. Note that this is equal to the probability of picking the S cellblock at random. In other words, the posterior probability for being in cellblock S, given the cell-number datum that you could be in either cellblock, is the same as the prior probability of randomly picking cellblock S—there is no observer selection effect.

Now let us formalize the problem for a general number of prisoners and cellblocks. You assign labels N_S and N_L to the number of cellblocks of each type, but all you know is that there is at least one cellblock (since you are in one), i.e., $N \equiv N_S + N_L \geqslant 1$. You also know that the prison is full and that each prisoner was assigned a random cell in the prison, with exactly one prisoner per cell. Let the ratio of cells in L and S cellblocks be

$$\rho \equiv \bar{n}_L/\bar{n}_S, \tag{1}$$

which is by assumption greater than 1. The bar just indicates we have normalized to the number of cellblocks. The total number of prisoners in all cellblocks of type $J = S$ or L is n_J, which is equal to the number of cells per cellblock of that type times the number of cellblocks of that type:

$$n_J = \bar{n}_J N_J. \tag{2}$$

Let us call the set of prisoners P (for "person," the set that will usually hold our observers) and the set of cellblocks W (for "world," since this problem is an analog to one of observers in worlds). W_S and W_L are the subsets of W containing all S and L cellblocks, respectively. Since there are only two types of cellblocks, the set W is the union of them: $W = W_S \cup W_L$. You assign some prior probability for what the fraction of small cellblocks $P(W_S) = N_S/N$ might be [we assume that the probability of picking any given cellblock is simply $1/N$, and these $P(W_S)$ and $P(W_L)$ are *fixed* inputs—we will explore varying ratios of them in Sec. IV]. Note that P is nested within W, i.e., every element of P (a prisoner) is associated with a particular element of W (a cellblock). The compound set PW_S contains the set of S cellblocks, and the set of prisoners in P who are in S cellblocks (see the Appendix for details on notation).

We will assume the SSA [7],

> SSA: One should reason as if one is a random sample from the set of all observers in one's reference class.

This is simply assuming *typicality*, that the probability of you being in a subset of a larger set is simply equal to the fraction of observers of the reference class (which we call set P) who are in that subset. For example, the probability to Be in subset P_x of set P is just $P(P_x|P) = n_x/n$.

You learn one datum, your cell number. Divide the datum into two categories: d if your cell number is $\leqslant \bar{n}_S$, and $\neg d$ if it is $> \bar{n}_S$. The corresponding subsets of P are P_d and $P_{\neg d}$ ($P = P_d \cup P_{\neg d}$). If your datum is $\neg d$, then you know for sure that you are in an L cellblock (because your cell number is greater than $\bar{n}_S$). The case of interest is when the datum is d, where you could still be in either type of cellblock. The question we want to answer in the prisoner problem is

> What is the posterior probability that a prisoner is in an S cellblock, given that they match datum d?

For convenience we define the number of people matching datum d to be $m \equiv n_d$ and the number of people matching datum d within a cellblock type J to be $m_J \equiv n_{d,J}$, where $J = L$ or S. All observers with cell numbers $\leqslant \bar{n}_S$ match datum d, so the number of people per cellblock matching datum d is $\bar{m} = \bar{n}_S$, and this holds for both S and L cellblocks, so

$$\bar{m} = \bar{m}_S = \bar{m}_L = \bar{n}_S. \tag{3}$$

We want to calculate the probability of you being in a cellblock type S (i.e., in subset PW_S of PW) given the datum, d, that you are in a low cell number (i.e., in subset P_dW of PW), which we write at the conditional probability $P(PW_S|P_dW)$. We will calculate this using Bayes's law, so we need the likelihood of matching the datum given that we are in a cellblock type S,

$$P(P_dW|PW_S) = \frac{m_S}{n_S} = \frac{\bar{m}_S}{\bar{n}_S} = 1, \tag{4}$$

and the probability [15] to Be in cellblock type S,

$$P(PW_S) = \frac{n_S}{n} = \frac{\bar{n}_S N_S}{\bar{n} N} = \frac{\bar{n}_S}{\bar{n}} P(W_S), \tag{5}$$

where $P(W_S)$ is the prior probability to Pick a cellblock of type S (which, assuming random typical selection, is equal to our prior value for fraction of worlds, N_S/N).

We need to pause here because Eq. (5), despite its simplicity, is the key to most of our results. We have simply taken the SSA at face value. Since the prisoner has an equal chance of being in any cell, the probability to Be in the subset of prisoners in S cellblocks is simply the fraction of prisoners in such cellblocks, n_S/n, which as we show in Eq. (5) is equal to the prior $P(W_S)$ weighted by the average number of prisoners $\bar{n}_S$ per cellblock of this type. We should at this point note the competing assumption, the self-indication assumption [7]:

> SIA: 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.

This *does* giving the weighting factor seen in Eq. (5), but it is a kludge factor because it gives that factor regardless of how the observer is selected, which, as we shall see, is inappropriate whenever the first selection is from a set that encloses the observer. (Some may take the SIA to mean that this weighting factor should be applied *where appropriate*—not in any situation where you are an observer. If so, then a way to think of our formalism is that it shows when that weighting factor is appropriate.) In contrast, we derived the weighting factor in Eq. (5) simply using typicality (the SSA) and the recognition that we are selecting the observer directly from set P. The effect from how the observer is selected is made transparent by our nested-set notation. There are a number of places in the literature which simply refer to "$P(S)$" and let it equal to the prior probability for picking a world type S, when to be a prisoner requires $P(PW_S)$ with its weighting factor $\bar{n}_S/\bar{n}$. Failing to include this factor leads to erroneous support for a doomsday effect.

Here is another way to understand this weighting factor. If you use the information that you are an observer in a random cell before also applying datum d, then you are more likely to be in an L cellblock than your prior for the fraction of L cellblocks would suggest. For example, if $P(W_S) = P(W_L) = 1/2$, then there are ρ times as many observers in L cellblocks as in S cellblocks, and so the probability of being in a cellblock type L (before knowing d) is ρ times that of being in a cellblock type S. This factor of $\bar{n}_S$ in Eq. (5) will exactly cancel a factor of $1/\bar{n}_S$ in the likelihood Eq. (4). [As we shall see in the next section, this factor is absent if there is a Pick on the world set W. We should also note that by our formulation of the problem we are assuming that the prisoner could be in both types of cellblocks. We will later consider the cases where there are mutually exclusive "universes" (Sec. V) and hypotheses (Sec. VI A).]

So the posterior probability of you being in a cellblock type S given datum d is given by Bayes's law,

$$\begin{aligned} P(PW_S|P_dW) &= \frac{P(P_dW|PW_S)P(PW_S)}{P(P_dW)} \\ &= \frac{P(P_dW|PW_S)P(PW_S)}{\sum_J P(P_dW|PW_J)P(PW_J)} \\ &= \frac{\frac{\bar{m}_S}{\bar{n}_S}\frac{\bar{n}_S}{\bar{n}}P(W_S)}{\sum_J \frac{\bar{m}_J}{\bar{n}_J}\frac{\bar{n}_J}{\bar{n}}P(W_J)} = \frac{\bar{m}_S}{\bar{m}}P(W_S) = P(W_S), \end{aligned} \tag{6}$$

where $J = S$ or L, $\sum_J \bar{m}_J P(W_J) = \bar{m}$, and $\bar{m} = \bar{m}_S = \bar{m}_L$. The right-hand side is the prior probability for picking a cellblock of type S—i.e., the probability before we have any observer information at all. As we noted before, the prior here to *pick* a world type S, $P(W_S)$, is a fixed value N_S/N, not updated by the datum. What is updated is our posterior probability to *be* in such a world. [Note that we can also write this more compactly using the shorthand notation described in Appendix, see Eq. (A16).] We can express the fact that there is no net observer selection effect by comparing the ratio of probabilities after (R_P) and before (R_W) observer information:

$$\begin{aligned} R_P &\equiv \frac{P(PW_L|P_dW)}{P(PW_S|P_dW)} = \frac{P(W_L)}{P(W_S)} \\ R_W &\equiv \frac{P(W_L)}{P(W_S)}, \quad R_{P/W} \equiv \frac{R_P}{R_W} = 1. \end{aligned} \tag{7}$$

In the prisoner problem, using observer information, which includes the effect of you being in a small cellblock, as well as the likelihood of you being in a low-numbered cell, you obtain the prior probability to Pick a cellblock type S. In short, in the prisoner problem, when your datum is d, there is no net observer selection effect ($R_{P/W} = 1$).

III. TO PICK: URN AND WARDEN PROBLEMS

Now let W be a set of urns, and P a set of ping-pong balls in them. Each urn contains either a large ($\bar{n}_L$) or small ($\bar{n}_S$) number of consecutively numbered balls—defining subsets W_L and W_S. You pick an urn at random and a ball at random from the urn. Before picking the ball, in fact before you actually picked an urn, you had a prior probability that the urn you picked is of type S, $P(W_S)$. After seeing the ball, *what is the posterior probability that the urn is type S?*, i.e.,

> What is the posterior probability that you pick an S urn and then a random ball in it, given that the ball you pick matches datum d?

Again, let us first use a numerical example to build intuition. Suppose there are two urns, one S and one L, with $\bar{n}_S = 2$ and $\bar{n}_L = 6$. You pick a random urn and then pick a random ball from it (we shall see that this is the same as the warden problem on the right side of Fig. 1). If the ball number is greater than 2, then the urn you picked was the L urn. So let us assume the same datum as before, that it is ball number 2, which corresponds to datum d. Now, before you knew the ball number, there was an equal chance that you picked the S or L urn. But once you have datum d, your posterior probability of having picked the S urn has greatly increased because all the balls in the S urn match d, whereas that is true only of 1/3 of the balls in the L urn. In fact, while your prior for picking the urns was equal, your posterior probability of picking the S urn is 3 times that of picking the L urn (3/4 vs. 1/4). Though the setup seems the same as in Sec. II, the fact that there was an initial selection of the urn makes all the difference.

Let us now go into the details. Obviously, if the ball's number is $>\bar{n}_S$, then you will know that it is an L urn and that posterior probability is 0. So let us assume that the datum d you get is that the ball's number is $\leqslant\bar{n}_S$. It is tempting to say that the situation is identical to the prisoner example and that we learn nothing about the urn. After all, both kinds of urns have the same number of balls with number less than $\bar{n}_S$. But the situation is different because *in order to pick the ball from the urn, we first had to pick the urn*. To denote that selection, we put a Pick sign "${}^{\mid}$" between sets (see Appendix for more on our set notation). So to Pick any ball from any urn is $P{}^{\mid}W$, and to Pick a ball matching datum d from an S urn is $P_d{}^{\mid}W_S$. Thus what we seek is $P(P{}^{\mid}W_S|P_d{}^{\mid}W)$, the probability of picking a ball from an S urn given that we picked a ball matching datum d.

The probability of matching datum d given the urn is type S is exactly the same as Eq. (4) because if it is *given* that you picked an S urn, then the Pick has no effect on the likelihood, it is "neutered" (see Appendix) and we put a slash through the Pick sign to indicate this:

$$P(P_d{}^{\nmid}W|P{}^{\nmid}W_S) = P(P_dW|PW_S) = \frac{m_S}{n_S} = \frac{\bar{m}_S}{\bar{n}_S} = 1, \tag{8}$$

and with $\bar{m}_S = \bar{n}_S$ (grouping all the balls matching datum d together), $P(P_d{}^{\mid}W|P{}^{\mid}W_S) = 1$. However, the probability of picking a ball from an urn of type S is *not* the same as Eq. (5) because there is no weighting for the number of balls. The probability of picking an S urn and then picking a ball from it is same as the prior probability for picking an S urn,

$$P(P{}^{\mid}W_S) = P(W_S). \tag{9}$$

Because of this, there is no factor of $\bar{n}_S$ in the numerator to balance the $1/\bar{n}_S$ rank factor in the likelihood, so Bayes's law does not just return the prior as it did in the Be case in Eq. (6):

$$\begin{aligned} P(P{}^{\mid}W_S|P_d{}^{\mid}W) &= \frac{P(P_d{}^{\nmid}W|P{}^{\nmid}W_S)P(P{}^{\mid}W_S)}{\sum_J P(P_d{}^{\nmid}W|P{}^{\nmid}W_J)P(P{}^{\mid}W_J)} \\ &= \frac{\frac{\bar{m}_S}{\bar{n}_S}P(W_S)}{\sum_J \frac{\bar{m}_J}{\bar{n}_J}P(W_J)} = \frac{P(W_S)}{\sum_J \frac{\bar{n}_S}{\bar{n}_J}P(W_J)} \\ &= \frac{P(W_S)}{P(W_S) + \frac{1}{\rho}P(W_L)}. \end{aligned} \tag{10}$$

[For shorthand notation, see Eq. (A17).] For $P(W_L)/\rho$ small, this goes to 1.

The posterior probability for L given d is

$$P(P^{\shortmid}W_L|P_d{}^{\shortmid}W) = \frac{\frac{1}{\rho}P(W_L)}{P(W_S)+\frac{1}{\rho}P(W_L)}, \tag{11}$$

which, for equal priors, goes to $1/\rho$ for $P(W_L)/\rho$ small. As in Sec. II, the prior here is a fixed input N_L/N that is unchanged by the datum. Our posterior is the probability of the *urn that we picked* to be type L. To see how data can update a multivalued prior with Pick selection, see Secs. V and X.

The ratios for P and W become

$$R_{P^{\shortmid}} \equiv \frac{P(P^{\shortmid}W_L|P_d{}^{\shortmid}W)}{P(P^{\shortmid}W_S|P_d{}^{\shortmid}W)} = \frac{1}{\rho}\frac{P(W_L)}{P(W_S)}$$

$$R_W \equiv \frac{P(W_L)}{P(W_S)}, \quad R_{P^{\shortmid}/W} \equiv \frac{R_{P^{\shortmid}}}{R_W} = \frac{1}{\rho}. \tag{12}$$

There is thus a very strong selection effect when one has to first Pick the urn ($R_{P^{\shortmid}/W} = 1/\rho$).

Of course balls are not people, so it is tempting to think that it is the nature of the elements of set P that causes the difference with the prisoner problem. To counter that, consider what we call the warden problem, where P is again a set of prisoners in cellblocks W. But this time, instead of the prisoner just *being* the observer within a cellblock, a warden selects a prisoner by first *picking* a random cellblock and then picking a random prisoner within the cellblock, all without noting which type of cellblock she has picked. So the question in the warden problem is

> What is the posterior probability that a warden picks an S cellblock and then a random prisoner in it, given that the prisoner they pick matches datum d?

Then all follows exactly as in the urn problem, and the posterior probability we seek is $P(P^{\shortmid}W_S|P_d{}^{\shortmid}W)$. The warden has a prior probability $P(W_S)$ for having picked a cellblock type S, the likelihood that she gets datum d given that she picked a cellblock type S is one (i.e., $P(P_d{}^{\shortmid}W|P^{\shortmid}W_S) = 1$), and, by Bayes's law, her posterior probability given datum d is given by Eq. (10), with a large selection effect, $R_{P^{\shortmid}/W} = 1/\rho$.

The reason the warden problem differs from the prisoner problem is that the warden has to first Pick a cellblock, whereas the prisoner is there without needing to be picked by anyone else. See Fig. 1. (It may help your intuition to imagine $\bar{n}_L$ huge, say, 2000 so $\rho = 1000$. The prisoner problem is unchanged since if you satisfy d, then you are still in cell 1 or 2 of your cellblock, but in the warden problem she is certain to pick cell 1 or 2 if she picks the S cellblock but there is only one chance in 1000 that she she will do that in the L cellblock.)

We note that if we try to use the SIA in this problem, we will get the wrong answer. If you are a prisoner and a warden picks your cell at random after having picked your cellblock at random, and you learn you match datum d, then you should conclude that you are likely in an S cellblock. But the SIA would have you weight your prior probability to be in a given cellblock by the number of cells, as in Eq. (5), falsely leading you to conclude that there is no OSE, whereas typicality (the SSA) gives you the correct unweighted prior of Eq. (9).

Just to highlight further, it is the Pick on the nesting set W that causes a change in the posterior probability. Consider the warden cafeteria problem, where all the prisoners are in a cafeteria, and the warden Picks a prisoner at random. If that prisoner is from a cell number $\leqslant \bar{n}_S$, then what is the probability that they came from an S cellblock? Now the selection is directly from set P, or equivalently, from inside of the nested set PW, so that the posterior probability is $P(PW_S|P_dW)$, just as in the Be case—there is no observer selection effect in the warden cafeteria problem. A Pick directly from the observer set is the same as a Be on that set (see Appendix). What causes a change in the posterior probability is a Pick on a set in which P is nested, such as W.

IV. INCLUSIVE SELECTION

However many nested sets we have, there are two possibilities: Either there is just a selection on the innermost set (a Be, unless there is a way to directly Pick from it as in the warden cafeteria problem), which we call *inclusive* selection, or there is also at least one selection on one of the enclosing sets [a Pick in all of our examples because we do not consider any sets enclosed by (to the left of) P], which we call *exclusive* selection. The selection in the prisoner problem is inclusive and in the warden problem it is exclusive.

Suppose we have a larger enclosing set, E, in which P and W are nested. For the prisoner and warden problems, this could be the set of all prisons, each of which has their own small-to-large cellblock ratio. We can even take E to encompass everything that we deem possible—such as a set of universes in all possible configurations. Then we define two possibilities for the reality:

> The *inclusiverse*: All things we deem possible are realized.
>
> An *exclusiverse*: Only some of the things we deem possible are realized.

The key question is whether all things to which we assign a nonzero probability actually occur (inclusive selection), or there are some mutually exclusive possibilities (exclusive selection). Perhaps a quantum example is useful. If one assumes that quantum theory is unitary and all pieces of the wave function with nonzero amplitude are realized, so that Schrödinger's cat is both alive and dead (as in the many-worlds case), then that is inclusive selection. If one assumes that the wave function collapses to a specific eigenvalue, so that Schrödinger's cat is alive or dead, not both, then that is an exclusive selection. In the rest of this section we study inclusive selection, though not its implications for reality.

Let us consider inclusive selection for the prisoner problem but with a much more modest set, where E is the set of all prisons we consider and the only selection is the self-selection of the prisoner. If we think that there are exactly two types of prisons, say, with all S cellblocks or all L cellblocks, then the key to inclusiveness is that we calculate probabilities under the assumption that both types of prisons exist—there is no Pick on the selection of E needed. We explicitly show the sum over subsets of E, e, so when we do the same calculation for the exclusive case, the difference will be apparent. For simplicity we will assume that the number of prisoners for any $J = S$ or L cellblock is the same across all prisons, so $\bar{n}_{J,e} = \bar{n}_J$, and similarly we assume the number of prisoners per cellblock matching datum d is the same, $\bar{m}_{J,e} = \bar{m}_J$. The subsets E_e

differ only in their fractions of S and L worlds. The likelihood for the inclusive case comes out the same as in the Be case, Eq. (4):

$$\begin{aligned} &P(P_dWE|PW_SE)\\ &\quad=\sum_e P(P_dWE_e|PW_SE_e)P(PW_SE_e|PW_SE)\\ &\quad=\frac{\bar{m}_S}{\bar{n}_S}\sum_e P(PW_SE_e|PW_SE)=\frac{\bar{m}_S}{\bar{n}_S}=1. \end{aligned}\tag{13}$$

There is no e dependence in the first term, since we assumed that $\bar{n}_S$ and $\bar{m}_S$ do not depend on e. The prior to Be in cellblock type S with inclusive selection of E is

$$\begin{aligned} P(PW_SE)&=\sum_e P(PW_SE_e|PWE_e)P(PWE_e)\\ &=\sum_e \frac{\bar{n}_{S,e}}{\bar{n}_{,e}}P(W_SE_e|WE_e)\frac{\bar{n}_{,e}}{\bar{n}}P(WE_e)\\ &=\frac{\bar{n}_S}{\bar{n}}\sum_e P(W_SE_e|WE_e)P(WE_e)=\frac{\bar{n}_S}{\bar{n}}P(W_SE), \end{aligned}\tag{14}$$

which is the same as Eq. (5), just the prior probability of picking a world of type S weighted by the number of observers per world type S. Note that a factor of $1/\bar{n}_{,e}$ and $\bar{n}_{,e}$ cancel here. Therefore, the posterior probability of you being in a cellblock type S given datum d with an inclusive selection of E is the same as Eq. (6),

$$\begin{aligned} P(PW_SE|P_dWE)&=\frac{P(P_dWE|PW_SE)P(PW_SE)}{\sum_J P(P_dWE|PW_JE)P(PW_JE)}\\ &=\frac{\frac{\bar{m}_S}{\bar{n}_S}\frac{\bar{n}_S}{\bar{n}}P(W_SE)}{\sum_J\frac{\bar{m}_J}{\bar{n}_J}\frac{\bar{n}_J}{\bar{n}}P(W_JE)}=\frac{\bar{m}_S}{\bar{m}}P(W_SE)\\ &=P(W_SE), \end{aligned}\tag{15}$$

just the prior probability of picking a world of type S, and we again get $R^E_{P/W}=1$ as in Eq. (7). There is no net observer selection effect for the prisoner problem in the inclusive case ($R^E_{P/W}=1$). Generalizing, if we are considering a problem where observers are selected only by being, and there is no other selection—all allowed possibilities are realized, as in the inclusiverse—then there is no OSE.

V. EXCLUSIVE SELECTION

Let us analyze the prisoner problem with exclusive selection. The key difference from the inclusive case is that we must Pick a subset E_e: Although we posit that there are multiple possibilities E_e, only one of them is actually realized. As we said in the previous section, if E is the set of everything possible, and we take reality to correspond to a smaller subset, then we live in an exclusiverse. But we will focus on a more mundane set: For the prisoner problem, those subsets of E are prisons.

The defining characteristic of these subsets E_e is the fraction of worlds of type S they contain, which we define as y. So the probability of picking an S world,

$$y\equiv P(W_SE_e|WE_e),\tag{16}$$

and a world of type L, $1-y=P(W_LE_e|WE_e)$, is the same for all elements of a given E_e. That is, E_e is completely specified by its y—in fact we will simply label these subsets by y. Again we assume for simplicity that the number of prisoners per type of world is independent of e: $\bar{n}_{J,e}=\bar{n}_J$ and $\bar{m}_{J,e}=\bar{m}_J$. But note that the average number of prisoners per cellblock in a given prison, $\bar{n}_{,e}$ varies from prison to prison:

$$\begin{aligned} \bar{n}_{,e}&=\bar{n}_SP(W_{S,e})+\bar{n}_LP(W_{L,e})\\ &\equiv\bar{n}_y=\bar{n}_S[y+\rho(1-y)]. \end{aligned}\tag{17}$$

The likelihood in the exclusive case is the same as in inclusive case Eq. (13) because the Pick of subset E_e on the first term in the sum is neutered:

$$\begin{aligned} &P(P_dW{}^{\mid}E|PW_S{}^{\mid}E)\\ &\quad=\sum_e P(P_dW{}^{\dagger}E_e|PW_S{}^{\dagger}E_e)P(PW_S{}^{\mid}E_e|PW_S{}^{\mid}E)\\ &\quad=\frac{\bar{m}_S}{\bar{n}_S}\sum_e P(PW_S{}^{\mid}E_e|PW_S{}^{\mid}E)=\frac{\bar{m}_S}{\bar{n}_S}=1. \end{aligned}\tag{18}$$

However, the prior is different because now we have to first Pick a subset E_e, and there is not a $\bar{n}_{,e}$ to cancel the $1/\bar{n}_{,e}$ as there was in Eq. (14),

$$\begin{aligned} P(PW_S{}^{\mid}E)&=\sum_e P(PW_S{}^{\dagger}E_e|PW{}^{\dagger}E_e)P(PW{}^{\mid}E_e)\\ &=\sum_e\frac{\bar{n}_{S,e}}{\bar{n}_{,e}}P(W_S{}^{\dagger}E_e|W{}^{\dagger}E_e)P(E_e)\\ &=\sum_y\frac{y}{y+\rho(1-y)}P({}^{\mid}y). \end{aligned}\tag{19}$$

For the last line, we have assumed again $\bar{n}_{S,e}=\bar{n}_S$, relabeled the subsets E_e by y, and used the definitions for y in Eq. (16) and $\bar{n}_{,e}$ in Eq. (17). The sum covers all values of y from 0 to 1 with nonzero $P({}^{\mid}y)$, which is the probability of picking an ensemble element of type y [it is shorthand for $P(PW{}^{\mid}E_y)$—see Eqs. (A13)–(A18)]. (Note that as with the warden problem, the SIA gives the wrong answer here because $P(E_e)$ should *not* be weighted by $\bar{n}_{,e}$ in Eq. (19) since we are first Picking subsets of E.) Similarly, for L,

$$\begin{aligned} P(P_dW{}^{\mid}E|PW_L{}^{\mid}E)&=\frac{\bar{m}_L}{\bar{n}_L},\\ P(PW_L{}^{\mid}E)&=\frac{\bar{n}_L}{\bar{n}_S}\sum_y\frac{1-y}{y+\rho(1-y)}P({}^{\mid}y). \end{aligned}\tag{20}$$

Let us use Bayes's law again to obtain the posterior probability of you being in a cellblock type S or L given datum d in the exclusive case, which has the same form as the inclusive case Eq. (15) except with Picks on E, which we obtain from Eqs. (18)–(20):

$$\begin{aligned} &P(PW_S{}^{\mid}E|P_dW{}^{\mid}E)\\ &\quad=\frac{P(P_dW{}^{\mid}E|PW_S{}^{\mid}E)P(PW_S{}^{\mid}E)}{\sum_J P(P_dW{}^{\mid}E|PW_J{}^{\mid}E)P(PW_J{}^{\mid}E)}\\ &\quad=\frac{\sum_y\frac{y}{y+\rho(1-y)}P({}^{\mid}y)}{\sum_y\frac{y+\frac{\bar{m}_L}{\bar{m}_S}(1-y)}{y+\rho(1-y)}P({}^{\mid}y)}=\frac{\sum_y\frac{y}{\rho-(\rho-1)y}P({}^{\mid}y)}{\sum_y\frac{1}{\rho-(\rho-1)y}P({}^{\mid}y)}, \end{aligned}\tag{21}$$

$$\begin{aligned} &P(PW_L{}^{|}E|P_dW{}^{|}E) \\ &\quad = \frac{P(P_dW{}^{|}E|PW_L{}^{|}E)P(PW_L{}^{|}E)}{\sum_J P(P_dW{}^{|}E|PW_J{}^{|}E)P(PW_J{}^{|}E)} \\ &\quad = \frac{\frac{\bar{m}_L}{\bar{m}_S}\sum_y \frac{1-y}{y+\rho(1-y)}P({}^{|}y)}{\sum_y \frac{y+\frac{\bar{m}_L}{\bar{m}_S}(1-y)}{y+\rho(1-y)}P({}^{|}y)} = \frac{\sum_y \frac{1-y}{\rho-(\rho-1)y}P({}^{|}y)}{\sum_y \frac{1}{\rho-(\rho-1)y}P({}^{|}y)}, \end{aligned} \tag{22}$$

where we use $\bar{m}_S = \bar{m}_L$ of Eq. (3) and we rewrote the denominators to collect the y dependence. We are again interested in the ratio of L to S posterior probabilities,

$$R_P^{|E} \equiv \frac{P(PW_L{}^{|}E|P_dW{}^{|}E)}{P(PW_S{}^{|}E|P_dW{}^{|}E)} = \frac{\sum_y \frac{1-y}{\rho-(\rho-1)y}P({}^{|}y)}{\sum_y \frac{y}{\rho-(\rho-1)y}P({}^{|}y)}. \tag{23}$$

We want to normalize this to

$$\begin{aligned} R_W^{|E} &\equiv \frac{P(W_L{}^{|}E)}{P(W_S{}^{|}E)} \\ &= \frac{\sum_e P(W_L{}^{\dagger}E_e|W{}^{\dagger}E_e)P(W{}^{|}E_e)}{\sum_e P(W_S{}^{\dagger}E_e|W{}^{\dagger}E_e)P(W{}^{|}E_e)} = \frac{\sum_y (1-y)P({}^{|}y)}{\sum_y yP({}^{|}y)}. \end{aligned} \tag{24}$$

We can see immediately that if there is only one value Y for which $P({}^{|}y = Y)$ is nonzero, then both $R_P^{|E}$ and $R_W^{|E}$ are equal to $(1-Y)/Y$ and their ratio, $R_{P/W}^{|E}$ is 1—no observer selection effect. That is because that is really the inclusive case—while there is a Pick on E, it is neutered, and all of the values (i.e., the one value) are realized. So for the exclusive case, there needs to be more than one allowed value of y.

So let us explore different assumptions for the function $P({}^{|}y)$, which, to remind you, is our prior probability for elements of E with S-world fraction y. For simplicity, let us define the probability density,

$$p({}^{|}y) \equiv P({}^{|}[y, y+dy])/dy, \tag{25}$$

where now y is not a set of discrete values but all real numbers in [0,1]. We can then write the sums in Eqs. (23) and (24) as integrals:

$$R_P^{|E} = \frac{\int_0^1 dy \frac{1-y}{\rho-(\rho-1)y}p({}^{|}y)}{\int_0^1 dy \frac{y}{\rho-(\rho-1)y}p({}^{|}y)}, \tag{26}$$

$$R_W^{|E} = \frac{\int_0^1 dy(1-y)p({}^{|}y)}{\int_0^1 dy\, yp({}^{|}y)}. \tag{27}$$

A. Near a single point

Let us first explore the case where we take y to have a nonzero probability near a single point Y, in particular that $p({}^{|}y)$ is constant over the range $Y-\sigma$ to $Y+\sigma$, where of course σ is no larger than Y or $1-Y$ so that the points are on the range 0 to 1:

$$p({}^{|}y)_{\text{near}} = \frac{1}{2\sigma}\{\Theta[y-(Y-\sigma)] - \Theta[y-(Y+\sigma)]\}. \tag{28}$$

[$\Theta(x)$ is the step function, equal to 0 for $x < 0$ and 1 for $x \geqslant 1$.] Plugging this into Eq. (27), for the prior ratio probabilities or picking L worlds to S worlds, we get

$$R_W^{|E} = \frac{\left[y - \frac{1}{2}y^2\right]_{Y-\sigma}^{Y+\sigma}}{\left[\frac{1}{2}y^2\right]_{Y-\sigma}^{Y+\sigma}} = \frac{1-Y}{Y}, \tag{29}$$

just as we obtained for a single point. (This is true because the integrand in the numerator and denominator of $R_W^{|E}$ are linear in y.) The expression for $R_P^{|E}$ is more complicated because of the denominator of the integrands. In the limit of $\sigma \to 0$, $R_P^{|E}$ is

$$R_P^{|E} \simeq \frac{1-Y}{Y}\left\{1 - \frac{1}{3}\sigma^2 \frac{\rho-1}{[\rho(1-Y)+Y]Y(1-Y)}\right\}, \tag{30}$$

and thus their ratio is

$$R_{P/W}^{|E} \simeq 1 - \frac{1}{3}\sigma^2 \frac{\rho-1}{[\rho(1-Y)+Y]Y(1-Y)}. \tag{31}$$

Thus if $p({}^{|}y)$ is nonzero within $\pm\sigma$ of a single point Y, then there is a small observer selection effect of order σ^2. In the limit that $\rho \to \infty$ [actually one must be careful when Y is near 1, so really we take $\rho(1-Y) \to \infty$],

$$R_{P/W}^{|E} \to 1 - \frac{1}{3}\sigma^2 \frac{1}{Y(1-Y)^2}. \tag{32}$$

So the closer we restrict our prior to be near a single point Y, the less $R_{P/W}^{|E}$ differs from 1, and this behavior is independent of ρ.

B. Flat prior

The simplest prior assumption is that every value of y is equally likely,

$$p({}^{|}y)_{\text{flat}} = 1. \tag{33}$$

From Eq. (27) this gives equal probability of picking S and L worlds,

$$R_W^{|E} = \frac{\left[y - \frac{1}{2}y^2\right]_0^1}{\left[\frac{1}{2}y^2\right]_0^1} = 1, \tag{34}$$

which we also could have obtained from Eq. (29) for $Y = \sigma = 1/2$. The posterior ratio of being in L and S worlds, $R_P^{|E}$, is thus unchanged when normalized to $R_W^{|E} = 1$, and for their ratio we obtain,

$$R_{P/W}^{|E} = \frac{1-(\ln\rho+1)/\rho}{\ln\rho - 1 + 1/\rho} \to \frac{1}{\ln\rho - 1}, \tag{35}$$

where we take the limit of $\rho \to \infty$ (this approximation is good only for $\rho \gtrsim 100$). So for a flat prior, we get an observer selection effect which goes roughly as $1/\ln\rho$, in between the original prisoner problem, $R_{P/W} = 1 = \rho^0$, and warden problem, $R_{P{}^{|}/W} = \rho^{-1}$.

If the point of choosing a flat prior is to minimize the effect of assumptions on the outcome, then it might make more sense to use inclusive selection instead of a flat-prior exclusive selection—to say that all values of y are realized rather than one of them is realized with equal probability for each. Assuming the latter leads to a small observer selection effect while the former does not.

C. Two separated points

To get a sense of how much the prisoner problem in the exclusive case can approach the warden problem, it suffices to consider a prior with nonzero probabilities at two points, $Y \pm \sigma$, where $0 < Y < 1$ and $0 < \sigma \leqslant \min(1/2, Y, 1 - Y)$, so that both points lie in the range [0,1]:

$$p(^{|}y)_{\text{two}} = \tfrac{1}{2}\{\delta(y - (Y - \sigma)) + \delta(y - (Y + \sigma))\}. \qquad (36)$$

$[\delta(x) = 1$ for $x = 0$ and is 0 otherwise.] Since the integrands in $R_W^{|E}$ are linear the σ terms cancel, and we again get $R_W^{|E} = (1 - Y)/Y$. For $R_P^{|E}$, we obtain,

$$R_P^{|E} = \frac{1 - Y(2 - 1/\rho) + (Y + \sigma)(Y - \sigma)(1 - 1/\rho)}{Y - (Y + \sigma)(Y - \sigma)(1 - 1/\rho)}. \qquad (37)$$

If we assume $Y = 1/2$, and define $k \equiv 2\sigma$, then $R_W^{|E} = 1$ and Eq. (37) reduces to

$$R_{P/W}^{|E}(Y = 1/2) = \frac{1 - k^2 + (1 + k^2)/\rho}{1 + k^2 + (1 - k^2)/\rho}. \qquad (38)$$

Note that $0 < k \leqslant 1$. For k near 0, $R_P^{|E}$ approaches 1—two points very close together is very much like the inclusive case. For $Y = 1/2$ and $k = 1$, i.e., when the two points are $y = 0$ and $y = 1$,

$$R_{P/W}^{|E}(y = 0 \text{ or } 1) = \frac{1}{\rho}. \qquad (39)$$

In other words, the prisoner problem in the exclusive case where the prior is that the prison is either all L cellblocks ($y = 0$) or all S cellblocks ($y = 1$) has the same observer selection effect as the warden problem in Eq. (12). By insisting on an either-or-Pick on the enclosing set E, we have, in essence, turned a Be for the prisoner into a Pick on which top-level subset she is in.

So we can go anywhere from no OSE, as in the prisoner case, to a warden-level $1/\rho$ OSE simply by adjusting our prior assumptions. In Fig. 2, we plot $R_{P/W}^{|E}$ as a function of Y for different values of k, which we more generally define as

$$k \equiv \begin{cases} \frac{\sigma}{Y} & Y \leqslant \frac{1}{2}, \\ \frac{\sigma}{1-Y} & Y \geqslant \frac{1}{2}. \end{cases} \qquad (40)$$

For Y near 0 or 1, or k near 0, $R_{P/W}^{|E} \simeq 1 = \rho^0$, and the exclusive case is like the inclusive one. The observer selection effect is maximized for $Y = 1/2$ and $k = 1$, yielding $R_{P/W}^{|E} = \rho^{-1}$ of Eq. (39).

VI. EXCLUSIVE THEORY SELECTION AND THE PRESUMPTUOUS PHILOSOPHER

A. Exclusive theory selection

Instead of taking E to be the top-level set, consider a set of theories, Θ. This set of theories might include very different hypotheses about reality, or they might simply specify different enclosed subsets, such as,

$$\Theta_L : \text{“All cellblocks are type } L\text{”}$$
$$\Theta_S : \text{“All cellblocks are type } S\text{”} \qquad (41)$$

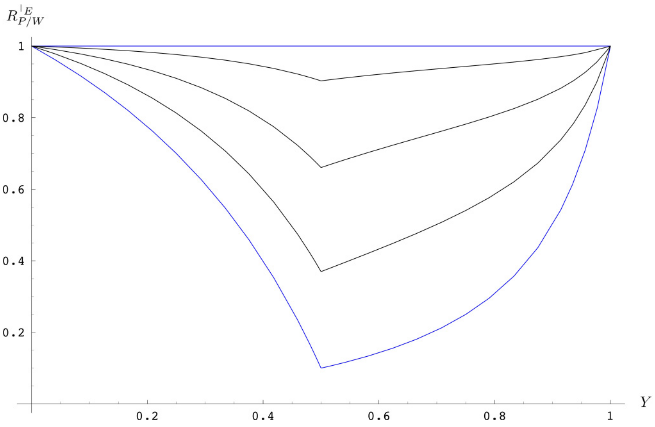

FIG. 2. How to interpolate between the prisoner (no OSE) and warden ($1/\rho$ OSE) cases: For exclusive selection over an ensemble $\{E_y\}$ (y is the fraction of worlds of type S in that ensemble element) which consists of two separated points $y = Y \pm \sigma$, we plot a measure of the OSE, $R_{P/W}^{|E}$ (the ratio of the ratios of posteriors to priors for L and S worlds for the exclusive Pick over ensemble E) versus Y for $\rho = 10$ (the ratio of the number of people per world of type L to that of type S). The OSE depends on how far apart the points are, which is characterized by $k \in (0, 1]$ defined in Eq. (40). Contours top to bottom are for $k = 0$, 0.25, 0.5, 0.75, and 1. There is no OSE for $k \to 0$ (akin to the prisoner case). The maximal OSE (minimal value of $R_{P/W}^{|E}$) is for $k = 1$ at $Y = 1/2$ (akin to the warden case), with a value $R_{P/W}^{|E}(Y = 1/2, k = 1) = 1/\rho = 0.1$.

These two theories could have been encoded in E: They are $E_{y=0}$ and $E_{y=1}$, respectively. But we tend to approach theories differently from ensembles, notably that usually one assumes that only one theory is true, that we have to Pick a theory before proceeding further. This is exclusive theory selection, and the probabilities are the same as in Sec. V. For example, if our prior for the two theories in Eq. (41) are equal, $R_\Theta = P(\Theta_L)/P(\Theta_S) = 1$, then

$$R_{P^{|}/\Theta} = \frac{P(PW^{|}\Theta_L|P_dW^{|}\Theta)}{P(PW^{|}\Theta_S|P_dW^{|}\Theta)} = \frac{1}{\rho}, \qquad (42)$$

just as in Eq. (39). [This is assuming typicality (the SSA). Again, the SIA gives the wrong answer because it does not take into account selections on enclosing sets, here the Pick selection on mutually exclusive theories.]

It is possible to have inclusive selection of a theory, where one assumes multiple theories are realized. For example, one could posit that prisons vary from country to country, so both theories in Eq. (41) would be realized somewhere. There is then no Pick on Θ, and one recovers the probabilities in the inclusive section, where there were no observer selection effects ($R_{P/\Theta} = 1$). One can even have a seemingly fundamental theory be part of an inclusive selection. For example, the landscape in string theory allows different regions of the larger universe to manifest different low energy theories with their own fundamental constants. If one posits that one can be an observer in any region of the landscape that has observers, then that is inclusive theory selection.

As stated, the main point of this paper is to show that the conclusions one draws depend on the assumptions made. If we assume exclusive selection, such as the theories in Eq. (41)

being mutually exclusive, then we will conclude that there are observer selection effects, but if we assume an inclusive case, such as half the prisons have only S cellblocks and half have only L cellblocks, we will conclude that there are no such effects.

B. Probing a multiverse?

Suppose we consider both possibilities about the selection from set P through set E: that it is inclusive, as discussed in Sec. IV, or exclusive, as discussed in Sec. V, and treat these as competing hypotheses, $\Theta_{\rm in}$ or $\Theta_{\rm ex}$. If we treat these hypotheses as mutually exclusive, with a Pick on set Θ, then the overall selection is exclusive. But let us focus on the rest of the selection, from P to E, which is inclusive or exclusive. We can then in principle use our data to alter our posterior probabilities for each hypothesis. Suppose we define E to be everything, so that the inclusive (exclusive) case corresponds to the inclusiverse (an exclusiverse). How do these terms relate to the term "multiverse"? If taken literally, then multiverse simply means that there are more realities than the one we perceive, either via something like parallel universes or just the universe being so large that realities similar to ours occur in some other part of it. That does not actually imply that all possible universes are realized. A set of a few parallel universes, which we will call *a partial multiverse*, is an exclusiverse, since not everything possible is realized. If *all* possibilities are realized, then to avoid ambiguity we will call it *the complete multiverse*. So

> *The inclusiverse* is the same as *the complete multiverse*: All things we deem possible are realized.
>
> *An exclusiverse* is the same as a universe or *a partial multiverse*: Some things we deem possible are not realized.

The question of this subsection is

> Can we determine whether we live in the inclusiverse or an exclusiverse simply by using a datum such as the date?

To get a handle on this, let us consider the prisoner problem again, where our selection in sets PW is a Be. Let PW again be embedded in a larger set E, which itself is considered in the context of one of two hypotheses,

$$\begin{aligned}\Theta_{\rm in} &: \text{“Inclusive selection on } E\text{”}\\ \Theta_{\rm ex} &: \text{“Exclusive selection on } E\text{”}\end{aligned}\tag{43}$$

We need new notation to combine these hypotheses in a single probability, with a "controlled-Pick" on E, so that there is a Pick on E for hypothesis ex, but not for hypothesis in. For this we put a left arrow pointing from Θ to the Pick on E:

$$\begin{aligned}P(PW\,{}^{|}\overleftarrow{E\Theta}_{\rm in}) &= P(PWE\Theta_{\rm in}),\\ P(PW\,{}^{|}\overleftarrow{E\Theta}_{\rm ex}) &= P(PW\,{}^{|}E\Theta_{\rm ex}).\end{aligned}\tag{44}$$

Using this notation, what we want to calculate is the posterior probability for hypotheses $h =$ in or ex given datum d:

$$\begin{aligned}P_{h|d} &\equiv P(PW\,{}^{|}\overleftarrow{E}\,{}^{|}\Theta_h|P_dW\,{}^{|}\overleftarrow{E}\,{}^{|}\Theta) = \frac{P_{d|h}P_h}{P_d}\\ &= \frac{P(P_dW\,{}^{|}\overleftarrow{E}\,{}^{\nmid}\Theta_h|PW\,{}^{|}\overleftarrow{E}\,{}^{\nmid}\Theta_h)P(PW\,{}^{|}\overleftarrow{E}\,{}^{|}\Theta_h)}{P(P_dW\,{}^{|}\overleftarrow{E}\,{}^{|}\Theta)}.\end{aligned}\tag{45}$$

If we define our prior probabilities for $h =$ in and ex to be α and β, respectively, i.e.,

$$P_{\rm in} \equiv P(PW\,{}^{|}\overleftarrow{E}\,{}^{|}\Theta_{\rm in}) \equiv \alpha, \quad P_{\rm ex} \equiv P(PW\,{}^{|}\overleftarrow{E}\,{}^{|}\Theta_{\rm ex}) \equiv \beta,\tag{46}$$

then our posteriors are simply

$$P_{{\rm in}|d} = \frac{\alpha P_{d|{\rm in}}}{\alpha P_{d|{\rm in}} + \beta P_{d|{\rm ex}}}, \quad P_{{\rm ex}|d} = \frac{\beta P_{d|{\rm ex}}}{\alpha P_{d|{\rm in}} + \beta P_{d|{\rm ex}}}.\tag{47}$$

Note that we also need priors for the probabilities of the elements of E. For simplicity, let us assume that the only ensembles with nonzero probability are $y = 0$ (all L-type cellblocks) or $y = 1$ (all S-type cellblocks), which we saw in Eq. (39) gives maximal OSE for the ex case. There is of course no OSE in the in case. For the inclusive case, let us assume equal probabilities for $y = 0$ and 1:

$$P(E_0\,{}^{\nmid}\Theta_{\rm in}|E\,{}^{\nmid}\Theta_{\rm in}) = P(E_1\,{}^{\nmid}\Theta_{\rm in}|E\,{}^{\nmid}\Theta_{\rm in}) = \tfrac{1}{2},\tag{48}$$

but for the exclusive case let us allow them to vary,

$$P(E_0\,{}^{\nmid}\Theta_{\rm ex}|E\,{}^{\nmid}\Theta_{\rm ex}) = q, \quad P(E_1\,{}^{\nmid}\Theta_{\rm ex}|E\,{}^{\nmid}\Theta_{\rm ex}) = p,\tag{49}$$

where $p + q = 1$. Our likelihoods are then

$$P_{d|in} = \frac{\bar{m}}{\bar{n}} = \frac{2}{\rho + 1}, \quad P_{d|{\rm ex}} = q\frac{\bar{m}}{\bar{n}_0} + p\frac{\bar{m}}{\bar{n}_1} = p + \frac{q}{\rho}.\tag{50}$$

We can then plug these likelihoods into Eq. (47) to obtain the posterior probabilities. It is clear that they depend on p (with $q = 1 - p$).

For $p = 1/2$, so that the $y = 0$ and $y = 1$ weights in the ex case match those of the in case in Eq. (48), we obtain posterior probabilities,

$$\begin{aligned}P_{{\rm in}|d} &\equiv \alpha'|_{p=1/2} = \frac{\alpha}{\alpha + \beta\frac{(\rho+1)^2}{4\rho}},\\ P_{{\rm ex}|d} &\equiv \beta'|_{p=1/2} = \frac{\beta}{\beta + \alpha\frac{4\rho}{(\rho+1)^2}}.\end{aligned}\tag{51}$$

Since α and β are $\geqslant 0$ and $\rho > 1$ [so that $(\rho + 1)^2 > 4\rho$], the denominator for α' (β') is larger (smaller) than 1, and datum d seems to decrease (increase) our credence in inclusive (exclusive) selection on E, except in the trivial case where α or β is zero. This would seem to argue that if E is a set of universes (not just prisons), we could use observer data to alter our probability that we live in the inclusiverse.

But there is a second prior in this problem, that of p (with $q = 1 - p$). We chose $p = 1/2$ to make the probabilities for $y = 0$ and $y = 1$ the same as those in the inclusive case. An equally reasonable hypothesis would be to set p equal to the value that gives the same value for datum d for each hypothesis, so that $P_{d|{\rm in}} = P_{d|{\rm ex}} = 2/(\rho + 1)$. With a little algebra, we see that this holds for

$$p = \frac{1}{\rho + 1}. \tag{52}$$

For this value of p, the denominators in Eq. (47) are 1 (since $\alpha + \beta = 1$), and

$$P_{\mathrm{in}|d} \equiv \alpha'|_{p=1/(\rho+1)} = \alpha, \quad P_{\mathrm{ex}|d} \equiv \beta'|_{p=1/(\rho+1)} = \beta, \tag{53}$$

so for this value of p we gain no information about hypotheses in and ex from datum d.

What happened? When we thought, due to Eq. (51), that we had obtained information about hypotheses in and ex from datum d, what we really learned about was the probability of getting datum d based on two factors—whether the selection from E was inclusive or exclusive—*and* the priors we had for the elements of E in each case. To the extent that d tells us anything about these cases, it is about a combination of these factors. We cannot disentangle these factors here. In general, one cannot claim that data tell us about whether we are in the inclusiverse (the complete multiverse) or not unless one can show that all other factors which separate the inclusiverse from exclusiverse hypotheses are fixed.

C. Presumptuous philosopher

In the Introduction, we noted that some authors argued against the doomsday argument by assuming the SIA: that we should weight the probability of some situation by the number of observers in it. As we have discussed, this is essentially a kludge, adding the factor that we found in Be choices without the clear-cut mathematical rationale we presented (based on applying the SSA—typicality—properly). This is perhaps why it has been referred to as "controversial" [11,12].

Nick Bostrom argues against the SIA with the following problem [7,9]. A philosopher is told that theories Θ_L and Θ_S have equal probabilities prior to taking into account any observer information. This is like the problem of exclusive theory selection we considered in Sec. VI A, except that there is no datum d favoring S over L. The philosopher states that there is no need to test which is right (and since this is exclusive selection, only one is right) because, by the SIA, Θ_L is ρ times more likely than Θ_S because there are ρ times as many observers in that case.

Bostrom is right that the philosopher is being presumptuous here, and this is a good argument against the SIA—that if one is to Pick between Θ_S and Θ_L, there should be no effect from there being more observers in the latter case, because we are *picking* a theory. This is simply an example of what we have found regarding the SIA—that it gives the wrong answer when there is a selection from an enclosing set, here Θ. But there is no reason to have invoked the SIA in the first place.

In short, the presumptuous philosopher has no bearing on our results because it argues against the SIA, which we did not use.

We note, however, that if the philosopher correctly uses the SSA and is asked about an inclusive problem, whether she is more likely to be in a domain of the inclusiverse governed by theory Θ_L or Θ_S, she would be correct to answer that she is more likely to be in the former due the SSA weighting by number of observers. In that case she is not presumptuous at all [16].

VII. TYPICALITY

All of the probabilities we have discussed thus far assume that the selection, Be or Pick, is *typical*, that, for example, if the fraction of observers in some subset P_a of P is n_a/n, then the probability of selecting a person in that subset is also n_a/n. Suppose we relax that assumption and allow *atypical selection*, where the probability of selecting a person from subset P_a differs from n_a/n—some values of a are intrinsically more likely to be selected than others [17]. For example, observers at CERN are not typical of Earth's population—they are more likely to be scientists than the population overall. Srednicki and Hartle [18] describe an atypical selection in their Eq. (6.1):

$$\text{“}P(q_1|T, \xi, D_0) = \sum_A \xi_A P(q1@A|T, D_0@A),\text{”} \tag{54}$$

where q_1 is a posterior result, T is a given theory, D_0 is data, ξ is a "xerographic distribution," which is a set of copies A of q_1 at different locations meeting data D_0, and ξ_A is the probability weight of xerographic occurrence A which is not necessarily what we would obtain from a typical selection. We need to translate this all into our notation.

A. Atypical notation

Let us define ξ^0 to be a *typical Be*, a typical selection on the set P (embedded in set W). We are interested in subsets P_a of P for some property a of the people in P:

$$\xi_a^0 \equiv P(P_a W) = \frac{\bar{n}_a}{\bar{n}}, \quad \xi_{a|d}^0 \equiv P(P_a W|P_d W) = \frac{\bar{m}_a}{\bar{m}}. \tag{55}$$

Now let us define an *atypical Be* using ξ to mark the atypical selection point,

$$\xi_a \equiv P({}^{\xi}P_a W), \quad \xi_{a|d} = P({}^{\xi}P_a W|{}^{\xi}P_d W), \tag{56}$$

which may not simply be a ratio of numbers of elements of set P. However, for a given atypical selection ξ on P, we will show that we can always find a new set $\tilde{P}$, with number of people per world $\tilde{n}$, on which a typical selection $\tilde{\xi}^0$,

$$\tilde{\xi}_a^0 \equiv P(\tilde{P}_a W) = \frac{\bar{\tilde{n}}_a}{\bar{\tilde{n}}}, \quad \tilde{\xi}_{a|d}^0 = P(\tilde{P}_a W|\tilde{P}_d W) = \frac{\bar{\tilde{m}}_a}{\bar{\tilde{m}}}, \tag{57}$$

gives the same answer. Here the tilde quantities are related to their counterparts by some scaling factors κ_a and $\kappa_{a|d}$:

$$\begin{aligned} \tilde{n}_a &\equiv \kappa_a n_a, \quad & \tilde{n}_{aK} &\equiv \kappa_a n_{aK}, \\ \tilde{m}_a &\equiv \kappa_{a|d} m_a, \quad & \tilde{m}_{aK} &\equiv \kappa_{a|d} m_{aK}. \end{aligned} \tag{58}$$

We claim that the atypical Be on P, ξ, is equal to the typical Be on $\tilde{P}$, $\tilde{\xi}^0$,

$$\xi_a = \tilde{\xi}_a^0, \quad \xi_{a|d} = \tilde{\xi}_{a|d}^0, \tag{59}$$

if we define κ_a as the ratio of atypical to typical selection,

$$\kappa_a \equiv c\frac{\xi_a}{\xi_a^0}, \quad \kappa_{a|d} \equiv c_d\frac{\xi_{a|d}}{\xi_{a|d}^0}, \tag{60}$$

where constants c and c_d are independent of a. We have the freedom to vary c and c_d because the overall numbers of people in $\tilde{P}$ do not matter, just the ratios we are interested in. However, they do affect the values for $\tilde{\bar{n}}$ and $\tilde{\bar{m}}$:

$$\tilde{\bar{n}} = \sum_a \tilde{\bar{n}}_a = \sum_a \kappa_a \bar{n}_a = c\bar{n} \sum_a \xi_a = c\bar{n},$$
$$\tilde{\bar{m}} = \sum_a \tilde{\bar{m}}_a = \sum_a \kappa_{a|d} \bar{m}_a = c_d \bar{m} \sum_a \xi_{a|d} = c_d \bar{m}, \quad (61)$$

using the fact that probabilities for even atypically selected people sum to 1. Note that we can choose to set c and c_d equal 1 and have $\tilde{\bar{n}} = \bar{n}$ and $\tilde{\bar{m}} = \bar{m}$, but we need not do this. Now we can show Eq. (59) does in fact hold,

$$\xi_a = \frac{1}{c}\kappa_a \xi_a^0 = \frac{1}{c}\kappa_a \frac{\bar{n}_a}{\bar{n}} = \frac{\tilde{\bar{n}}_a}{c\bar{n}} = \frac{\tilde{\bar{n}}_a}{\tilde{\bar{n}}} = \tilde{\xi}_a^0,$$
$$\xi_{a|d} = \frac{1}{c_d}\kappa_{a|d} \xi_{a|d}^0 = \frac{1}{c_d}\kappa_{a|d} \frac{\bar{m}_a}{\bar{m}} = \frac{\tilde{\bar{m}}_a}{c_d\bar{m}} = \frac{\tilde{\bar{m}}_a}{\tilde{\bar{m}}} = \tilde{\xi}_{a|d}^0, \quad (62)$$

and we can write our atypical selection on P as a typical selection on $\tilde{P}$ with number of elements defined by Eq. (58) with κ defined in Eq. (60).

B. Posterior probability

We can now write Srednicki and Hartle's Eq. (54) in our notation. We want the posterior probability $P(PW_K|P_dW)$ but with an atypical Be, i.e., $P({}^{\xi}PW_K|{}^{\xi}P_dW)$:

$$P({}^{\xi}PW_K|{}^{\xi}P_dW) = \sum_a \xi_{a|d} P(P_aW_K|P_{ad}W)$$
$$= \left(\sum_a \xi_{a|d} \frac{\bar{m}_{aK}}{\bar{m}_a}\right) P(W_K) = \frac{\tilde{\bar{m}}_K}{\tilde{\bar{m}}} P(W_K), \quad (63)$$

which we write as a typical Be on set $\tilde{P}$ defined by Eqs. (58) and (60). This is the same expression as for a Be in Eq. (6) with the elements from set $\tilde{P}$. Note that if we condition on a subset a, the selection within that subset is typical (all atypicality comes from nontrivial weighting of the different subsets P_a), thus $P({}^{\xi}P_aW_K|{}^{\xi}P_{ad}W) = P(P_aW_K|P_{ad}W)$.

C. Atypical example

Let us see how this atypical notation works in an example using prisoners of two types. Suppose half the cellblocks are filled with humans ($a = h$) and half filled with zombies ($a = z$). Humans are distributed as in the prisoner problem, $\bar{n}_{hL} = \rho\bar{n}_{hS}$ and $\bar{m}_{hL} = \bar{m}_{hS}$. Zombies have the same distribution in cells, $\bar{n}_{zL} = \rho\bar{n}_{zS}$, but let us assume that all zombies who can think well enough to formulate a question think they meet datum d, i.e., $\bar{m}_{zL} = \rho\bar{m}_{zS}$. If you think it is equally likely that you are a human or a zombie (because half the prisoners are humans and half zombies), and for simplicity you assume $P(W_S) = P(W_L) = 1/2$, then you calculate the typical Be posterior probabilities,

$$P(PW_S|P_dW) = \frac{\bar{m}_S}{\bar{m}}P(W_S) = \frac{2}{3+\rho}, \quad (64)$$
$$P(PW_L|P_dW) = \frac{\bar{m}_L}{\bar{m}}P(W_L) = \frac{1+\rho}{3+\rho}. \quad (65)$$

Thus, unlike the prisoner problem, there *is* an observer selection effect $R_{P/W} = (1+\rho)/2$, favoring that you are in W_L, because there are more zombies matching d in W_L.

But suppose you think it is quite unlikely that you are a zombie, say, because zombies do not usually use Bayesian reasoning. For simplicity, you take $\kappa_{h|d} = 1$ and set $\kappa_{z|d}$ to be some very small number κ—one zombie out of every κ thinks well enough to calculate the probabilities we have been discussing (the ratio of chances you are a zombie to you are a human is κ, not 1). Then you calculate the atypical Be,

$$P({}^{\xi}PW_S|{}^{\xi}P_dW) = \frac{\tilde{\bar{m}}_S}{\tilde{\bar{m}}}P(W_S) = \frac{1+\kappa}{2+\kappa(1+\rho)}, \quad (66)$$
$$P({}^{\xi}PW_L|{}^{\xi}P_dW) = \frac{\tilde{\bar{m}}_L}{\tilde{\bar{m}}}P(W_L) = \frac{1+\kappa\rho}{2+\kappa(1+\rho)}. \quad (67)$$

There is still an observer selection effect, $R_{P/W} = (1+\kappa\rho)/(1+\kappa)$, favoring W_L, but note that when $\kappa \to 0$, $R_{P/W} \to 1$, because there is no OSE due to the human prisoners. If you *assume* you are not a zombie, then you take $\kappa = 0$ and all probabilities spring from P_h—in fact if you are going to do that, you might as well drop the label h and ignore the zombies.

D. Redefine the conditional

Another way of addressing an atypical selection which is due to different subsets a meeting the conditional with different relative frequencies is to redefine the conditional so the weights are the same. For example, in the case above, we deweighted zombies by a factor κ because only that fraction of zombies could formulate the question. So why not limit the sets P and P_d to the subset P_Q of P of people who have formulated the Bayesian question in the first place? As we discuss in the Appendix, adding such a conditional is not just another label but actually redefining the set P as set $[P_Q]$. Then all we need to do is define set $\tilde{P} \equiv [P_Q]$, and typical selection on $\tilde{P}$ gives the probabilities for those atypical people who ask the question.

E. Boltzmann brains

Normal observers are necessarily far from equilibrium and experience an arrow of time of increasing entropy [19]. Fortunately, the observable Universe is in a relatively low entropy state [20,21]. How did it get that way? Ludwig Boltzmann argued that a low-entropy "world" could arise as a stupendously rare fluctuation within a higher-entropy world [13,22]. The prevailing theory of cosmology is more subtle: that our Universe began within a patch of smooth spacetime, which inflated for a time at an exponential rate [23] (for a review, see Ref. [24]). Though inflation has ended here, it has likely not stopped everywhere in the larger Universe. Further, our observable Universe has seemingly entered another era of exponential expansion and seems slated to approach de Sitter space (a spacetime with a positive cosmological constant Λ and vanishing matter density) asymptotically.

If so, then the empty places greatly outnumber the places where normal observers can live. Further, de Sitter space is a thermal state (with a temperature which depends only on the cosmological constant: $T = \sqrt{\Lambda/12\pi^2}$) [25] and thus seems subject to worlds fluctuating into existence via stupendously rare fluctuations. And one may not need such a large fluctuation, the size of a galaxy or a planet, to create observers; one may need only "Boltzmann brains" [26–28], which are spontaneously formed configurations of matter that, for a brief period, are self-aware, including ones that think they are having the thoughts you are having now. Such events are still extremely improbable, occurring at a rate $\sim e^{-\Delta S}$, where ΔS is the reduction in entropy that the fluctuation represents. For a brain-sized object, the timescale to form them, τ_{BB}, will be enormous—of order $e^{10^{70}}$. (Note that the units do not actually matter with numbers this large—switching from Planck times to Hubble times changes the googol-sized exponents by only about 140.) But this is small compared to the timescale for a Hubble volume to fluctuate into existence, τ_{HV} of order $e^{10^{122}}$. This is time enough to form googolplexes of Boltzmann brains, far more than the number of normal observers [13].

One might ask why this is a problem. We do not seem to be Boltzmann brains. In fact, we need to assume that we are normal observers in order to do science. And if one conditions on the assumption that we are normal observers, then the probability of us being a freak observer is zero, no matter how common they are [$P(\text{freak}|\text{normal}) = 0$]. The problem is that if freak observers outnumber us by a large-enough factor, say, a googolplex, there are many, many of them that think that they are experiencing any given moment that any normal observer does, and it is *not* safe to assume that you are a normal observer. So the problem is one of consistency: You need to assume that your observations reflect reality to do science, and thus it is a problem if the resulting science says that this assumption is very likely to be false. The problem is especially acute if there is an *infinite* volume of spacetime which could spawn Boltzmann brains, and only a finite volume containing normal observers. This possibility led Don Page to argue that the Universe must decay rapidly, via bubbles of vacuum decay [29], so as to avoid any infinite patches of spacetime, leading him to predict a lifetime of our Universe shorter than about 20 billion years [30]. Many papers have been written with less drastic proposed solutions, such as having the physical "constants" vary over time [31].

We want to know whether our analysis of typicality has any impact on the Boltzmann brain problem. Since freak observers may be fooled into thinking that they are normal only for a small fraction of their "life," we use *observer moments* instead of observers. Let us assume that there are two types of observer moments per comoving Hubble volume, normal (n) and freak (f), with $\bar{n}_f = \rho\bar{n}_n$ for some constant ρ which now can be any nonnegative real number, and $\bar{n} = \bar{n}_n + \bar{n}_f$ is the total number of observer moments per comoving Hubble volume. The probability to Be a normal observer moment is just the fraction of observer moments per comoving Hubble volume which are normal:

$$P(P_n) = \frac{\bar{n}_n}{\bar{n}} = \frac{1}{1+\rho}, \qquad (68)$$

which is not close to 1 unless $\rho \to 0$. But what we really want is the fraction of observer moments in which the observer is self-aware and could ask a question like "Am I normal?" in the first place. The typical freak observer moment which superficially seems like a normal observer moment might not pass that test. Let us assume that freak observer moments are κ times likely as normal moments to do so. Then we are interested in the atypical selection $P({}^{\xi}P_n)$, which is a typical selection on set $\tilde{P}$, scaled from P by κ on the freak observer moments,

$$P({}^{\xi}P_n) = P(\tilde{P}_n) = \frac{\tilde{\bar{n}}_n}{\tilde{\bar{n}}} = \frac{1}{1+\kappa\rho}. \qquad (69)$$

This probability can go to 1 even if ρ is large if κ is sufficiently small. But if ρ is huge, as the recurrence time of de Sitter space argues, then the probability of being in a normal observer moment is near 1 only if there is an argument that κ is zero.

Boddy *et al.* [32] make such a case. They argue that if the theory is unitary ("many worlds"), de Sitter space is in a stationary state. Fluctuations do occur, including ones which correspond to Boltzmann brains, but they do not actually correspond to self-aware freak observer moments because nothing happens in a stationary state—there is no decoherence corresponding to the splitting of worlds. If true, then this is akin to setting $\kappa = 0$, since being a self-aware freak observer moment is not only atypical, it does not happen. Obviously if $\kappa = 0$, then $P({}^{\xi}P_n) = 1$ independent of how big ρ is.

How might this argument be affected by the fact that our Universe contains matter? Well, rarely, stable matter could play the role of an "environment" by interacting with a Boltzmann brain, causing decoherence. Such atypical Boltzmann brains might thus actually be self-aware. How rare is rare? An upper bound to the fraction κ of such atypical matter-interacting fluctuations is the fraction of Hubble volumes which contain even a single matter particle. Let us define the entropy of a Hubble-volume-sized fluctuation entropy change,

$$\mathcal{S} \equiv 10^{122}, \qquad (70)$$

so that the fluctuation time τ_{HV} for Hubble volumes is $\sim e^{\mathcal{S}}$ and the fluctuation time for Boltzmann brains τ_{BB} is "about" $e^{\sqrt{\mathcal{S}}}$ (more accurately, $\sim e^{\mathcal{S}^{0.57}}$). Then the number of freak observers is huge: $\bar{n}_f \sim \tau_{\mathrm{HV}}/\tau_{\mathrm{BB}} \sim e^{\mathcal{S}}$. The number of normal observers per comoving Hubble volume is proportional to the volume of spacetime in which they can occur. A healthy upper bound on $\bar{n}_n$ is $\mathcal{S}$ (e.g., 10^{20} moments/lyr^3 s $\times$ 10^{31} lyr^3 $\times$ 10^{64} yrs $\times$ 10^7 s/yr), so that

$$\rho \equiv \frac{\bar{n}_f}{\bar{n}_n} \sim e^{\mathcal{S}}, \qquad (71)$$

i.e., the number of freak observer moments is so vast that the number of normal observer moments is irrelevant. Then the probability of being normal vanishes: $P(P_n) \simeq 0$ to a *very* good approximation, yielding a seemingly serious consistency problem. *But* only fraction κ of freak observers actually can be self-aware by the argument above, where κ must be smaller than the fraction of Hubble volumes with any matter in them. de Sitter space expands exponentially fast, so soon there is fewer than one particle per Hubble volume. By the time of the first Boltzmann brains, the fraction of Hubble volumes with a single matter particle is

$$\kappa < e^{-\tau_{\rm BB}} < e^{-e^{\sqrt{S}}}. \tag{72}$$

This is exponentially smaller than ρ is big, and $\kappa\rho$ *does* go to zero so that the relevant probability that we are normal observers, $P({}^{\xi}P_n)$, goes to 1. In summary, by this argument Boltzmann brains are overwhelmingly plentiful, but those which are atypically self-aware are very rare and thus not a problem. That matter effects are negligible is unlikely to come as a surprise to those already convinced by the arguments of Ref. [32]. We do think it is interesting that there is a typicality factor so strong that it overwhelms even an exponentially large factor like the ratio of freak to normal observers ($\kappa\rho \ll 1$).

F. Scarce observers

Thus far we have assumed that observers in models are not rare. In fact, we have assumed that there is one observer per "cell." What if we relax this assumption and assume cells are filled only with probability $p_{\mathcal{F}}$? Hartle, Hertog, and Srednicki show that there is a different kind of OSE called "first-person probabilities" [33]. Consider a set of models Θ_K. If $p_{\mathcal{F}}$ is small enough, then it is possible for there to be no observers in some or all of them (we do not necessarily think that assuming "scarce observers" is a reasonable hypothesis, we are merely considering the consequences of that assumption). First-person probabilities weight models by the probability, $p^{\geqslant 1}$, that there is at least one observer in the model—one cannot be an observer in a model if there are no observers in it. If there are n_J observer locations (e.g., cells in a prison block or Hubble volumes in a Universe) which contain observers with probability $p_{\mathcal{F}}$, then the probability that there are no observers in the model is $(1-p_{\mathcal{F}})^{n_K}$, and the probability that there is at least one observer in the model is [33]

$$p_K^{\geqslant 1} = 1-(1-p_{\mathcal{F}})^{n_K}. \tag{73}$$

Now the inclusive probability $P(P\Theta_K|P\Theta)$ (i.e., multiple theories are realized—a theoryverse) is not affected by $p_K^{\geqslant 1}$ because we are conditioning on there being one observer (the "$P\Theta$"), and the weighting by the number of observers in each model, $p_{\mathcal{F}}n_K$, already takes that into account. So we have

$$P(P\Theta_K)_{p_{\mathcal{F}}} = \frac{p_{\mathcal{F}}n_K P(\Theta_K)}{\sum_J p_{\mathcal{F}}n_J P(\Theta_J)} = \frac{n_K}{\langle n\rangle}P(\Theta_K), \tag{74}$$

where $\langle n\rangle = \sum_J n_J P(\Theta_J)$ is the average number of observer cells per model. Models with more observer cells are favored because it is more likely for an observer to be in such a model, as expected from our previous results. In a cosmological model this corresponds to *volume weighting* [34] where models with greater volume for observers are favored.

What about the exclusive probability $P(P^{|}\Theta_K|P^{|}\Theta)$, which is how one generally selects between competing models? Condition "$P^{|}\Theta$" ensures that there is at least one observer in one of the models, but to ensure that a given model meets that criterion, we need to weight the models by $p_K^{\geqslant 1}$ [33]:

$$P(P^{|}\Theta_K)_{p_{\mathcal{F}}} = \frac{[1-(1-p_{\mathcal{F}})^{n_K}]P(\Theta_K)}{\sum_J[1-(1-p_{\mathcal{F}})^{n_J}]P(\Theta_J)}. \tag{75}$$

There are two interesting limits: where observers are common or rare. First, if $p_{\mathcal{F}}n_K$ is large for some models and tiny in others, then $p_K^{\geqslant 1}$ are close to 1 for the former models, and they have observers. Define these models that certainly have observers by subset $\Theta_{\rm obs}$ and normalization factor $\mathcal{N} \equiv \sum_{J\in\Theta_{\rm obs}} P(\Theta_J)$. Then the probability becomes

$$P(P^{|}\Theta_K)_{\rm common} \simeq \frac{1}{\mathcal{N}}P(\Theta_K). \tag{76}$$

Note that models either "pass" (are in $\Theta_{\rm obs}$) or "fail" (are not in $\Theta_{\rm obs}$). If all models we consider pass ($\Theta_{\rm obs} = \Theta$), then $\mathcal{N} = 1$, and we obtain the usual expression for a Pick probability.

If, on the other hand, all the $p_{\mathcal{F}}n_K$ are small, so there are no models that certainly have observers ($\Theta_{\rm obs} = \emptyset$), then $p_K^{\geqslant 1} \simeq p_{\mathcal{F}}n_K$ [because $(1-p)^n = 1-np+\mathcal{O}((np)^2)$] and the Pick probability becomes

$$P(P^{|}\Theta_K)_{\rm rare} \simeq \frac{p_{\mathcal{F}}n_K P(\Theta_K)}{\sum_J p_{\mathcal{F}}n_J P(\Theta_J)} = \frac{n_K}{\langle n\rangle}P(\Theta_K). \tag{77}$$

This is the same as the inclusive probability! Even though we are Picking between mutually exclusive models K, there is nonetheless a volume weighting factor, not just a pass-fail selection, due to it being less likely that scarce observers are in a model with few places for them to be. So this "first-person" effect of Hartle, Hertog, and Srednicki is somewhat orthogonal to the observer effect we have been discussing: Ours assumes observers in every "cell", $p_{\mathcal{F}} = 1$, and comes from the difference between inclusive and exclusive selection, while theirs assumes the limit where observers are scarce, $p_{\mathcal{F}} \ll 1$, and is the *same* for inclusive and exclusive selection in that limit.

This "first-person" analysis can be used in the context of freak observers. Suppose we consider two models, S and L, which differ only in the volume of spacetime in which freak observers occur. We could assign probability p_n for "you" to arise normally per unit volume of spacetime and p_f for a "freak" observer that thinks they are you (i.e., after any typicality effects have been folded in). Let the volume of spacetime where normal observers can arise be m_K, and the volume where freaks could arise be n_K, which is usually much larger. We want the case where you exist within the model $[1-(1-p_n)^{m_K}]$, *and* that no freak versions of you exist, $(1-p_f)^{n_K}$ (as we argued before, you want to rule out cases where you might be a freak observer for self-consistency). Let us refer to this as "$1n, 0f$." Then the ratio of exclusive probabilities is

$$\begin{aligned} R^f_{P^{|}\Theta} &\equiv \frac{P(P_{1n,0f}{}^{|}\Theta_L)}{P(P_{1n,0f}{}^{|}\Theta_S)} \\ &= \frac{[1-(1-p_n)^{m_L}]}{[1-(1-p_n)^{m_S}]}\frac{(1-p_f)^{n_L}}{(1-p_f)^{n_S}}\frac{P(\Theta_L)}{P(\Theta_S)} \\ &= (1-p_f)^{n_L-n_S}R_\Theta \\ &\to e^{-p_f(n_L-n_S)}R_\Theta, \end{aligned} \tag{78}$$

where $R_\Theta \equiv P(\Theta_L)/P(\Theta_S)$ and we have assumed $m_S = m_L$ (i.e., that the models do not differ in the volume of spacetime available to normal observers). The last line follows for large n.

We can neglect n_S for $n_S \ll n_L$. Then there are two interesting limits. If $p_f n_L$ is small, then freak observers are scarce, and the "first-person" ratio $R_{P\Theta}$ is only slightly smaller than the "third-person" one:

$$R^f_{P|\Theta}\Big|_{p_f n_L \to 0} \simeq (1 - p_f n_L) R_\Theta. \tag{79}$$

This is a slight preference for S models over L ones, but for $p_f n_L \ll 1$ the preference is negligible. The other limit of interest is when both models have problems with freak observers because $p_f n_K$ is large. Then each theory is deweighted by the factor $(1 - p_f)^{n_K}$ which goes to 0, but the factor for L falls much faster and we have,

$$R^f_{P|\Theta}\Big|_{p_f n_L \to 1} \simeq e^{-p_f n_L} R_\Theta, \tag{80}$$

strongly favoring S over L. So under the criterion of "no freaks like me," if there are *no* models without significant probability for freak observers, then the ones which minimize the volume for them to spawn are strongly preferred. Of course, any model which has no freak observers would, by that criterion, be preferred over those.

VIII. GOTT ANALYSIS

J. Richard Gott III wrote about what seems to be an entirely different kind of observer selection effect [14]. He argued that simply by knowing how long some finite-lifetime entity has been observed, one can bound the probability of it lasting a long time. For example, if you live at time t after the start of a civilization, his argument says that simply assuming you are a random observer implies that the probability of the civilization lasting $40t$ is only $1/40$ or 2.5%.

There are a number of problems with this argument, as we shall see. The first is that Gott's analysis did not make use of a prior [35], which Gott then addressed [36]. This point was echoed by Carleton Caves [37], who found that the prior probability for a world having lifetime T needed to obtain Gott's result is the Jeffereys prior, which goes as $1/T$. However, as we shall see, this corresponds to a Pick selection. The prior needed to obtain the probability Gott finds to Be in a civilization lasting time T is *not* the Jeffereys prior, but a prior that goes as $1/T^2$ [38,39]. Caves argued that the analysis was also flawed because it assumed that the observer had to live only during the time span of the "world," and that once one relaxes that assumption, the effect goes away. (This is really about what set of observer moments it is reasonable for one to consider that the moment at hand is randomly drawn from. For Gott's example of the Berlin wall, one could assert that his observation of the wall was drawn randomly from possible moments during the existence of the wall when he could ponder the question of the duration of its existence rather than a random moment from his lifetime that predates and postdates the wall. It is then a question of whether that assumption is reasonable. It is certainly problematic in many cases. For example, it is hard to argue that the observer moment in which you ponder the lifetime of an architectural construction is randomly drawn from all the moments during its existence if you were born before it was built—for a long-lived construction you are necessarily seeing only its earliest moments.) But it should not be a problem in the narrow case of interest to us: where we assign probabilities for the lifetime of the world in which we were born—we are necessarily alive only during the world in which we are born, and so random observer moments in our lifetime are necessarily within the time window of the world's existence.

We will first explain the Gott argument in his notation and then ours. Then we will show how to incorporate a prior, derive results for different priors, and determine which one gives Gott's results. Then we show that Gott's results do not actually represent an OSE, and we trace the source of the effect. Finally, we consider the exclusive case, where one lifetime is picked.

A. Gott's argument

Suppose we are a random intelligent observer of some "world" of lifetime T which has existed so far for time t. We do not know T and we want to know if knowing t tells us anything about T, other than $T \geqslant t$. Gott gives a few examples [14], but they are of two types: things on which our existence does not depend, such as the time span for which the Berlin wall existed, and things on which it does depend, such as the civilization in which we were born. We will not consider the former further, except to note that the second critique of Caves may apply to those situations. Thus, since we assume we live during the world, we can without loss of generality define Gott's quantities as

$$t_{\text{begin}} \to 0 \quad t_{\text{end}} \to T \quad t_{\text{now}} \to t \tag{81}$$

$$t_{\text{future}} \to T_{\text{fut}} \equiv T - t, \tag{82}$$

where we take as a precondition that t is in the range $[0, T]$. This world could refer to our planet (in which case $t \sim 10^9$ years), the era of *homo sapiens* ($t \sim 10^5$ years), our civilization ($t \sim 10^4$ years), or civilization since Bayesian questions like this have been asked ($t \sim 40$ years). One could even try to argue that it refers to the metastable electroweak vacuum ($t \sim 10^{10}$). Now, going back to the original assumption, it is not at all clear that we qualify as a random observer in *any* of these "worlds," but nevertheless let us assume that we do.

First, Gott argues each value of t in the range $[0, T]$ is equally likely. This is true if there is an equal number of observers at each time t in $[0, T]$ (unreasonable in most cases—really t and T are better thought of as the current and final tally of observers in the world) and one selects them at random. This can be loosely written,

$$\text{“}P(t) = \text{const}/T.\text{”} \tag{83}$$

Further, this means that t/T is a random number between 0 and 1, so

$$\text{“}P(t/T) = \text{const.”} \tag{84}$$

Finally, if we sum up the probabilities for our expectation for the remaining time left for the world, $T_{\text{fut}} \equiv T - t$, then we obtain that it is overwhelmingly likely to be of roughly of order t (neither much greater nor smaller than t),

$$\text{“}P\left(\frac{1}{39}t < T_{\text{fut}} < 39t\right) = 0.95,\text{”} \tag{85}$$

or, focusing on the upper end and using $T_{\text{fut}} \equiv T - t$ to write this more generally,

$$\text{“}P(T > Kt) = 1/K,\text{”} \tag{86}$$

where $K > 1$. Note that for $K = 40$, we get the probability of $T_{\text{fut}} = T - t$ being greater than $39t$ is 1/40, or 2.5%, in agreement with Eq. (85) (the upper and lower tails are equally probable). Further, note that these are scale-invariant probabilities: They depend on the ratio t/T independent of whether the scale is decades or millennia.

Gott seemingly found a way to argue that our datum t not only tells us something about our world's eventual lifetime T but argued that T is unlikely to be more than a few times t, no matter the scale.

Is this right?

B. Our argument

As usual, we have a set of observers P and a set of worlds W. As Gott does, we will for simplicity assume that the number of observers at each time is the same. We will use the compact notation outlined at the end of Appendix, i.e.,

$$P(P_\alpha W_\beta) \equiv P(\alpha\beta), \quad P(P_\alpha {}^{|}W_\beta) \equiv P(\alpha {}^{|}\beta), \tag{87}$$

where α and β can be “null,” e.g., $P(T|t) \equiv P(PW_T|P_tW)$ and $P({}^{|}T|t^{|}) \equiv P(P^{|}W_T|P_t{}^{|}W)$. Let us then define the probability density to Be in a world at time t (for a moment lasting dt):

$$p(t) \equiv P(P_{[t,t+dt]}W)/dt. \tag{88}$$

The probability density to Be in a world of lifetime T (one again needs a finite range $[T, T + dT]$) and to Pick a world of lifetime T are

$$p(T) \equiv P(PW_{[T,T+dT]})/dT, \tag{89}$$

$$p({}^{|}T) \equiv P(P^{|}W_{[T,T+dT]})/dT. \tag{90}$$

Note that the probability density to Be in a world is weighted as before by the total number of observers who will ever live in the world, which by assumption is proportional to T, so

$$p(T) \sim T p({}^{|}T). \tag{91}$$

What we are going to do is start with a prior probability density for our world having lifetime T, $p({}^{|}T)$, the likelihood density of being in our world at time t given that it will exist for time T, $p(t|T)$, and we will use Bayes's theorem to calculate the posterior probability density of our world living time T given our datum t, $p(T|t)$.

The likelihood density is, as Gott said, a constant, independent of t,

$$p(t|T) \equiv P(P_{[t,t+dt]}W|PW_{[T,T+dT]})/dt = \frac{1}{T}. \tag{92}$$

Note that if we integrate this probability density over all values of t in $[0, T]$, $P((0 \leqslant t \leqslant T)|T) = \int_0^T p(t|T)dt$ we get 1. This is essentially the same expression as Eq. (83) which we used to express Gott's words, except that here we are explicitly writing a likelihood density conditioned on lifetime T.

The key problem with Gott's analysis is that he jumps right to a probability for t/T without a prior. Let us examine three possible priors, and see which gives the results Gott found. We need the prior probability density for Picking a world of lifetime T, $p({}^{|}T)$, because it should contain all factors *other* than our existence. This is parallel to what we did in the prisoner scenario, though there we needed only probabilities $P(W_S)$ and $P(W_L)$, whereas here we need a function of T over its range. This brings up an important point: We need to define minimum and maximum plausible values of lifetime T for the world we are in, T_- and T_+ respectively. They allow us to properly normalize our expressions, but $T_\pm$ play a more subtle role, too, as we shall see. It must end up being the case that T_+ is greater than both t and T, and that T_- be smaller than T, so if we really tried to define $T_\pm$ without any idea of the timescales involved, we might fail in that. And our expectations for the timescale might change with t. For example, today we might see $T_+ = 10^6$ years as reasonable, but if civilization somehow survives for a million years, then that T_+ will be too low. This is less of an issue for T_+ because we will be able take it to infinity in our final expressions. But T_- is trickier.

Three reasonable choices for our prior $p({}^{|}T)$ are constant, $\sim 1/T$ (Jeffereys), and $\sim 1/T^2$. The normalized priors to Pick a world of lifetime $T \in [T_-, T+]$ are as follows:

$$p({}^{|}T)\big|_{\text{const}} = \frac{1}{T_+ - T_-}, \quad p({}^{|}T)\big|_{1/T} = \frac{1}{T\ln(T_+/T_-)},$$
$$p({}^{|}T)\big|_{1/T^2} = \frac{1}{T^2(1/T_- - 1/T_+)}, \tag{93}$$

which lead to corresponding probability densities to Be in such a world [again assuming the number of observers at each time is constant and Eq. (91)]:

$$p(T)\big|_{\text{const}} = \frac{T}{\frac{1}{2}(T_+^2 - T_-^2)}, \quad p(T)\big|_{1/T} = \frac{1}{T_+ - T_-},$$
$$p(T)\big|_{1/T^2} = \frac{1}{T\ln(T_+/T_-)}. \tag{94}$$

Next we plug the likelihood density $p(t|T)$ in Eq. (92) and our Be priors $p(T)$ in Eq. (94) into Bayes's theorem,

$$p(T|t) = \frac{p(t|T)p(T)}{p(t)}. \tag{95}$$

We can calculate $p(t)$ by integrating $p(t|T)p(T)dT$ over T. We need to be a little careful about the limits of integration because we have defined $T \geqslant t$ and $T \geqslant T_-$, but at the moment it is ambiguous whether t is greater than T_- or not. So let us define the lower limit on T to be the maximum of the two: $t_m \equiv \max(t, T_-)$. For the three different priors, we obtain three posterior probability densities for $T \in [t_m, T_+]$:

$$p(T|t)\big|_{\text{const}} = \frac{1}{T_+ - t_m} \to\sim \text{const},$$
$$p(T|t)\big|_{1/T} = \frac{1}{T\ln(T_+/t_m)} \to\sim \frac{1}{T},$$
$$p(T|t)\big|_{1/T^2} = \frac{1}{T^2(1/t_m - 1/T_+)} \to \frac{t_m}{T^2}, \tag{96}$$

where the right-hand side is the limit where $T_+ \to \infty$. Note that these are the same expressions as the priors in Eq. (93) with T_- replaced by t_m. In other words, the only effect of the datum here is the trivial replacement of the lower bound on T because it is necessarily at least equal to t. So if we quantify the OSE by taking the ratio of the posterior to the prior,

$$R_T \equiv \frac{p(T|t)}{p({}^{|}T)}, \tag{97}$$

then we obtain for the three priors,

$$R_T\big|_{\text{const}} = \frac{T_+ - T_-}{T_+ - t_m} \to 1, \quad R_T\big|_{1/T} = \frac{\ln(T_+/T_-)}{\ln(T_+/t_m)} \to 1,$$
$$R_T\big|_{1/T^2} = \frac{1/T_- - 1/T_+}{1/t_m - 1/T_+} \to \frac{t_m}{T_-}, \tag{98}$$

where again the right-hand side is for $T_+ \to \infty$. In that limit, the first two priors yield $R_T = 1$ even if we include the replacement effect of $T_- \to t_m$. To evaluate the third prior, we need to discuss the value of t_m. There are three possible values:

(i) $t < T_-$, so $t_m = T_-$, and our lower bound on T does *not* increase.

(ii) $t = T_-$, so $t_m = t = T_-$, and our lower bound on T does *not* increase.

(iii) $t > T_-$, so $t_m = t$, and our lower bound on T *does* increase.

The first case means that prior to our using our datum t we assumed that the minimum value of T was larger, asserting that there is zero probability for our world to end between now, t, and T_-. The third case means that prior to taking note of t, we thought that the lower bound on T was T_-, and so datum updates our knowledge, raising that lower bound—yet somehow we are still confident in our prior assumed probability density despite being wrong about its endpoint. The second case strikes us as the most reasonable, because we should already know that $T_- \geqslant t$ and cannot know that $T_- > t$, so we should assume $T_- = t$. Nevertheless, let us consider all three cases.

For the first two cases, $t \leqslant T_-$, all three priors lead to $R_T = 1$. For $t > T_-$ and the $1/T^2$ prior, $R_T = t/T_-$, which is >1. This is an upward shift due to the fact that the posterior probability density is nonzero over a smaller range, $[t, T_+]$, than the prior probability density $[T_-, T_+]$. We will call this a "boundary condition OSE." It is *not* due to the number of elements in the set of observers, P, as in OSEs we considered previously. Rather, it is simply due to raising the lower bound on T from T_- to t.

So, given that there is only at best a boundary condition OSE here, can we reproduce Gott's result? We can. To compare to Gott's result, we have to integrate these functions of T from Kt to T_+ for fixed t (and assume $Kt \in [T_-, T_+]$). This yields probabilities for T in the range of Kt to T_+:

$$P(T > Kt|t)\big|_{\text{const}} = \frac{T_+ - Kt}{T_+ - t_m} \to 1,$$
$$P(T > Kt|t)\big|_{1/T} = \frac{\ln(T_+/Kt)}{\ln(T_+/t_m)} \to 1,$$
$$P(T > Kt|t)\big|_{1/T^2} = \frac{1}{K}\frac{t_m}{t}\frac{T_+ - Kt}{T_+ - t_m} \to \frac{1}{K}\frac{t_m}{t}, \tag{99}$$

where we again take the limit that $T_+ \to \infty$. We see that for the constant and Jeffereys priors, the probability of $T > Kt$ goes to 1. This is not surprising; if we assume the maximum on T is much greater than Kt, then the probability that $T > Kt$ approaches 1, unless our prior falls very fast. For the prior $p({}^{|}T) \sim 1/T^2$ it *does* fall fast enough. If $t \geqslant T_-$, then $t_m = t$ so that

$$P(T > Kt|t)\big|_{1/T^2,\, t \geqslant T_-} \to \frac{1}{K}, \tag{100}$$

and we have obtained Gott's expression in Eq. (86). (For $t < T_-$, this integrated probability is *larger*. We shall see what that means shortly.)

So even though there is only a boundary-condition OSE, we have reproduced the result of Gott, seemingly disfavoring long-term worlds. How is that possible?

C. Why does Gott seem to find an OSE?

To answer this, consider the situation *before* we know datum t and where we Pick a world at random. We know by assumption that with probability 1, $T \in [T_-, T_+]$ (integrate $p({}^{|}T)$ from T_- to T_+ and we get 1). Suppose we ask what the probability is for this world to last K times its minimum, i.e., for $T > KT_-$. We simply integrate $p({}^{|}T)$ from KT_- to T_+. This gives

$$P({}^{|}(T > KT_-))\big|_{1/T^2} = \frac{1}{K}\frac{T_+ - KT_-}{T_+ - T_-} \to \frac{1}{K}. \tag{101}$$

For fixed KT_- and $T_+ \to \infty$ this gives $1/K$. In other words, the effect that Gott found has nothing to do with the datum t but just the rapidly falling prior to which his result corresponds.

Still, it is useful to define a metric which manifestly shows that there is no OSE. For that, let us define the ratio of probability densities integrated over T. Dividing Eqs. (99) by (101) we see that for the $1/T^2$ prior,

$$R_{\int T}\big|_{1/T^2} \equiv \frac{P(T > Kt|t)}{P({}^{|}(T > KT_-))}\Big|_{1/T^2} = \frac{t_m}{t}. \tag{102}$$

For $t \geqslant T_-$, the cases where we obtained Gott's result, we see that this equals 1—that the posterior probability is the same as we obtained using the prior lower bound, and there is no OSE of any kind. For the case $t < T_-$ this ratio is *larger* than 1 [note that the right-hand side cannot exceed K because if $Kt < T_-$ then $P(T > Kt|t) = 1$]. What that means is that from our prior, we assumed that large T worlds were disfavored, but on learning that $t < T_-$, our expectation is *less* negative due to not having reached the lower bound in the world's lifetime, T_-.

So in the inclusive case, there is no $1/T$ OSE. For a fast falling prior we can obtain Gott's $1/K$ result, but it is not an OSE either, just a manifestation of the fast-falling prior we assumed. The only OSE that remains in any of these cases is if we assumed a fast-falling $1/T^2$ prior, thinking that worlds with $T > KT_-$ were very unlikely, but then finding out that $t < T_-$, making our posterior probability *less* dire than our prior.

D. Picking hypothesis T

Suppose that instead of Being in a set of worlds of various lifetimes T, we assert that there is precisely one world, with one future and one lifetime T_*, and we have a set of hypotheses Θ_T for what T_* is. This is an exclusive case, and we are interested in the posterior probability density,

$$\begin{aligned} p({}^{|}T|t^{|}) &\equiv P(P^{|}\Theta_{[T,T+dT]}|P_{[t,t+dt]}{}^{|}\Theta)/dT \\ &= \frac{p(t^{\dagger}|{}^{\dagger}T)p({}^{|}T)}{p(t^{|})}. \end{aligned} \tag{103}$$

The key difference from our analysis above is that the prior that goes into Bayes's theorem is the Pick probability density $p({}^{|}T)$ instead of the Be probability density $p(T)$ [and the corresponding denominator $p(t^{|})$]. The likelihood is not affected, as in the warden case, because the Pick is neutered. The upshot is that the posterior probabilities go as $\sim 1/T$ times those in the Be case in Eq. (96),

$$\begin{aligned} p({}^{|}T|t^{|})\big|_{\text{const}} &= \frac{1}{T\ln(T_+/t_m)} \to\sim \frac{1}{T}, \\ p({}^{|}T|t^{|})\big|_{1/T} &= \frac{t_m}{T^2}\frac{T_+}{T_+ - t_m} \to \frac{t_m}{T^2}, \end{aligned} \tag{104}$$

which means there *is* an OSE for this pick-a-hypothesis-T_*:

$$R_{|T} \equiv \frac{p({}^{|}T|t^{|})}{p({}^{|}T)} \to\sim \frac{1}{T}. \tag{105}$$

Specifically,

$$\begin{aligned} R_{|T}\big|_{\text{const}} &= \frac{1}{T}\frac{T_+ - T_-}{\ln(T_+/t_m)} \to\sim \frac{1}{T}, \\ R_{|T}\big|_{1/T} &= \frac{1}{T}\frac{\ln(T_+/T_-)}{1/t_m - 1/T_+} \to\sim \frac{t_m}{T}. \end{aligned} \tag{106}$$

But as with R_T, $R_{|T}$ is not an ideal metric of OSE, so we should consider the probabilities resulting from integrating over T:

$$\begin{aligned} P({}^{|}(T > Kt)|t^{|})\big|_{\text{const}} &= \frac{\ln(T_+/Kt)}{\ln(T_+/t_m)} \to 1, \\ P({}^{|}(T > Kt)|t^{|})\big|_{1/T} &= \frac{1}{K}\frac{t_m}{t}\frac{T_+ - Kt}{T_+ - t_m} \to \frac{1}{K}\frac{t_m}{t}, \end{aligned} \tag{107}$$

and we obtain the same $1/K$ expression as Gott, now for the $1/T$ prior and $T_- = t$ [the expression is the same as the Gott case, but his description of the problem seems like a Be and thus corresponds to Eq. (100)].

As we did in the Be case, we define an OSE metric as the ratio of integrated probability densities,

$$R_{\int |T} \equiv \frac{P({}^{|}(T > Kt)|t^{|})}{P({}^{|}(T > KT_-))}, \tag{108}$$

which yields for the two priors we consider here,

$$R_{\int |T}\big|_{\text{const}} \to 1, \quad R_{\int |T}\big|_{1/T} \to \frac{1}{K}\frac{t_m}{t}. \tag{109}$$

What this means is that there is a true OSE in the $t \geqslant T_-$ Pick case for the $1/T$ prior which manifests itself as a factor of $1/K$ in that ratio of the integrated posterior to prior probability densities. In other words, the posterior probability density falls with T faster than the prior probability density due to an OSE, which manifests itself in $R_{\int |T}$ being smaller than one. If $t < T_-$, then this is mitigated by the T_-/t factor and is completely erased if $Kt < T_-$, yielding $R_{\int |T} = 1$.

For the constant prior case, there is an OSE in the ratio of probability densities ($R_{|T} \sim 1/T$) but it is washed out when one integrates over T (the posterior probability density falls faster with T than the prior probability density, but both fall slowly enough that their integrated probabilities go to 1, hence their ratio, $R_{\int |T}$, is also 1).

So in the exclusive case there is a real OSE but only if the prior falls fast enough and t is not much less than T_-.

IX. DOOMSDAY ARGUMENT

We are now finally ready to discuss the doomsday argument. The question is

> Do observer selection effects increase the probability that our world will be short-lived?

First, this is a very strange thing to ask. This would entail laying out all the factors which we might use to assign a probability for the world ending soon and separate out the datum of what year it is. But all of the factors are intertwined. For the purpose of the argument below, we need to make the somewhat unreasonable assumption that we can put all factors (e.g., our estimate for the probability of nuclear war) other than that datum into some prior—which is somewhat unreasonable because such a calculation usually depends on temporal information (e.g., the survival probability per year was surely lower in the early days of nuclear weapons than at other times). In any case, we make this assumption for the arguments below.

As it is usually stated, the question is whether the probability that we live in a short-lived world (world type S) or a long-lived one (world type L) is changed given the information about the date (datum d). Clearly this is a Be selection—we are born in this world without the need for that world to be picked. So the zeroth-order analysis is that the case is like our very first example, the prisoner problem, where there was no OSE and thus no doomsday effect. The posterior probability of being in a short-lived world is just given by Eq. (6) and equals the prior probability of picking such a world, so that the ratio of posterior probabilities to their priors, $R_{P/W}$, is 1:

$$P(PW_S|P_dW) = P(W_S), \tag{110}$$

$$R_{P/W} = 1. \tag{111}$$

But we need to be careful just what our assumptions are regarding any larger sets PW are embedded in. For example, if we treat the world types as mutually exclusive hypotheses for short-lived and long-lived worlds, Θ_S and Θ_L, then there is a Pick at that level and there is an OSE akin to that in Eq. (10),

$$P(PW^{|}\Theta_S|P_dW^{|}\Theta) = \frac{P(\Theta_S)}{P(\Theta_S) + \frac{1}{\rho}P(\Theta_L)}, \tag{112}$$

$$R_{P^{|}/\Theta} = \frac{1}{\rho}. \tag{113}$$

Note that here we are saying that either hypothesis Θ_S or Θ_L is realized but not both. This is reasonable only if one assumes that there is only one relevant planet (the Earth) because there are no relevant exoplanets (we are not asking about the inhabitants of inhabitable worlds, just of the Earth), nor copies of the Earth nor multiple futures of this one Earth (in a partial or complete multiverse of some sort, such as in unitary quantum mechanics). Again, there is an OSE given these assumptions because we are saying that there are multiple hypotheses (Θ_S and Θ_L), but only one of them can be realized.

This also assumes that we are typical observers. This, too, can depend on assumptions or on how the problem is stated. For example, by saying that you are equally likely to be any human throughout history fails to take into account the fact that only a tiny fraction of humans throughout history might have asked the doomsday question, at least as stated. For example, humans before 1763 could not have phrased a question in terms of Bayes's theorem [40], and the question "Will our civilization last until the year 2500?" will become moot in 500 years. Similarly, the question "Will our civilization last another 100 years?" changes character as the centuries we survive accrue, since a century becomes a smaller and smaller fraction of the civilization's lifetime. We need to phrase the question in such a way that it would be just as reasonable for a current and future inhabitant of the civilization to ask it.

We argue that the question framed by Gott is actually best, because "Will our world last K times its present age?" is somewhat timescale invariant. There are still issues with assigning a starting point for the world, and a prior probability density for a world of lifetime T, $p({}^{|}T)$ (e.g., neglecting the problem of lumping all other factors into the prior in a time-independent way), but at least it is reasonable for future observers to ask that same question.

So, to be specific, we should ask whether the current age of our world, t, should affect our estimate for the lifetime of the world, T. As we discussed in Sec. VIII the selection in *PW* is a Be, and there is just a boundary condition OSE: the effect of replacing the lower bound on T, T_-, with t, for $t > T_-$. We further argued that it is not reasonable to have chosen T_- either greater or smaller than t, and that for $T_- = t$, the prior and posterior probability densities are equal, so there is no OSE at all:

$$p(T|t)\big|_{T_-=t} = p({}^{|}T) \Rightarrow R_T\big|_{T_-=t} = 1. \tag{114}$$

We then integrate these probability densities over T to obtain the probability of Being in a world with $T > Kt$ given t. As we said in Sec. VIII this goes to 1 unless the prior falls quickly, see Eq. (99). Even in the case of such a fast falling prior, the $1/K$ effect is *not* an OSE but just an artifact of that prior. We quantified that by taking the ratio of integrated probabilities in Eq. (102),

$$R_{\int T}\big|_{T_-=t} = 1, \tag{115}$$

which shows that there is no OSE at all in the Be case.

Is there any somewhat reasonable set of assumptions which leads to a doomsday effect? Yes. If we assert, as we did in Sec. VIII D, that there is a unique lifetime for the world, T_*, and we have hypotheses T for what that T_* is, then there is a Pick on the nested set, $P\,{}^{|}\Theta$, and there *is* an OSE given by Eq. (105):

$$p({}^{|}T|t\,{}^{|}) \sim \frac{1}{T}p({}^{|}T) \Rightarrow R_{{}^{|}T} \sim \frac{1}{T}. \tag{116}$$

But even then, if we choose a constant prior probability density $p({}^{|}T)$, then the posterior probability that the world will last K times longer than it has so far goes to 1 as in Eq. (107). However, if we start with a $1/T$ prior, then the OSE is *not* washed out in Eq. (107), and the OSE survives in the ratio of integrated probabilities, Eq. (109):

$$R_{\int {}^{|}T}\big|_{1/T,\ T_-=t} \rightarrow \frac{1}{K}. \tag{117}$$

This *is* a doomsday effect. It says that given the assumptions above, even if we include our timescale in setting the minimum lifetime ($T_- = t$), integrate our probability densities over T, and normalize to that integrated probability for the prior, there *is* an OSE in the Pick case for a falling prior—that our datum t, by itself, should cause us to reduce our posterior probability that our world will last substantially longer than it has.

So, in summary, there *can* be a doomsday effect, but to have one requires a set of assumptions like this:

(i) All factors other than the current age of the world, t, can be separated out into a prior, which is a simple function of the world's lifetime T.

(ii) You are typical of observers throughout the lifetime of the world, including in what question is being asked.

(iii) There is exactly one true value of the lifetime, T_*, because you consider only one world with one fixed future—so you view the values of T to be mutually exclusive hypotheses for the value of T_*, resulting in a Pick. It is not enough to assume an exclusiverse; it has to a be universe with only one manifestation of the world so that there is only one true lifetime T_*.

(iv) The prior probability density falls as a function of T so that the integration over T does not wash out the OSE.

Absent a set of assumptions like these, there is no doomsday effect. All of these strike us as somewhat unreasonable, except the last. Thus, one can probably not argue that our "world," be it the era of Bayesian reasoning or of the stable electroweak vacuum, is doomed to end soon on the basis of datum t.

X. UNIVERSAL DOOMSDAY ARGUMENT

In addition to the doomsday argument, which concerns our world, some authors have discussed a "universal doomsday" argument [10,11], which says that not only does our datum imply that our world is doomed to die sooner than our priors for its lifetime, due to some OSE, but that all worlds are also doomed to die out sooner due to our datum. Some authors argue that "universal doomsday" can occur even when the doomsday effect is not present. This cannot be. If there is a doomsday effect due to a temporal datum, that lowered posterior probability *can* affect our posterior probability for the lifetimes of other worlds, but it should be clear that if there is no doomsday effect, if we gain no information from our datum about our own world, then our posteriors for other worlds must be unchanged as well.

What we are interested in is how the datum affects an ensemble of worlds, E, as we consider in the inclusive and exclusive cases of Secs. IV and V. In particular, here are the posterior probability densities for ensembles of type y given datum d, in the inclusive case where there is no doomsday effect, and in the exclusive case where there can be one:

$$p(y|d) \equiv P(PWE_{[y,y+dy]}|P_dWE)/dy, \tag{118}$$

$$p({}^{|}y|d^{\,|}) \equiv P(PW^{\,|}E_{[y,y+dy]}|P_dW^{\,|}E)/dy. \tag{119}$$

We ask whether these differ from the prior probability density for y,

$$p({}^{|}y) \equiv P(PW^{\,|}E_{[y,y+dy]})/dy. \tag{120}$$

Universal doomsday is the claim that it does. If the probability distribution function for y changes, then so does our estimate for the average fraction y of worlds of type S. Our prior estimate is the average of y weighted by the prior $p({}^{|}y)$,

$$\langle y\rangle \equiv \int_0^1 y\,p({}^{|}y)dy. \tag{121}$$

After taking our datum d into account, our posterior estimates for that average in the inclusive and exclusive cases are weighted by the posterior probability distribution functions $p(y|d)$ and $p({}^{|}y|d^{\,|})$, respectively,

$$\langle y\rangle_d \equiv \int_0^1 y\,p(y|d)dy, \tag{122}$$

$$\langle y\rangle_{d^{|}} \equiv \int_0^1 y\,p({}^{|}y|d^{\,|})dy. \tag{123}$$

For reasons that will become clear in a moment, let us define metrics for universal doomsday,

$$R_W^{\mathrm{UD}} \equiv \frac{1-\langle y\rangle}{\langle y\rangle}, \quad R_P^{({}^{|})\mathrm{UD}} \equiv \frac{1-\langle y\rangle_{d({}^{|})}}{\langle y\rangle_{d({}^{|})}},$$

$$R_{P/W}^{({}^{|})\mathrm{UD}} \equiv \frac{R_P^{({}^{|})\mathrm{UD}}}{R_W^{\mathrm{UD}}}, \tag{124}$$

where the "${}^{|}$" is there in the exclusive case but not the inclusive case.

It turns out we have already come across these averages. The prior average fraction $\langle y\rangle$ in Eq. (121) is equal to the prior probability of worlds of type S:

$$P(W_S{}^{|}E) = \int_0^1 P(W_S{}^{\dagger}E|W^{\,\dagger}E_y)p(W^{\,|}E_y)dy$$
$$= \int_0^1 y\,p({}^{|}y)dy = \langle y\rangle. \tag{125}$$

Note that if we assume that $\bar{N}_y/\bar{N} = 1$, i.e., that the ensembles differ only by fraction of worlds type S, y, not their number, then $p(W_S{}^{|}E) = p(W_SE)$, so that this is the prior probability of worlds of type S in both the exclusive and inclusive cases. What about $\langle y\rangle_d$ and $\langle y\rangle_{d^{|}}$? They turn out to be simply equal to the posterior probabilities for being in an S world, given datum d, in the inclusive and exclusive cases, respectively:

$$P(PW_SE|P_dWE)$$
$$= \int_0^1 P(PW_SE|P_dWE_y)p(P_dWE_y|P_dWE)dy$$
$$= \int_0^1 y\,p(y|d)dy = \langle y\rangle_d, \tag{126}$$

$$P(PW_S{}^{|}E|P_dW^{\,|}E)$$
$$= \int_0^1 P(PW_S{}^{\dagger}E|P_dW^{\,\dagger}E_y)p(P_dW^{\,|}E_y|P_dW^{\,|}E)dy$$
$$= \int_0^1 y\,p({}^{|}y|d^{\,|})dy = \langle y\rangle_{d^{|}}. \tag{127}$$

These are just the expressions for the posterior probabilities for worlds of type S. In fact we see that

$$p(y|d) = p({}^{|}y),$$
$$p({}^{|}y|d^{\,|}) = \frac{p({}^{|}y)}{y+(1-y)\rho}\left\langle\frac{1}{y+(1-y)\rho}\right\rangle^{-1}. \tag{128}$$

Thus, we see that the metrics for universal doomsday are exactly the same as for doomsday,

$$R_W^{\mathrm{UD}} = R_W^{({}^{|})E}, \quad R_P^{({}^{|})\mathrm{UD}} = R_P^{({}^{|})E}, \quad R_{P/W}^{({}^{|})\mathrm{UD}} = R_{P/W}^{({}^{|})E}. \tag{129}$$

In the inclusive case, $\langle y\rangle_d = \langle y\rangle$, $R_P^E = R_W^E$, and $R_{P/W}^E = 1$ for both doomsday and universal doomsday. So one cannot have one without the other. For the exclusive case, $\langle y\rangle_{d^{|}} \neq \langle y\rangle$, and $R_P^{|E} \neq R_W^{|E}$, but the values for these metrics and $R_{P/W}^{|E}$ for universal doomsday and doomsday are the same. There is a fundamental reason for this: Any doomsday effect, from our data on being in a world selected from ensemble E, can be written as a universal doomsday change in our weighting of the ensemble, i.e., taking $p({}^{|}y) \to p(({}^{|})y|d({}^{|}))$. So universal doomsday and doomsday are two different ways of expressing the same effect or lack thereof.

XI. SLEEPING BEAUTY PROBLEM

Let us apply what we have learned to an observer thought experiment called the Sleeping Beauty problem [41], which has generated disagreement to the point that philosophers have separated into two camps called "halfers" [42–44] and "thirders" [41,45–47]:

Suppose Sleeping Beauty is put to sleep on Sunday. She is woken on Monday, questioned, then put back to sleep, and all her memories of that day are deleted. A fair coin is flipped. If it lands tails, then she is also woken on Tuesday and again questioned, put back to sleep, and her memory deleted. If it lands heads, then she is not woken on Tuesday. In either case she awakes on Wednesday after the experiment concludes. Beauty is aware of all of the above. She is asked each time she is woken for the probability that the coin flip results in "heads."

So-called halfers argue that she should answer "1/2" (each time) because it is a fair coin and she learns nothing from being awakened, and the question is the same as "what is the probability you are in a heads world?" (i.e., a world where the coin landed heads). So-called thirders argue that she should say "1/3" because there is one observer moment associated with a head flip, which we will call Mon-H, and there are two associated with tails, Mon-T and Tue-T, and the question is effectively the same as "What is the probability you are in a heads observer moment?" There are a number of other papers advocating one side or the other, but none of them specify whether the situation corresponds to inclusive or exclusive selection, which we will see is key. A number of authors assume the SIA, which as we have pointed out is an unfortunate kludge that leads to the presumptuous philosopher problem. All authors seem to argue that if Beauty learns that it is Monday, her estimate for "heads" should go up. As we will see, that is not always true. There are also arguments about what wagers she should be willing to accept and whether that reasoning should affect her probability estimate, which we address at the end of the section.

For our formalism, we need two sets. We need a set of worlds, $W = \{W_H, W_T\}$, in which the coin came up H or T. For a fair coin, the probability of picking each world is the same: $P(W_H) = P(W_T) = 1/2$. Nested inside W is the set, P, of Sleeping Beauty observer moments, $P = P_{\text{Mon},H} \cup P_{\text{Mon},T} \cup P_{\text{Tue},T} = \{\text{Mon-}H,\ \text{Mon-}T,\ \text{Tue-}T\}$, where the first element belongs to P_H (which is nested in W_H) and the other two to P_T (nested in W_T). If Beauty does not know the day, then all three of these observer moments are indistinguishable to her.

First, let us look at Beauty's viewpoint within the inclusive case. The probability that she should assign for the coin coming up heads within the world associated with her observer moment is given by the Be probability for a heads observer moment,

$$P(PW_H|PW) = P(P_H) = \frac{n_{,H}}{n} = \frac{1}{3}. \tag{130}$$

That is, in the inclusive case "she" is in all three observer moments, only one of which is a heads observer moment.

If she learns the day is Monday, then the set of observer moments is $[P_{\text{Mon}}]$ instead of P, and her probability for "heads" increases because "she" could be in only two Monday observer moments:

$$P([P_{\text{Mon}}]W_H \mid [P_{\text{Mon}}]W) = \frac{[n_{\text{Mon}}]_{,H}}{[n_{\text{Mon}}]} = \frac{1}{2}. \tag{131}$$

Thus, in the inclusive case, learning that it is Monday *does* increase her probability estimate that the coin came up heads, and both of these probabilities correspond to those of the thirder camp.

Next, let us look at Beauty's viewpoint with exclusive selection. If she does not know the day, then her probability estimate is the same as that of an outside observer, such as the coin flipper, where a single world (coin flip) result is Picked first:

$$P(P\,^{|}W_H|P\,^{|}W) = P(W_H) = \tfrac{1}{2}. \tag{132}$$

In other words, if she assumes there is one world, it has a 1/2 chance of being an H world, and her being awake in an observer moment and not knowing the day brings her no new information. This is the halfer point of view.

Now suppose she learns it is a Monday. One might think that this information should increase her credence in "heads." And in fact, if you were to Pick a single recording of a random day in the experiment (Mon-H in an H world, Mon-T or Tue-T in a T world), and the recording turned out to be from a Monday, you should increase your credence that the coin came up heads, as the halfer camp claims,

$$\begin{aligned}
&P(\text{"Record Picked= Mon-}H\text{"}|\text{"Mon"})\\
&\quad = P(P\,^{|}W_H \mid P_{\text{Mon}}\,^{|}W)\\
&\quad = \frac{P(P_{\text{Mon}}\,^{\dagger}W|P\,^{\dagger}W_H)P(W_H)}{P(P_{\text{Mon}}\,^{|}W|P\,^{|}W)}\\
&\quad = \frac{P(W_H)}{\sum_{F=\{H,T\}} P(P_{\text{Mon},F}|P_{,F})P(W_F)}\,\frac{1/2}{1/2+1/4} = \frac{2}{3},
\end{aligned} \tag{133}$$

but that is not what Beauty does. Instead, if the coin comes up tails, then she experiences *both* Mon-T and Tue-T, so the fact that one of them is on a Monday adds no new information. In our formalism, the way to see this is that the set of observer moments is $[P_{\text{Mon}}]$ instead of P, and her estimate for the probability of heads is just

$$P([P_{\text{Mon}}]\,^{|}W_H \mid [P_{\text{Mon}}]\,^{|}W) = P(W_H) = \tfrac{1}{2}. \tag{134}$$

So if Beauty assumes exclusive selection, learning that it is Monday does not increase her credence that she is in an H world because she is sure to experience a Monday whatever the coin flip. (The reader might note that if Beauty learns that it is a Tuesday, then she should assign zero probability to H, but that fact does not affect her probability for H in the case where she learns it is Monday because in a tails world "she" experiences both days.) This is good, because if she knows it is a Monday, then the amnesia drug is irrelevant, it is the same situation if you ask anyone what the odds a fair coin will come up heads, and there had better be no difference between inclusive and exclusive selection: They both conclude that the probability of heads is 1/2, as they do in Eqs. (131) and (134).

Now, it is interesting to consider what happens if we run the experiment multiple times, once a week for w weeks. We will assume she does not know the day, so the amnesia drug does matter. If Beauty knows the week, then she can treat each of the w experiments like a copy of the original experiment, and she should come to the thirder (halfer) probability in the inclusive (exclusive) case. If she does *not* know the week, then the inclusive probability is unchanged, but something interesting happens in the exclusive case: We get the result Nick Bostrom calls a "hybrid model" [48].

In this exclusive situation, there is one fixed set of coin flips $F = \{F_1, F_2, \ldots, F_w\}$ which actually occurs. The set of worlds can be broken into $2w$ subsets specifying exactly one flip, such as W_{F_1}, where the coin in week 1 came up heads for $F_1 = H_1$ and tails for $F_1 = T_1$, and we do not specify what happened in the other weeks. We can also break W down into subsets with the flips in multiple weeks specified, including the 2^w subsets where they are all specified: $W_{F_1F_2\ldots F_w}$. There is a third way to partition the set W, by the total number of heads, h, in set F, W_h. If $w = 1$, then we have $P(P\,^{|}W_{H_1}) = 1/2$ because she is in either W_{H_1} or W_{T_1} with equal probability. But, if $w > 1$, although she reasons she can experience exactly one sequence of coin flips, e.g., $\{H_1, T_2\}$, then she also reasons that in that world she should lump observer moment Mon$_1$-H_1 with Mon$_2$-T_2 and Tue$_2$-T_2, since she has no way to tell them apart.

So for sequences with half the flips heads, $h = w/2$, she will come up with a probability of 1/3 for the coin having been heads in a given observer moment. For a sequence with a total of h heads out of w flips, the probability of her being in a heads observer moment is $h/[h + 2(w - h)]$. Thus she just needs to weight this probability by the probability that the sequence that occurs has h heads, $P(W_h)$, which is $\frac{1}{2^w}\binom{w}{h}$:

$$\begin{aligned}
&P\big(\big(P_1\,{}^{|}W_{H_1}\big) \vee \cdots \big(P_w\,{}^{|}W_{H_w}\big)\big|P\,{}^{|}W\big)\\
&\quad= \sum_{h=0}^{w} P\big(\big(P_1\,{}^{|}W_{H_1}\big) \vee \cdots \big(P_w\,{}^{|}W_{H_w}\big)\big|P\,{}^{|}W_h\big)P(W_h)\\
&\quad= \sum_{h=0}^{w} \frac{h}{h + 2(w - h)}\frac{1}{2^w}\binom{w}{h}. \qquad (135)
\end{aligned}$$

For $w = 1$, this is 1/2, for $w = 2$ it is 5/12, which is midway between 1/2 and 1/3, and for $w = 10$, the probability of heads drops to about 0.35. For larger and larger $w, P(W_h)$ is approximately a narrower and narrower Gaussian centered on $h = w/2$, and the probability for Beauty's heads observer moments gets closer and closer to 1/3. In other words, exclusive selection with a large number of indistinguishable trials becomes indistinguishable from inclusive selection.

Let us consider what happens if we ask Beauty to wager on whether the coin will come up heads or tails. Can she distinguish whether she is in a reality that corresponds to the inclusive or exclusive case? The answer is no, because they lead to the same result, though for seemingly different reasons. Suppose she is offered x:1 odds that the coin landed heads. We will consider the cases where she bets at every awakening, or only on Mondays. First, consider how Beauty would see the situation on Wednesday, after the experiment is over. Whether she is in the inclusive or exclusive case, she calculates that she has a 1/2 chance of being in a world where the coin came up heads and she won x on Monday and a 1/2 chance of being in a world where the coin came up tails and she lost 1 on both Monday and Tuesday, so she calculates her average winnings to be

$$\Delta = \tfrac{1}{2}(x - 2). \qquad (136)$$

Thus, she will break even ($\Delta = 0$) if she is given 2:1 odds. If the betting occurs only on Mondays, then, whether she is in the inclusive or exclusive case, she calculates that she has a 1/2 chance of being in a world where the coin came up heads and she won x on Monday and a 1/2 chance of being in a world where the coin came up tails and she lost 1 on Monday. Thus Beauty after the experiment calculates her average winnings on Mondays to be

$$\Delta_{\text{Mon}} = \tfrac{1}{2}(x - 1), \qquad (137)$$

and she will break even ($\Delta_{\text{Mon}} = 0$) on Monday bets if she is given even money, 1:1 odds.

How can her winnings be the same for the inclusive or exclusive case when her credence for heads differs for them (if she does not know the day)? If she assumes she is in the exclusive case, then her reasoning during the experiment is exactly the same as afterward. She has a 1/2 chance of being in a world where the coin comes up heads and she wins x on Monday and a 1/2 chance of being in a world where the coin comes up tails and she loses 1 on both Monday and Tuesday. Thus she calculates her winnings for betting each day [on Mondays] to be Eq. (136) [Eq. (137)]. The exclusive case and Wednesday results are the same because they both refer to head and tail worlds.

If she assumes she is in an inclusive case, then "she" is in all three of the observer moments, {Mon-H, Mon-T, Tue-T}, and so if she bets in each, her winnings per observer moment are

$$\Delta^{\text{moment}} = \tfrac{1}{3}(x - 2). \qquad (138)$$

If she bets only on the two Monday moments, then her winnings per observer moment are

$$\Delta_{\text{Mon}}^{\text{moment}} = \tfrac{1}{2}(x - 1). \qquad (139)$$

But to compare apples to apples, we need to know what she thinks the winnings per *world* will be, which just changes the normalization factor for Eq. (138) by the number of observer moments per world, which is 3/2: $\Delta = \frac{3}{2}\Delta^{\text{moment}} = \frac{1}{2}(x - 2)$. For the Monday case, the number of observer moments and worlds is the same, so $\Delta_{\text{Mon}} = \Delta_{\text{Mon}}^{\text{moment}} = \frac{1}{2}(x - 1)$, and we again get Eqs. (136)–(137).

So an inclusive Beauty calculates the same winnings per world as an exclusive Beauty. Inclusive Beauty needs 2:1 odds to break even because she wins in only one observer moment of three. Exclusive Beauty needs 2:1 odds to break even because although she has a 1/2 probability of a heads world picked out by the coin flip, whenever she is in a tails world she loses twice. What this means is that there is no practical difference between the inclusive and exclusive cases in this thought experiment and no way to tell them apart.

The question, "What credence do you assign to heads?" has answer "1/3" if Beauty sees herself as being in all three observer moments and "1/2" if she sees herself as living in an H world or a T world. So, in the end, the only difference between inclusive Beauty (thirder position) and exclusive Beauty (halfer position) is that the former sees "herself" in all three observer moments with equal probability and the latter sees "herself" in one of two worlds with equal probability. For the halfer, the person in Mon-T and Tue-T is the same, a temporal continuation of one being, but not the same person as Mon-H because they are mutually exclusive timelines. For the thirder, all three observer moments correspond to the same person, an inclusive viewpoint. Neither of these is inherently right or wrong; it is a matter of how we define "self"— we do not give an answer about which camp is "right" because they are each right given a reasonable set of assumptions. We can analyze the problem with either definition, but there is no physical difference between them, as shown by the identical betting odds for the halfer and thirder viewpoints.

Note that one can rephrase the single-run Sleeping Beauty problem as several equivalent problems, such as the sailor's child problem [49], but the answer is the same: For the inclusive case the probability is 1/3, and for the exclusive case it is 1/2, and there is no way to tell them apart with betting.

Finally, it is possible to construct a similar Gedanken-experiment where betting *can* distinguish between inclusive and exclusive cases. Motivated by Nick Bostrom's incubator problem [7], Scott Aaronson suggests the following scenario [16]: If a fair coin comes up heads, then Beauty H-One is cloned into existence; if tails, then Beauties T-One and T-Two are cloned into existence. If you find yourself to be one of these people, then what odds would you need to bet that the coin comes up heads? One needs to be extra careful when observers are created like this. In the exclusive case, if H, then you are H-One and you win x; if T, then you are either T-One or T-Two, and you lose 1, so $x = 1$, you are willing to take 1:1 odds. For the inclusive case, you need to specify your assumptions about personhood. H-One wins x, and T-One and T-Two each lose 1, but which of them are "you"? Here are three possibilities:

(i) You are exactly one of the three. You have $1/3$ chance of winning x and $2/3$ chance of losing 1, so $x = 2$, you need 2:1 odds.

(ii) You are one person each world. If heads, then you are H-One; if tails, then you are *one* of T-One or T-Two. You have $1/2$ chance of winning x and $1/2$ chance of losing 1, so $x = 1$, you need 1:1 odds.

(iii) You are all three. You have $1/3$ chance of winning x and $2/3$ chance of losing 1, so $x = 2$, you need 2:1 odds.

So with the first and the third assumptions, the inclusive case *differs* from the exclusive one, whereas it does not for the second assumption. As we have stressed throughout this work, carefully specifying assumptions is crucial.

XII. HEURISTIC SUMMARY AND FUTURE DIRECTIONS

We fully recognize that some readers interested in the topic of observer selection effects are not used to as much math as we used. To that end, we provide a heuristic summary of our main results. We end by pointing to some directions in which this line of research may proceed.

Our central goal was to study the claim that there is a doomsday effect—that by taking into account one's temporal location in a world that datum leads one to conclude that the world will end sooner than one otherwise would have thought. Along the way, we built the tools needed to investigate that claim, laid out arguments about when the doomsday effect holds, and discussed related issues, such as the problems in cosmology due to Boltzmann brains.

Throughout the paper, we discussed probabilities of selecting "people" from some set P. Usually the people were the observers in the problem. The key distinguishing element about whether there is an OSE or not is if the selection is a Pick or a Be—whether one first *picks* a "world" that the person belongs to or whether no such picking is needed because the person just *is* in the world.

In Sec. II, we explored the latter via the prisoner problem. If you are a prisoner in a cell, then no one has to select that cell, cellblock, or prison for you to experience an observer moment there. You just *are* there. As a result, you are more likely to Be in a cellblock type L, which has more prisoners than a cellblock of type S, and that effect exactly cancels the effect of learning rank information d, which would otherwise favor you in being in a cellblock type S (see the left half of Fig. 1).

Contrast that to Sec. III, where we considered the warden problem, where a warden has to *pick* a cellblock before selecting a prisoner. This is the way things usually work when not selecting observers: When the entity being selected is in an enclosing set, such as a prisoner in a cell within a cellblock, to select them one has to pick the outer set, such as the cellblock, first. The effect of this Pick is to nullify the counteracting effect, seen in the Be case, due to the number of prisoners. The result is that the rank information d *does* tell you that if you are picked by the warden, you are more likely to be in a cellblock type S (see the right half of Fig. 1).

Actually, to be more precise, the issue is whether there is *any* selection beyond the one needed on the innermost (leftmost, in our notation) set and not whether that selection is a Be or Pick. If the selection on the leftmost set is the only one, then we call it *inclusive* selection. If there is a selection on one or more of the enclosing sets, then we call it *exclusive* selection. In most of the inclusive cases we considered the selection of the innermost set was a Be. This is unsurprising, because in order to physically select elements of a set within some set of "worlds," one usually must pick the "world" (urn, cellblock, or civilization) first. (We did give a counterexample, the warden cafeteria problem, where the warden directly picks a prisoner in the cafeteria, circumventing the enclosing set W (the prisoners are still labeled by the "world" that they belong to, just not constrained to be selected via that world). And it is also possible to have a Be selection on a set other than the leftmost set by making P an enclosing set for some other set which the observer picks from, and then the situation will necessarily be exclusive.)

We then explored the concepts of inclusive and exclusive selection by extending our analysis of the prisoner problem to the largest physical enclosing set in the problem, which we call E. For our problem, this corresponds to a set of prisons containing various fractions of S and L cellblocks. In the inclusive case (Sec. IV), the only selection is on the leftmost set (a Be selection of set P). We then considered exclusive selection (Sec. V), where there is selection on E in addition to the Be selection on P. As in the prisoner problem, we found that there is no OSE in the inclusive case. In the exclusive case, there is an OSE, but its magnitude depends on our prior assumptions. One can find effects which range from nearly no OSE to an OSE as large as in the warden case (see Fig. 2). The larger the differential between the choices one picks from, the larger the OSE. We can generalize E to comprise "everything," a set of all possible universes. Inclusive selection then corresponds to *the inclusiverse*, which we also later called *the complete multiverse*, which simply means that we assume all possibilities are realized. Exclusive selection corresponds to *an exclusiverse*, where only some possibilities are realized.

Next, in Sec. VI, we added an enclosing set of theories, Θ. We tend to view theories and hypotheses as mutually exclusive: One must *pick* one and then analyze the resulting scenario. But that Pick introduces an OSE because now the selection is exclusive, so one should be careful not to promote coexisting possibilities to hypotheses, such as "I am in an S cellblock." Instead, one should say that there are multiple physical cellblocks, and we are in one of them with some probability for being in an S cellblock. If we really want to have coexisting hypotheses, then we would need to have inclusive selection on Θ, a "theoryverse" if you will. That is not as unreasonable as it seems. For example, the string landscape predicts multiple coexisting theories. Another avenue we took in this set-of-theories analysis was to ask if we can probe whether we live in the inclusiverse or an exclusiverse. It is not generally possible, because it is usually impossible to disentangle other effects. We also briefly discussed the presumptuous philosopher problem. It is not a problem for us because we do *not* make use of something called the self-indication assumption and argue against its use. (We noted in several places that if we use the SIA—where a weighting factor for observers is put in by hand instead of it arising naturally out of typicality and keeping track of how observers are selected—then we get the wrong answer when there is exclusive selection. The presumptuous philosopher problem is an example of this.)

Thus far, we had assumed that whatever selection was done, was "typical," that is, corresponding to what one would get by random selection of a given subset of entities from a set. We relaxed that assumption and found that any atypical selection can be made typical by a simple redefinition of the relevant sets. This allowed us to address the question of Boltzmann brains, which are hypothetical freak observer moments which arise from very rare fluctuations. They are a problem in a stupendously large universe where it is possible for them to dominate normal observers, which are confined to a small subset of the spacetime. This is a consistency problem because we must assume that we are not freak observers for us to argue that we have a correct understanding of the world, so that understanding is inconsistent if it predicts that we are freak observers. We examined an argument by Boddy *et al.* [32] that there are no self-aware freak observers because at late times the Universe will be an empty exponentially expanding de Sitter space with no decoherence to split into "many worlds." We argued that there could be decoherence effects from diluted matter, but an upper bound on the typicality of that is so small that it counters the huge number of future freak observers such that, by this argument, there are essentially no self-aware freak observers. We also used the analysis of Hartle, Hertog, and Srednicki to demonstrate a "first-person probability" effect which is somewhat orthogonal to ours—that when models with observers are scarce, models with more places for them to be are favored, even with exclusive selection. Conversely, if all viable models allow potentially many freak observers, then those with fewer places for those freak observers to fluctuate into existence are favored.

We then considered the analysis of J. Richard Gott III in Sec. VIII, which seems to constitute a different kind of OSE. He argued that one can bound the probability of a world lasting time T using an observer's time t since the start of the world; this is strange because the selection seems to be inclusive: just the Be selection of the observer. One problem is that his original treatment did not include a prior, which is essential. We showed that one needs a fast falling ($\sim 1/T^2$) prior to reproduce his results. Then there is an effect, but it is *not* an OSE, rather just an artifact of the fast-falling prior. However, if we consider a scenario with a Pick selection of a unique lifetime for the world, and the prior falls with T, then there *is* an OSE.

All of this prepared us to address, in Sec. IX, the doomsday question, "Do observer selection effects increase the probability that our world will be short-lived?" The answer is "probably not." One must first write the question in a scale-invariant way, by which we mean that it makes just as much sense to ask at any timescale during the world. A question that could work is "Will our world last K times its present age?," which naturally leads to using the formalism we developed in Sec. VIII for the Gott analysis. There are scenarios where it is reasonable for the selection there to be exclusive, and it *is* possible to conclude that there is a doomsday effect but only under a set of assumptions akin to those listed at the end of Sec. IX.

Several papers have argued for a universal doomsday effect, which says that our data imply that worlds on average are probably more short-lived than we would have estimated without our data. We showed that universal doomsday and doomsday are inextricably linked because if our expectation for the fraction of short-lived worlds changes as a result of our data, so does our expectation for the lifetime of our world and vice versa. So the assumptions needed for a universal doomsday effect are the same as those needed for a doomsday effect.

We then applied our formalism to a somewhat different scenario called the Sleeping Beauty problem. Beauty is woken once or twice during an experiment, depending on a coin flip, and her memory of each awakening is deleted. What probability should she assign to the coin having come up "heads"? This would seem to be trivial but has led to philosophers dividing into two camps, "halfers," who would assign probability $1/2$, and "thirders," who would assign probability $1/3$. It turns out that they are both right. The problem is that the question is insufficiently clearly posed and each answer is right, given a particular question. If Beauty views "herself" as occupying the three equally likely observer-moments, the inclusive case, then she agrees with the thirders. If, on the other hand, she views "herself" as being in one of two possible timelines—in the one waking session of the "heads world" *or* the two waking sessions of the "tails world"—then she will agree with the halfers. These are both reasonable ways of interpreting who "she" is. They might also be interpreted as implying whether the world is a multiverse (in the inclusive case) or not (in the exclusive case), though this is an extrapolation—all she is really doing is assuming one or the other definition of self. Anyway, the two cases are physically indistinguishable. For example, we showed that both cases lead to precisely the same betting outcomes, though Beauty arrives at the same correct odds of winning in each case for different reasons. We also discussed multiple trials, and the creation of observers, which may help extend the formalism of the paper to more general problems.

So we have explored multiple ways in which it matters how observers are selected. The key factor is whether the selection is inclusive or exclusive. There can be an OSE in the latter case but not the former, at least for the problems we considered. Inclusive selection means that all events considered actually occur, though you may not experience them, such as prisoners being in an S and an L cellblock. Exclusive selection means assigning nonzero probabilities to some events which do not occur, such as picking an S or L cellblock. So

> Observer selection effects arise from assuming that there are some possibilities which are not realized.

Among other things, to have a doomsday effect requires such an exclusive selection, which we wrote as, "There is exactly one true value of the lifetime, T_*, because you consider only one world with one fixed future." It is thus crucial that one carefully lays out all of one's assumptions, because whether there is an OSE or not depends on them.

Finally, we lay out some possible future directions for this work.

A simple direction to go in is to relax some of the assumptions we made, such as ρ being constant across the ensemble of possibilities or that the subsets are nonoverlapping (see Appendix) to generalize our results.

Almost all of our analysis was classical. It would be interesting to explore further the quantum context. One consequence is clear: If quantum theory corresponds to something like the many-worlds interpretation, then we are in a multiverse with inclusive selection of events. If there is "wave-function collapse," so that there is only one reality, then there is an exclusive selection. But a comprehensive evaluation of our discussion in the quantum context may turn up interesting results. For example, what of quantum observers, which comprise superpositions of observer states?

Another avenue of inquiry is how to analyze a theoryverse, such as the string landscape. Is it reasonable to assume the inclusive case? In other words, should we sum probabilities of "observers like us" from different parts of the string landscape which contain observers similar to us despite operating with different physical laws? If so, then it is not the probability of a given vacuum in the landscape that matters but that probability times its effective number of observer moments.

Finally, while we discussed atypical observers, and the problem of Boltzmann brains, there is perhaps more to learn from studying what one might call "freak observers," any observer who happens to experience freakish conditions. There are many metrics for "number of observers" in addressing the problem of Boltzmann brains, and it would be useful to see if our results shed any light on them. Also, in a multiverse there are otherwise normal observers who happen to experience statistical fluctuations of many standard deviations who draw erroneous conclusions. How do we treat such observers, especially with the recognition that it is not impossible in a multiverse that we are one of them?

ACKNOWLEDGMENTS

We thank Scott Aaronson, Jim Hartle, Carl Caves, Steve Carlip, and the anonymous referees for useful suggestions.

APPENDIX: NOTATION

Consider two sets, A and B. We will write AB to mean the compound set that consists of set B, and of set A that is *nested* in B, by which we mean that every element of A is associated with exactly one element of B. If, for example, A is a set of nuts and B is a set of jars, then AB is a set of jars with nuts in them. Formally, every element $A_i \in A$ has a secondary label j which corresponds to a specific element $B_j \in B$. So $A_{i,j}$ is an element of A which is associated with (or, usually, "in") an element B_j of B, and $A_{,j}$ denotes all elements in A which correspond to a given B_j. But we do not usually refer to labels for individual elements. Instead, we focus on subsets. Let us define subsets A_a and B_b of sets A and B by properties a and b, such as the subset of all nuts which are peanuts or cashews or the subset of large or small jars. We will assume that these subsets are nonoverlapping and form a complete basis, i.e.,

$$A = \bigcup_a A_a, \quad A_a \neq \emptyset, \quad A_a \cap A_{a' \neq a} = \emptyset, \tag{A1}$$

and the same for B and B_b (in our example above, all the nuts are peanuts or cashews and all the jars large or small). Further, we can define $A_{a,b}$ to be the subset of A whose elements all belong to A_a and correspond to some element in subset B_b, e.g., all cashews in small jars, $A_{c,S}$, are in the set of cashews A_c and are "in" a small jar (they correspond to an element in B_S). Note that the set A is the union of all its nonoverlapping subsets: $A = \bigcup_{a,b} A_{a,b}$. Further, the subset $A_{,b}$ is the union of all subsets corresponding to label b, independent of a, i.e., $A_{,b} = \bigcup_a A_{a,b}$. For example, $A_{,S}$ is the set of all nuts in small jars, which is the union of peanuts in small jars ($A_{p,S}$) and cashews in small jars ($A_{c,S}$).

Let us define the number of elements of A, A_a, $A_{,b}$, and $A_{a,b}$, to be n, n_a, $n_{,b}$, and $n_{a,b}$, and the number of elements of B and B_b to be N and N_b. Note that since A is the union of nonoverlapping subsets $A_{a,b}$, we have $n = \sum_a n_a = \sum_b n_{,b} = \sum_{a,b} n_{a,b}$, and since B is the union of the nonoverlapping subsets B_b, we have $N = \sum_b N_b$. We also define the number of elements in a subset *normalized* by the number of elements in its next enclosing set with an overbar:

$$\bar{n} \equiv \frac{n}{N} = \sum_b \bar{n}_{,b} \frac{N_b}{N}, \tag{A2}$$

$$\bar{n}_a \equiv \frac{n_a}{N} = \sum_b \bar{n}_{a,b} \frac{N_b}{N}, \tag{A3}$$

$$\bar{n}_{,b} \equiv \frac{n_{,b}}{N_b}, \tag{A4}$$

$$\bar{n}_{a,b} \equiv \frac{n_{a,b}}{N_b}. \tag{A5}$$

[Note that all $N_b \neq 0$ by definition, see Eq. (A1).] For example, $\bar{n}_{c,S} = n_{c,S}/N_S$ is the average number of cashews per small jar, which is the number of cashews in small jars divided by the number of small jars, and $\bar{n}_c$ is the average number of cashews per jar, which is the sum of the average number of cashews in each type of jar weighted by the fraction of jars that are of that type: $\bar{n}_c = \sum_J^{L,S} \bar{n}_{c,J}(N_J/N)$ (J is summed over S and L).

In most of the problems we consider, the leftmost set will be P, a set of people, and the set it is nested in, W, is a set of worlds of some kind. The main subset of the leftmost set we will be interested in is "d," those people matching datum d. Since we will often contrast the number of people, n, with the number of people matching datum d, n_d, and that is the only subset we need to worry about, we drop the comma before nesting subset label b, and define m:

$$n_b \equiv n_{,b}, \quad m \equiv n_d, \quad m_b \equiv n_{d,b}. \tag{A6}$$

We are interested in the probability of selecting an element of some set that belongs to a subset of that set. We will assume that the selection is random and the same for each element, so that the probability of selection is equal to the fraction of elements in the subset (if this is not the case, we can always make it so by weighting the number of elements of the subsets by some scaling factors—see Sec. VII on typicality). Let us define $P(A_a)$ to mean "the probability that a randomly selected element of set A belongs to subset A_a." Note that $P(A) = 1$, since an element selected from A belongs to A by definition. So $P(A_a) = P(A_a|A)$ because the conditional A just means that "an element was randomly selected from A," which is already part of the definition of A_a. With these assumptions,

$$P(A_a) = P(A_a|A) = \frac{n_a}{n} = \frac{\bar{n}_a}{\bar{n}}, \quad P(B_b) = \frac{N_b}{N}. \tag{A7}$$

Note that we can thus replace (N_b/N) in Eqs. (A2) and (A3) with $P(B_b)$. For example, if A is the set of cards in a deck, then $P(A_{\text{clubs}}) = 1/4$ and $P(A_{\text{aces}}) = 1/13$.

So long as we are selecting from one set only, there is no ambiguity. But if we are selecting from compound set AB with set A nested in set B, there are two possibilities: Either we first select an element of B_j of B, and then an element $A_{i,j}$ which corresponds to (is "in") element B_j, which we call to *Pick*, *or* we directly select the element A_i, despite being nested in set B, which we define as to *Be*. One has to *pick* a nut from a jar: Select a jar B_j and then select a nut from within the jar. But if the elements of A are themselves observers, say, prisoners in specific cellblocks, then there is another way to select: You can *be* a prisoner in a cellblock without having to perform a cellblock selection—you are just there. (It is possible to Pick directly from set A even if it is nested in B if the correspondence between $A_{i,j}$ and B_j is not really to be "in" it. For example, set B could correspond to a label, S or L, we place on each nut, and we toss them all together and randomly select one. No jar selection is needed to do that, yet the nesting is preserved by the labeling. We mention this briefly in Sec. III with the warden cafeteria problem.)

Be probabilities are simple, just the fraction of elements in the inner set meeting the criteria:

$$\begin{aligned}
P(AB) &= P(A) = \frac{n}{n} = 1,\\
P(AB_b) &= P(A_{,b}) = \frac{n_{,b}}{n} = \frac{\bar{n}_{,b}}{\bar{n}}P(B_b),\\
P(A_aB) &= P(A_a) = \frac{n_a}{n} = \frac{\bar{n}_a}{\bar{n}},\\
P(A_aB_b) &= P(A_{a,b}) = \frac{n_{a,b}}{n} = \frac{\bar{n}_{a,b}}{\bar{n}}P(B_b).
\end{aligned} \tag{A8}$$

Pick probabilities are weighted by the selection that first must be made on set B. We use a superscripted vertical bar ${}^{|}$ to indicate a Pick from the set immediately to its right. It is akin to a conditional within the statement, e.g., "$A_a{}^{|}B_b$" means "we pick an element of type b from set B and then from the elements of A corresponding to that element of B we select an element of A that is in subset A_a." This is the same as saying "we picked an element in A_a from A given that we picked an element of B_b from B." If there are no subset labels indicated to the left of a Pick, then the situation is as if we are ignoring that set. So $P(A{}^{|}B_b) = P(B_b)$ because after we pick an element type b from B with probability $P(B_b)$, it is certain that the element we pick from A is from subset A (which is just the whole set A). (We assume that there is some such element of A, i.e., $A_{a,b} \neq \emptyset$.) If there are subsets specified to the left of the Pick, such as in $P(A_a{}^{|}B_b)$, then we can write it as a product of conditional probabilities defined below, $P(A_a{}^{|}B_b|A{}^{|}B) = P(A_a{}^{\nmid}B_b|A{}^{\nmid}B_b)P(A{}^{|}B_b)$. Note that we have put a slash through the Picks in the first term of the right-hand side. We will call such Picks *neutered* because we are conditioning on the fact that an element was chosen from subset B_b, and thus no action is needed before selecting the element from A. Thus, the probability with a neutered Pick is the same as for a Be, e.g.,

$$P(A_a{}^{\nmid}B_b|A{}^{\nmid}B_b) = P(A_aB_b|AB_b) = \frac{\bar{n}_{a,b}}{\bar{n}_{,b}}. \tag{A9}$$

For example, the probability of picking a small jar and then picking a cashew *given* that one picked a small jar, is the same as picking a cashew *given* that one picked a small jar. So the Pick probabilities are

$$\begin{aligned}
&P(A{}^{|}B) = 1\\
&P(A{}^{|}B_b) = P(B_b),\\
&P(A_a{}^{|}B) = \sum_b P(A_a{}^{|}B_b) = \sum_b \frac{\bar{n}_{a,b}}{\bar{n}_{,b}}P(B_b),\\
&P(A_a{}^{|}B_b) = P(A_a{}^{\nmid}B_b|A{}^{\nmid}B_b)P(A{}^{|}B_b) = \frac{\bar{n}_{a,b}}{\bar{n}_{,b}}P(B_b).
\end{aligned} \tag{A10}$$

The astute reader may wonder why the selection on the leftmost set differs from the selection of the sets to its right. Actually, it does not, and we could put a "${}^{|}$" to the left of every leftmost set. But our notation *assumes* that there is a selection on the leftmost set. So really "${}^{|}$" means a selection done on a set other than the leftmost set. [Note that one can have a set to the left of an observer, and then one needs to insert a selection "${}^{|}$" to the left of the observers set, e.g., $C{}^{|}P$, where C are cards and P are observers, and although that observer is Be selected (i.e., just *is*), this is exclusive selection since there is a selection other than on the innermost set.]

Let us explore conditional probabilities, such as the ones we employed above, where there is one set of selections *given* another. Here are the nontrivial possibilities (keeping in mind that $P(A_aB|AB_b) = P(A_aB_b|AB_b)$, etc.):

(i) $P(A_aB|AB_b)$: the probability that we select an element of type a from A nested in B *given* that we select an element of A that corresponds to an element of B of type b.

(ii) $P(AB_b|A_aB)$: the probability that we select an element of A that corresponds to an element of B of type b *given* that we select an element of type a from A nested in B.

(iii) $P(A_a{}^{|}B|A{}^{|}B_b)$: the probability that we select an element of B and then select an element type a from A which is associated with that element of B *given* that we select an element of B of type b and then select an element of A associated with that element of B.

(iv) $P(A{}^{|}B_b|A_a{}^{|}B)$: the probability that we select an element of B of type b and then select an element of A associated with that element of B *given* that we select an element of B and then select an element type a from A which is associated with that element of B.

For example, $P(A{}^{|}B_S|A_c{}^{|}B)$ is the probability to pick a small jar and then pick a nut from that jar *given* that we pick some jar and then pick a cashew from it. There are actually only three nontrivial possibilities because the first and the third are equal since the selection in the third is neutered:

$$P(A_aB|AB_b) = P(A_a{}^{\nmid}B|A{}^{\nmid}B_b) = \frac{P(A_{a,b})}{P(A_{,b})} = \frac{\bar{n}_{a,b}}{\bar{n}_{,b}},$$
$$P(AB_b|A_aB) = \frac{P(A_{a,b})}{P(A_a)} = \frac{\bar{n}_{a,b}}{\bar{n}_a}P(B_b),$$
$$P(A{}^{|}B_b|A_a{}^{|}B) = \frac{P(A_a{}^{|}B_b)}{P(A_a{}^{|}B)} = \frac{\frac{\bar{n}_{a,b}}{\bar{n}_{,b}}P(B_b)}{\sum_{b'}\frac{\bar{n}_{a,b'}}{\bar{n}_{,b'}}P(B_{b'})}. \quad \text{(A11)}$$

In Eq. (A10) we showed that $P(A_a{}^{|}B_b)$ is not in general equal to $P(B_b)$, because the selection of an element of type a adds a nontrivial weighting factor. That is because there is an implied conditional $A{}^{|}B$: We take it as a given that we pick some element of B and then some element associated with that element from the whole set A, i.e., $P(A_a{}^{|}B_b)$ means $P(A_a{}^{|}B_b|A{}^{|}B)$. But sometimes we want to redefine the set A we select from so that it is some subset of qualifying elements. For example, if our jars contain peanuts, cashews, and pebbles, but our selection process ensures that only nuts are picked, then we are really concerned with the subset $A_{\rm nut}$ of cashews and peanuts. To help clarify such situations, we write redefined sets with square brackets $[A_{\rm re}]$. This new set then has subsets $[A_{\rm re}]_{a,b}$, and we can write the number of elements in these as $[n_{\rm re}]$ and $[n_{\rm re}]_{a,b}$, and so on. Now set $[A_{\rm re}]$ acts like A did in Eq. (A10),

$$P([A_{\rm re}]{}^{|}B) = 1,$$
$$P([A_{\rm re}]{}^{|}B_b) = P([A_{\rm re}]{}^{|}B_b|[A_{\rm re}]{}^{|}B) = P(B_b),$$
$$P([A_{\rm re}]_a{}^{|}B) = P([A_{\rm re}]_a{}^{|}B|[A_{\rm re}]{}^{|}B)$$
$$= \sum_b \frac{[\bar{n}_{\rm re}]_{a,b}}{[\bar{n}_{\rm re}]_{,b}}P(B_b),$$
$$P([A_{\rm re}]_a{}^{|}B_b) = P([A_{\rm re}]_a{}^{|}B_b|[A_{\rm re}]{}^{|}B)$$
$$= \frac{[\bar{n}_{\rm re}]_{a,b}}{[\bar{n}_{\rm re}]_{,b}}P(B_b), \quad \text{(A12)}$$

since one selects some element of $[A_{\rm re}]$ with certainty.

Now, one might object that there is a lot of redundant information in the above notation, namely the set labels A and B. We think it is important to retain those labels if there is any confusion about which sets are considered, which subset labels correspond to which set, and which sets have a Pick on them—an issue if there are more than two nested sets. But if there are only two nested sets which are the same throughout some calculation, and the subscript labels are unique to a set, we can use a compact notation by omitting the set names while

TABLE I. Summary of our major results using compact notation where the list of sets provides a key for the location of the Picks. Worlds $J = S$ or L. For three or more sets we use a double-Pick mark to avoid ambiguity. For "probing a multiverse" $h =$ in or ex (it is probably advisable not to use compact notation for four sets with controlled-Picks). The weighted averages are $\langle f(y)\rangle \equiv \int_0^1 f(y)\,p({}^{|}y)dy$, $\langle f(y)\rangle_{d^{(|)}} \equiv \int_0^1 f(y)\,p({}^{(|)}y|d^{(|)})dy$. For the Gott case we take $t_m = t$.

Section	Description	Sets	Input	Output	Result
II & IX	Be selection	$P_d{}^{|}W_J$	$P(S) = \frac{\bar{n}_S}{\bar{n}}P({}^{|}S)$	$P(S\|d) = P({}^{|}S)$	$R_{P/W} = 1$
III	Pick selection	$P_d{}^{|}W_J$	$P({}^{|}S)$	$P({}^{|}S\|d^{|}) = \frac{P({}^{|}S)}{P({}^{|}S)+\frac{1}{\rho}P({}^{|}L)}$	$R_{P^{|}/W} = \frac{1}{\rho}$
IV	Inclusive selection	$P_d{}^{\|}W_J{}^{|}E_y$	$P(S) = \frac{\bar{n}_S}{\bar{n}}P({}^{\|}S)$	$P(S\|d) = P({}^{\|}S) = \langle y\rangle$	$R^E_{P/W} = 1$
V	Exclusive selection	$P_d{}^{\|}W_J{}^{|}E_y$	$P(S^{|}) = \left\langle\frac{y}{\rho-(\rho-1)y}\right\rangle$	$P(S^{|}\|d^{|}) = \left\langle\frac{y}{\rho-(\rho-1)y}\right\rangle\left\langle\frac{1}{\rho-(\rho-1)y}\right\rangle^{-1}$	$R^{|E}_{P/W} \in \left[\frac{1}{\rho}, 1\right]$
VI.A & IX	Excl. theory selection	$P_dW{}^{|}\Theta_J$	$P({}^{|}S)$	$P({}^{|}S\|d^{|}) = \frac{P({}^{|}S)}{P({}^{|}S)+\frac{1}{\rho}P({}^{|}L)}$	$R_{P^{|}/\Theta} = \frac{1}{\rho}$
VI.B	Probing a multiverse	$P_dW\overleftarrow{{}^{\|}E_y}{}^{|}\Theta_h$	$P_h = P({}^{\|}\overleftarrow{{}^{|}h})$, $p = P({}^{\|}1^{\nmid}\|{}^{\|\nmid}\mathrm{ex})$	$P_{h\|d} = P({}^{\|}\overleftarrow{{}^{|}h}\|\,d^{\|}\overleftarrow{{}^{|}}\,)$	Prior-dependent
VII.E	Atypical freak pbservers	${}^{\xi}P_n$	$P(n) = \frac{1}{1+\rho}$	$P({}^{\xi}n) = \frac{1}{1+\kappa\rho}$	Need $\kappa\rho \ll 1$
VII.F	Rare observers	$P{}^{|}\Theta_J$	$p_J^{\geqslant 1} = 1-(1-p_{\mathcal{F}})^{n_J}$	$P({}^{|}J)_{\rm rare} = P(J)_{p_{\mathcal{F}}} = \frac{n_J}{\langle n\rangle}P({}^{|}J)$	Rare-Pick = Be
VII.F	Rare freak observers	$P_{0f}{}^{|}\Theta_J$	$P(0f^{|}\|{}^{|}J) = (1-p_f)^{n_J}$	$R^f_{P^{|}\Theta} = (1-p_f)^{n_L-n_S}R_\Theta$	$R^f_{P^{|}/\Theta} \to e^{-p_fn_L}$
VIII & IX	Incl. Gott &	$P_t{}^{|}W_T$	$p({}^{|}T)\sim\frac{1}{T^2} \Rightarrow p(T)\sim\frac{1}{T}$	$p(T\|t) \to \frac{t}{T^2}$	$R_T \to 1$
	No doomsday			$P((T > Kt)\|t) \to \frac{1}{K}$	$R_{\int T} \to 1$
VIII & IX	Excl. Gott &	$P_t{}^{|}W_T$	$p({}^{|}T)\sim\frac{1}{T}$	$p({}^{|}T\|t^{|}) \to \frac{t}{T^2}$	$R_{{}^{|}T} \to \frac{t}{T}$
	doomsday			$P({}^{|}(T > Kt)\|t^{|}) \to \frac{1}{K}$	$R_{\int{}^{|}T} \to \frac{1}{K}$
X	Incl. universal doomsday	$P_d{}^{\|}W_J{}^{|}E_y$	$p({}^{|}y) \Rightarrow \langle y\rangle$	$\langle y\rangle_d = P(S\|d) = \langle y\rangle$ (same as Sec. IV)	$R^{\rm UD}_{P/W} = R^E_{P/W}$
X	Excl. universal doomsday	$P_d{}^{\|}W_J{}^{|}E_y$	$p({}^{|}y) \Rightarrow \langle y\rangle$	$\langle y\rangle_{d^{|}} = P(S^{|}\|d^{|}) =$ (same as Sec. V)	$R^{|UD}_{P/W} = R^{|E}_{P/W}$
XI	Beauty thirder/incl.	$P_{\rm Day}{}^{|}W_{\rm flip}$	Three observer moments	$P(H) = \frac{1}{3}$	Need 2:1 odds
XI	Beauty halfer/excl.	$P_{\rm Day}{}^{|}W_{\rm flip}$	Two observer timelines	$P({}^{|}H) = \frac{1}{2}$	Need 2:1 odds
XI	Mon. Beauty thirder/incl.	$[P_{\rm Mon}]{}^{|}W_{\rm flip}$	Two observer moments	$P(H) = \frac{1}{2}$	Need 1:1 odds
XI	Mon. Beauty halfer/excl.	$[P_{\rm Mon}]{}^{|}W_{\rm flip}$	Two observer timelines	$P({}^{|}H) = \frac{1}{2}$	Need 1:1 odds

preserving the order of any subscript labels and selection bars:

$$P(\alpha\beta) \equiv P(A_\alpha B_\beta), \quad P(\alpha\,{}^{|}\beta) \equiv P(A_\alpha\,{}^{|}B_\beta), \tag{A13}$$

where α and β can be "null," e.g., $P(b|a) \equiv P(AB_b|A_aB)$ and $P({}^{|}b|a\,{}^{|}) \equiv P(A\,{}^{|}B_b|A_a\,{}^{|}B)$. For example, in compact notation, using Eq. (A8)–(A10),

$$P(b) \equiv P(AB_b) = \frac{\bar{n}_{,b}}{\bar{n}}P(B_b) = \frac{\bar{n}_{,b}}{\bar{n}}P({}^{|}b), \tag{A14}$$

and Bayes's law with a Pick is

$$P({}^{|}b|a\,{}^{|}) = \frac{P(a\,{}^{|}|\,{}^{|}b)P({}^{|}b)}{P(a\,{}^{|})}. \tag{A15}$$

We use the more verbose notation in most of the main text for clarity. Here are the terse versions: The posterior probability for a Be, Eq. (6), becomes

$$P(S|d) = \frac{P(d|S)P(S)}{P(d)} = P({}^{|}S), \tag{A16}$$

and the posterior probability for a Pick, Eq. (10), becomes

$$P({}^{|}S|d\,{}^{|}) = \frac{P(d\,{}^{\dagger}|\,{}^{\dagger}S)P({}^{|}S)}{P(d\,{}^{|})} = \frac{P({}^{|}S)}{P({}^{|}S) + \frac{1}{\rho}P({}^{|}L)}. \tag{A17}$$

We can use our compact formalism for three or more nested sets, but there is then an ambiguity about the location of the Pick. Does $P({}^{|}c)$ mean $P(A\,{}^{|}BC_c)$ or $P(AB\,{}^{|}C_c)$? To avoid this ambiguity, we use a double-Pick mark (and if needed a triple-Pick mark) on inner sets, so $P({}^{\|}c) \equiv P(A\,{}^{|}BC_c)$ and $P({}^{|}c) \equiv P(AB\,{}^{|}C_c)$. For example, the probabilities in Sec. V using sets $P_d\,{}^{\|}W_S\,{}^{|}E_y$ are

$$\begin{aligned} &P(d\,{}^{|}) \equiv P(P_dW\,{}^{|}E), \quad P(S\,{}^{|}) \equiv P(PW_S\,{}^{|}E),\\ &P({}^{|}y) \equiv P(PW\,{}^{|}E_y) = P(E_y),\\ &P({}^{\|}S) \equiv P(P\,{}^{|}W_SE) = P(W_SE). \end{aligned} \tag{A18}$$

We conclude with Table I, which summarizes our main results in compact notation.

[1] E. Noether, Invariante variationsprobleme, Nachr. Ges. Wiss. Goettingen, Math. Phys. Kl. **1918**, 235 (1918) [Invariant variation problems, Transport Theory and Statistical Physics **1**, 186 (1971)].

[2] B. Carter, Large number coincidences and the anthropic principle in cosmology, Symp. Int. Astronom. Union **63**, 291 (1974).

[3] S. Weinberg, The cosmological constant problem, Rev. Mod. Phys. **61**, 1 (1989).

[4] B. Carter and W. H. McCrea, The anthropic principle and its implications for biological evolution [and discussion], Philos. Trans. R. Soc. A **310**, 347 (1983).

[5] D. Dieks, Doomsday–Or: The dangers of statistics, Philos. Quart. **42**, 78 (1992).

[6] N. Bostrom, Investigations into the doomsday argument (unpublished).

[7] N. Bostrom, *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge, London, 2002).

[8] K. D. Olum, The doomsday argument and the number of possible observers, Philos. Quart. **52**, 164 (2002).

[9] N. Bostrom and M. M. Cirkovic, The doomsday argument and the self-indication assumption: Reply to Olum, Philos. Quart. **53**, 83 (2003).

[10] J. Knobe, K. D. Olum, and A. Vilenkin, Philosophical implications of inflationary cosmology, Br. J. Philos. Sci. **57**, 47 (2006).

[11] A. Gerig, K. D. Olum, and A. Vilenkin, Universal doomsday: Analyzing our prospects for survival, J. Cosmol. Astropart. Phys. 05 (2013) 013.

[12] J. Garriga and A. Vilenkin, Prediction and explanation in the multiverse, Phys. Rev. D **77**, 043526 (2008).

[13] S. M. Carroll, Why Boltzmann brains are bad, arXiv:1702.00850.

[14] J. R. G. III, Implications of the Copernican principle for our future prospects, Nature **363**, 315 (1993).

[15] This is not a prior probability for a world of type S, but rather the probability of being in a world of type S—so making use of the information of the fraction of observers in such worlds but prior to making use of rank information data.

[16] Scott Aaronson (private communication).

[17] J. B. Hartle and M. Srednicki, Are we typical? Phys. Rev. D **75**, 123523 (2007).

[18] M. Srednicki and J. Hartle, Science in a very large universe, Phys. Rev. D **81**, 123524 (2010).

[19] This is true almost by definition—we define an observer as a self-aware entity that can process external information (such as datum d). But it is an interesting question whether participation in the arrow of time is a requirement for consciousness—see Scott Aaronson, "Could a Quantum Computer Have Subjective Experience?", https://www.scottaaronson.com/blog/?p=1951.

[20] R. Penrose, Singularities and time-asymmetry, in *General Relativity, an Einstein Centennary Survey*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, UK, 1979).

[21] R. M. Wald, The arrow of time and the initial conditions of the universe, Stud. Hist. Philos. Sci. B **37**, 394 (2006).

[22] L. Boltzmann, Zu Hrn. Zermelo's Abhandlung: Ueber die mechanische Erklärung irreversibler Vorgänge, Ann. Phys. **296**, 392 (1897); translated in *Kinetic Theory*, edited by S. G. Brush (Oxford University Press, Oxford, 1966), p. 412.

[23] A. H. Guth, Inflationary universe: A possible solution to the horizon and flatness problems, Phys. Rev. D **23**, 347 (1981); A. A. Starobinsky, A new type of isotropic cosmological models without singularity, Phys. Lett. B **91**, 99 (1980); A. D. Linde, A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems, *ibid.* **108**, 389 (1982); A. Albrecht and P. J. Steinhardt, Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking, Phys. Rev. Lett. **48**, 1220 (1982).

[24] D. Baumann, Inflation, in *Physics of the Large and the Small: Proceedings of the Theoretical Advanced Study Institute in*

Elementary Particle Physics (TASI'09) (World Scientific Publishing Company, Hackensack, NJ, 2011), pp. 523–686.

[25] G. W. Gibbons and S. W. Hawking, Cosmological event horizons, thermodynamics, and particle creation, Phys. Rev. D **15**, 2738 (1977).

[26] L. S. Schulman, *Time's Arrows and Quantum Measurement* (Cambridge University Press, Cambridge, UK, 1997).

[27] L. Dyson, M. Kleban, and L. Susskind, Disturbing implications of a cosmological constant, J. High Energy Phys. 10 (2002) 011.

[28] A. Albrecht and L. Sorbo, Can the universe afford inflation? Phys. Rev. D **70**, 063528 (2004).

[29] S. Coleman, Fate of the false vacuum: Semiclassical theory, Phys. Rev. D **15**, 2929 (1977).

[30] D. N. Page, Is our universe likely to decay within 20 billion years? Phys. Rev. D **78**, 063535 (2008); Is our universe decaying at an astronomical rate? Phys. Lett. B **669**, 197 (2008).

[31] S. Carlip, Transient observers and variable constants or repelling the invasion of the Boltzmann's brains, J. Cosmol. Astropart. Phys. 06 (2007) 001.

[32] K. K. Boddy, S. M. Carroll, and J. Pollack, De Sitter space without dynamical quantum fluctuations, Found. Phys. **46**, 702 (2016).

[33] J. Hartle and T. Hertog, The observer strikes back, in *The Philosophy of Cosmology* (Cambridge University Press, New York, 2017), pp. 181–205; M. Srednicki and J. Hartle, The xerographic distribution: Scientific reasoning in a large universe, in *Proceedings of the 6th International Symposium on Quantum Theory and Symmetries (QTS609)* [J. Phys.: Conf. Ser. **462**, 012050 (2013)].

[34] D. N. Page, Space for both no-boundary and tunneling quantum states of the universe, Phys. Rev. D **56**, 2065 (1997).

[35] P. Buch, Future prospects discussed, Nature **368**, 107 (1994).

[36] J. R. Gott, Future prospects discussed—Gott Replies, Nature **368**, 108 (1994).

[37] C. M. Caves, Predicting future duration from present age: A critical assessment, Contemp. Phys. **41**, 143 (2000).

[38] Caves later [39] gave a geometric argument that the prior must go $\sim 1/T^2$ to obtain Gott's result, though he did not differentiate between a Pick and Be prior.

[39] C. M. Caves, Predicting future duration from present age: Revisiting a critical assessment of Gott's rule, arXiv:0806.3538.

[40] T. Bayes, Lii. an essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S., Philos. Trans. **53**, 370 (1763).

[41] A. Elga, Self-locating belief and the sleeping beauty problem, Analysis **60**, 143 (2000).

[42] D. Lewis, Sleeping beauty: Reply to Elga, Analysis **61**, 171 (2001).

[43] F. Arntzenius, Reflections on sleeping beauty, Analysis **62**, 53 (2002).

[44] J. Pust, Horgan on sleeping beauty, Synthese **160**, 97 (2008).

[45] D. Papineau and V. Durà-Vilà, A thirder and an Everettian: A reply to Lewis's 'Quantum sleeping beauty', Analysis **69**, 78 (2009).

[46] J. S. Rosenthal, A mathematical analysis of the Sleeping Beauty problem, Math. Intell. **31**, 32 (2009).

[47] T. Horgan, Synchronic Bayesian updating and the Sleeping Beauty problem: Reply to Pust, Synthese **160**, 155 (2008).

[48] N. Bostrom, Sleeping Beauty and self-location: A hybrid model, Synthese **157**, 59 (2007).

[49] R. M. Neal, Puzzles of anthropic reasoning resolved using full non-indexical conditioning, arXiv:math/0608592.