Testing typicality in multiverse cosmology

Feraz Azhar[*]

Department of History and Philosophy of Science, University of Cambridge,
Free School Lane, Cambridge, CB2 3RH, United Kingdom

In extracting predictions from theories that describe a multiverse, we face the difficulty that we must assess probability distributions over possible observations, prescribed not just by an underlying theory, but by a theory together with a conditionalization scheme that allows for (anthropic) selection effects. This means we usually need to compare distributions that are consistent with a broad range of possible observations, with actual experimental data. One controversial means of making this comparison is by invoking the 'principle of mediocrity': that is, the principle that we are typical of the reference class implicit in the conjunction of the theory and the conditionalization scheme. In this paper, I quantitatively assess the principle of mediocrity in a range of cosmological settings, employing 'xerographic distributions' to impose a variety of assumptions regarding typicality. I find that for a fixed theory, the assumption that we are typical gives rise to higher likelihoods for our observations. If, however, one allows both the underlying theory and the assumption of typicality to vary, then the assumption of typicality does not always provide the highest likelihoods. Interpreted from a Bayesian perspective, these results support the claim that when one has the freedom to consider different combinations of theories and xerographic distributions (or different 'frameworks'), one should favor the framework that has the highest posterior probability; and then from this framework one can *infer*, in particular, how typical we are. In this way, the invocation of the principle of mediocrity is more questionable than has been recently claimed.

I. INTRODUCTION

A generic prediction of the theories that underpin our current understanding of the large-scale structure of the universe, is that the observable universe is not all that exists, and that we may be part of a vast landscape of (as yet) unobserved domains where the fundamental constants of nature, and perhaps the effective laws of physics more generally, vary. The predominant approach to characterizing this variability rests on theory-generated probability distributions that describe the statistics of constants associated with the standard models of particle physics and cosmology. The hope remains that plausible descriptions of such multi-domain universes (henceforth 'multiverses'), generated, for example, from inflationary cosmology [1–3] or the string theory landscape [4–7], will yield prescriptions for calculating these distributions in unambiguous ways. Subsequent comparisons with our observations would allow us to ascertain which multiverse models are indeed favored.

To be more precise, one expects theories that describe a multiverse to set down a likelihood for observations we might make, given both the theory under consideration, and conditions that restrict the vast array of domains to ones in which we might arise. This latter conditionalization is naturally couched in terms of conditions necessary for the existence of 'us', as defined by relevant features of the theory. The need for such 'anthropic' conditionalization, as captured, for example, in what has become known as Carter's 'Weak Anthropic Principle' [8], is predicated on the presumption that most of the domains described by theories of the multiverse will not give rise to the specialized structures we see around us, nor indeed to complex biological life [9]. Under this scenario any observation we might make conditionalized on theory alone, would prove to be unlikely; and one should therefore restrict one's attention to relevant domains so as to secure relevant probabilities for possible observations.

An appropriate conditionalization scheme might make our observations more likely: but how likely should they be, before we can count them as having been successfully predicted by the conjunction of a theory and a conditionalization scheme? One proposed solution to this problem is known as the 'principle of mediocrity' [10]: in more current terminology, it assumes that we should reason as though we are typical members of a suitable reference class (see also [11–13]). Under this assumption, for appropriately conditionalized distributions, as long as our observations are within some 'typical' range according to the distribution, we can count them as being successfully predicted. The assessment of this assumption is the topic of this paper, and constitutes a central concern in extracting predictions from any theory of the multiverse.

The principle of mediocrity, or the 'assumption of typicality'—as it will also be referred to in this paper—is not without its critics [14–16]. A key issue involves how one defines the reference class with respect to which we are typical [17]. This problem is even more stark given our ignorance of who or what we are trying to characterize, and the precise physical constraints we need to implement in order to do so [14, 18].

Rather than assessing this principle from a primarily conceptual point of view, we propose to test it *quantitatively*. In particular, we investigate how well it does in

terms of accounting for our data in comparison to other assumptions regarding typicality, in a restricted set of multiverse cosmological scenarios. We do this by extending the program of Srednicki and Hartle [19] to explore a variety of assumptions regarding typicality, building these assumptions into likelihoods for possible observations through the use of 'xerographic distributions' (in the terminology of Srednicki and Hartle [19]). The goal then is to find the conjunction of a theory and a xerographic distribution (which they call a 'framework') that gives rise to the highest likelihoods for our data.

I will show that (1) for a fixed theory, the assumption that we are typical gives rise to higher likelihoods for our observations; but (2) if one allows both the underlying theory and the assumption of typicality to vary, then the assumption of typicality *does not* always provide the highest likelihoods. Interpreted from a Bayesian perspective, these results provide support for the claim that one should try to identify the framework with the highest posterior probability; and then from this framework, one can *infer* how typical we are.

The structure of this paper is as follows. In section II, I outline the general formalism within which I will be investigating assumptions regarding typicality, including the introduction of a statement of the principle of mediocrity adapted to our specific purposes. Section III introduces the multiverse model we will analyze (which is a generalization of the cosmological model of Srednicki and Hartle [19]), derives the central equations for relevant likelihoods from which we will eventually test assumptions regarding typicality, and shows that these likelihoods reduce to the results of Srednicki and Hartle [19] under the appropriate simplifying assumptions. Explicit tests of the principle of mediocrity are presented in section IV, and we conclude in section V with a discussion of the context in which one should interpret the results of these tests. So I turn first to a description of the general formalism within which I will be working.

II. XEROGRAPHIC DISTRIBUTIONS

A. Generalities

I begin by outlining the formalism of Srednicki and Hartle [19], recasting relevant parts of their discussion to suit our computations in the next section.

In general multiverse scenarios, it is possible that any reference class of which we believe we are a member, may have multiple members. Indeed, it is plausible that our accumulated data D_0, which gives a detailed description of our physical surroundings, might be replicated at different spacetime locations in the multiverse. A theory $\mathcal{T}$ describing this multiverse scenario, will, in principle, generate a likelihood for this data which we will denote by $P(D_0|\mathcal{T})$. This corresponds to a 'third-person' likelihood in the terminology of Srednicki and Hartle [19]—that is, a likelihood that does not include any information about

which member of our reference class we might be. The quantity that takes this *indexical* information into account is a 'first-person' likelihood and will be denoted by $P^{(1p)}(D_0|\mathcal{T},\xi)$, in accordance with the notation of Srednicki and Hartle [19]. The added ingredient here is the *xerographic distribution* ξ, a probability distribution that we specify *by assumption*, that encodes our belief about which member of our reference class we happen to be. Its functional form is independent of a given theory $\mathcal{T}$, and together with such a theory, constitutes a 'framework' $(\mathcal{T},\xi)$ (in the notation of [19]). Thus the transition from a third-person likelihood $P(D_0|\mathcal{T})$ to a first-person likelihood $P^{(1p)}(D_0|\mathcal{T},\xi)$ is effected by two ideas: (i) the conditionalization scheme, which (as mentioned in section I) specifies our reference class, and (ii) a probability distribution over members of our reference class.

In the case where there exist L members of our reference class at spacetime locations x_l for $l = 1, 2, \ldots, L$, we let the probability that we are the member at location x_l be denoted by ξ_l. So the xerographic distribution is just the sequence of probabilities $\xi := \{\xi_l\}_{l=1}^{L}$, and will always be chosen so that it is normalized to unity: $\sum_{l=1}^{L} \xi_l = 1$. We will assume throughout this paper that the total number of members L is finite.

The assumption that we are a typical member of this reference class, is then the statement that the probability that we are any one of these members is the same, and thus the xerographic distribution is given by the uniform distribution: $\xi_l = \frac{1}{L}$. Correspondingly, the assumption that we are atypical of this reference class will be given by xerographic distributions that deviate from the uniform distribution.

How then do we propose to compute the first-person likelihood $P^{(1p)}(D_0|\mathcal{T},\xi)$? To do so, we will introduce a few conventions that will form the basis of the general discussion here, and also the basis of the more specific examples we will pursue in the following two sections.

Assume then that there exists a finite set of N distinct domains in a multiverse, within each of which 'observers' may exist with some probability that may depend on the particular domain being considered. We will only track the existence or non-existence of observers in each domain, without concerning ourselves with issues such as: whether or not these observers are anything like 'us'[1], precisely where in these domains these observers might be located, and how many observers might exist in a domain. We thereby consider only whether a domain has observers in it or does not. Thus, there exists a total of 2^N possible configurations of observers in domains across the entire multiverse. We will denote each such configuration by an N-dimensional vector $\vec{\sigma}$ of binary digits,

---

[1] The simplicity of the scenarios we will consider makes this less egregious an assumption than it would otherwise be. Indeed, as mentioned in section I, definitions of observers and the subsequent specification of appropriate reference classes is a thorny issue, but we will not need to engage with it here (see [16] for further discussion).

where $\sigma_i = 1$ denotes the existence of observers in domain i, and $\sigma_i = 0$ denotes there are no observers in that domain. The set of all such configurations $\vec{\sigma}$ will be denoted by $\mathcal{K}$, and we will denote the probability of a configuration $\vec{\sigma}$ by $P(\vec{\sigma})$.

We live inside one of these domains and observe some data. Let D_0 denote the data that there exist observers who see this same data. We will take a theory $\mathcal{T}$ to describe an *observable* fact about each domain: in the model introduced in section III, this will be the value of a binary quantity. So the probabilities $P(\vec{\sigma})$ will be given in general, independently of $\mathcal{T}$. But $\mathcal{T}$ will determine the subset $\mathcal{K}_{D_0}(\mathcal{T})$ of those configurations $\vec{\sigma}$ in which D_0 is observed ($\mathcal{K}_{D_0}(\mathcal{T}) \subset \mathcal{K}$). The sum of the probabilities of the configurations $\vec{\sigma}$ belonging to the subset $\mathcal{K}_{D_0}(\mathcal{T})$, is just the theory-generated third-person likelihood for our data: $P(D_0|\mathcal{T}) = \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(\mathcal{T})} P(\vec{\sigma})$. This specifies the likelihood that there is at least one observing system in the multiverse that witnesses the data D_0.

To derive a first-person likelihood, we need to assume a xerographic distribution by firstly specifying a reference class that could plausibly describe 'us'. Two natural reference classes to consider (which we will further develop in section III), following [19], are (i) the reference class of all observers who witness our data D_0, and (ii) the reference class of all observers (who do not necessarily see our data D_0). In either case, for any particular possible observer configuration $\vec{\sigma}$, the xerographic distribution encodes the probability that we are the reference-class member at some specified location. Owing to the simplicity of our model, a 'location' will correspond to a multiverse domain. In general, for a given $\mathcal{T}$ and given $\vec{\sigma} \in \mathcal{K}_{D_0}(\mathcal{T})$, only a subset of locations $L_{D_0}(\vec{\sigma}, \mathcal{T})$, will contain observers who see our data D_0. From these considerations, we can write down the appropriate first-person likelihood $P^{(1p)}(D_0|\mathcal{T}, \xi)$ as follows:

$$P^{(1p)}(D_0|\mathcal{T}, \xi) := \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(\mathcal{T})} P(\vec{\sigma}) \sum_{l' \in L_{D_0}(\vec{\sigma}, \mathcal{T})} \xi_{l'} . \quad (1)$$

To be clear, *for each configuration $\vec{\sigma}$* in the above sum over configurations, one needs to compute the appropriately normalized xerographic distribution, subsequently summing that distribution over only those locations that could indeed correspond to us.[2]

We note that for the reference class of all observers who witness our data D_0 (called (i) above): for each theory $\mathcal{T}$ and each configuration $\vec{\sigma} \in \mathcal{K}_{D_0}(\mathcal{T})$, the subset of locations that contain observers who witness our data D_0,

─────────

[2] The notation adopted in Eq. (1) is not to be confused with the claim that the theory in any way *determines* the xerographic distribution—it does not. We are simply spelling out how our assumptions regarding which possible member of a suitable reference class we might be, enter into the determination of first-person likelihoods. This will become clearer when we consider concrete cosmological scenarios below—see Eq. (5) for a preview.

namely $L_{D_0}(\vec{\sigma}, \mathcal{T})$, is the entirety of the set of locations over which the xerographic distribution can be nonzero. Thus, by the normalization condition the xerographic distribution satisfies, we have, in Eq. (1): $\sum_{l' \in L_{D_0}(\vec{\sigma}, \mathcal{T})} \xi_{l'} = 1$. For this reference class therefore, the first-person likelihood is independent of the functional form of the xerographic distribution, and reduces to the appropriate third-person likelihood: $P^{(1p)}(D_0|\mathcal{T}, \xi) \longrightarrow P(D_0|\mathcal{T})$.

B. Preferred xerographic distributions

The likelihood introduced in Eq. (1) can be analyzed from a Bayesian perspective, which, under the appropriate conditions, allows us to pick out a preferred xerographic distribution. As detailed by Srednicki and Hartle [19], one can compute the posterior probability $P^{(1p)}(\mathcal{T}, \xi|D_0)$ by Bayes' theorem

$$P^{(1p)}(\mathcal{T}, \xi|D_0) = \frac{P^{(1p)}(D_0|\mathcal{T}, \xi)P(\mathcal{T}, \xi)}{\sum_{(\mathcal{T}, \xi)} P^{(1p)}(D_0|\mathcal{T}, \xi)P(\mathcal{T}, \xi)}, \quad (2)$$

where $P(\mathcal{T}, \xi)$ is the prior probability of the framework $(\mathcal{T}, \xi)$. We will be working below in a simplified setting where we assume we have a set of equally plausible frameworks at our disposal. That is, each framework enters into our Bayesian analysis with an equal prior. This implies the posterior probability is simply proportional to the likelihood:

$$P^{(1p)}(\mathcal{T}, \xi|D_0) \propto P^{(1p)}(D_0|\mathcal{T}, \xi), \quad (3)$$

and the question of which framework is to be preferred (by virtue of having the highest posterior probability), becomes a question of which framework gives rise to the highest likelihood $P^{(1p)}(D_0|\mathcal{T}, \xi)$. It is from this preferred framework that one can select a preferred xerographic distribution. Given that we will use these distributions to encode assumptions regarding typicality, this selection will allow us to infer, in particular, how typical we are.

C. Testing mediocrity

How then do we propose to test the principle of mediocrity? In the language introduced above, we can formulate a more precise version of the principle of mediocrity as follows:

> **PM**: We are typical of the entirety of the reference class of observers in the multiverse who measure our data D_0. That is, in cases where there are (a finite number of) L observers who measure D_0, situated at spacetime locations x_l for $l = 1, 2, \ldots, L$, the probability that we are any one of these observers is $\frac{1}{L}$. The corresponding xerographic distribution is given by $\xi_l = \frac{1}{L}$, for $l = 1, 2, \ldots, L$.

In what follows, we will denote the xerographic distribution implementing **PM** by ξ^{PM}. Any other xerographic distribution, with either a non-uniform distribution over the reference class referred to in **PM**, or else any distribution over a reference class that is not the one referred to there, constitutes some form of *non-mediocrity*.

With the assumptions of section II B in mind, we propose to test the principle of mediocrity by comparing likelihoods $P^{(1p)}(D_0|\mathcal{T}, \xi^{\mathrm{PM}})$ against $P^{(1p)}(D_0|\mathcal{T}^\star, \xi^\star)$, where $\xi^\star \neq \xi^{\mathrm{PM}}$, and where we allow for the possibility that the underlying theory $\mathcal{T}$ can vary as well.

III. EXTENDING THE COSMOLOGICAL MODEL OF SREDNICKI AND HARTLE

To explicate the schema of section II, and to probe the plausibility of **PM** in more concrete settings, we now construct a generalization of the cosmological toy model presented by Srednicki and Hartle [19] (see also their earlier paper [16]). In section III D, we will demonstrate that our results for likelihoods for this extended model reproduce theirs under the appropriate simplifying assumptions.

A. Model preliminaries

Let $\mathcal{V} = \{1, 2, \ldots, N\}$ label N distinct domains in a multiverse, each of which is assumed to have one of two 'colors', *red* or *blue*, corresponding to two possible values of some physical observable. The precise interpretation of this observable will not matter, and we will rely only on the fact of it taking two distinct values.

Observers may exist in these domains with some probability, where we assume this probability is independent of the color of any domain. As outlined in section II A, we characterize this probability by first introducing a vector of observer occupancy (or a 'configuration') via the notation $\vec{\sigma} := (\sigma_1, \sigma_2, \ldots, \sigma_N)$, where, for $i = 1, 2, \ldots, N$, $\sigma_i = 1$ denotes the existence of observers in domain i, and $\sigma_i = 0$ denotes there are no observers in domain i. There will, of course, be 2^N such configurations $\vec{\sigma}$, the set of which we label $\mathcal{K}$.

We will further assume that the probability of observers in a domain is independent of that for all the other domains, but also that these probabilities are *not* generally the same.[3] This implies that the probability of $\vec{\sigma}$, denoted by $P(\vec{\sigma})$, factorizes into a product of

―――――

[3] The general results below (in section IV) do not depend on any assumption that the different domains receive independent but unequal (or even equal) probabilities. However, it is natural to assume that (a) the existence of observers in different domains constitute probabilistically independent events (e.g., due to separate processes of evolution), and (b) that these probabilities can be unequal, reflecting the fact that different domains can vary in their hostility to life.

marginals $P_i(\sigma_i)$: $P(\vec{\sigma}) = \prod_{i=1}^N P_i(\sigma_i)$. If we let p_i denote the probability of the existence of observers in domain i, then $(1 - p_i)$ is the probability of no observers in that domain, and since $\sigma_i = 1$ or 0, we can write $P_i(\sigma_i) = p_i^{\sigma_i}(1 - p_i)^{1 - \sigma_i}$, giving

$$P(\vec{\sigma}) = \prod_{i=1}^N p_i^{\sigma_i}(1 - p_i)^{1 - \sigma_i}. \tag{4}$$

As to our data, we assume: we exist within one of these domains and observe the color red. That is, our data D_0 is

D_0: There exists a domain with observers in it who see *red*.

To write down the first-person likelihood for our data in accord with Eq. (1), we also need to specify the theories and the xerographic distributions we are interested in.

The theories we consider are ones that specify the color of each of the N domains. In particular, each theory will be denoted by a vector $\mathcal{T} = (T_1, T_2, \ldots, T_N)$, where, for $i = 1, 2, \ldots, N$, $T_i = 1$ when the theory predicts that domain i is red, and $T_i = 0$ when the theory predicts it is blue.

The xerographic distributions we consider will be defined to take the value zero, for each domain outside some nonempty subset (among all possible (nonempty) subsets) of domains $\mathcal{V} = \{1, 2, \ldots, N\}$ in the multiverse. To fix notation, if we let $\mathcal{C}$ denote the power set of $\mathcal{V}$ excluding the empty set, i.e., $\mathcal{C} = 2^{\mathcal{V}} \backslash \emptyset$, then a nonempty subset of domains $c \in \mathcal{C}$, outside which a xerographic distribution must be zero, will be denoted by $c = \{v_1, v_2, \ldots, v_M\}$ (where, of course, $1 \leq M \leq N$). The v_α's are simply integers labeling different domains in $\mathcal{V}$. We will explicitly write xerographic distributions as ξ_c, with the subscript c indicating the subset over which they can be nonzero (so this subset c is a superset of the support of a xerographic distribution). For the sake of simplicity, we will assume that the elements of c are listed in increasing order, though nothing physical in what follows will depend on this.

Then for each possible c, our xerographic distributions will fall into two classes:

C1 The first class of xerographic distributions assumes that we are typical among instances of observers who see our data D_0, and will be denoted by ξ_c^{typD}.

C2 The second class of xerographic distributions assumes we are typical among instances of observers, regardless of what data they see (i.e., regardless of whether they measure red or blue for their particular domain). These distributions will be denoted by ξ_c^{typO}.

Note that only *one* of the xerographic distributions imposes **PM**, namely, $\xi^{\mathrm{PM}} := \xi_{\mathcal{V}}^{\mathrm{typD}}$. That is, the principle of mediocrity is represented by the distribution that assumes we are typical among *all* instances of our data

over the multiverse (note that $c = \mathcal{V}$ for this distribution). Under the assumptions laid out in section II C, any other xerographic distribution encodes some form of *non-mediocrity*.

With these preliminaries in mind, we can construct the analog of Eq. (1) for our multiverse model. We will do so separately for **C1** and **C2** above. We turn first to **C2**: where we will construct the first-person likelihood for our data D_0, assuming we are typical among instances of observers, regardless of their data, as encoded by ξ_c^{typO}.

B. Construction of $P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typO}})$

As indicated in the outer sum of Eq. (1), the definition of the first-person likelihood for our data D_0, sums over the subset $\mathcal{K}_{D_0}(\mathcal{T})$ of configurations that contain our data, according to theory $\mathcal{T}$. Keeping in mind that any xerographic distribution must be zero outside a nonempty subset $c = \{v_1, v_2, \ldots, v_M\}$ of multiverse domains $\mathcal{V}$, we let $\mathcal{K}_{D_0}(c, \mathcal{T}) \subset \mathcal{K}$ denote the corresponding subset, i.e. the subset of all configurations $\vec{\sigma}$ that contain instances of our data D_0, according to theory $\mathcal{T}$, within the domains specified by c.

Next we need to characterize the xerographic distribution ξ_c^{typO}. For a configuration $\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})$, the total number of observers within $c = \{v_1, v_2, \ldots, v_M\}$ is simply $\sum_{\alpha=1}^{M} \sigma_{v_\alpha}$; and the xerographic distribution corresponding to ξ_c^{typO}, assigns to the location x_l of each observer-system in c, the value $\xi_l = 1/(\sum_{\alpha=1}^{M} \sigma_{v_\alpha})$ (recalling that a location in our model is simply a multiverse domain). In the computation of the first-person likelihood, we sum the xerographic distribution over only those locations that contain instances of our data D_0. So to that end, following our notation in Eq. (1), for each configuration $\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})$, let $L_{D_0}(c, \vec{\sigma}, \mathcal{T})$ denote the subset of locations in c, where our data D_0 exists, according to $\mathcal{T}$. Note that the total number of instances of our data in c is the just the size of this set $|L_{D_0}(c, \vec{\sigma}, \mathcal{T})|$. This can be written as $|L_{D_0}(c, \vec{\sigma}, \mathcal{T})| = \sum_{\beta=1}^{M} \sigma_{v_\beta} T_{v_\beta}$: the dummy variable v_β on the right hand side shows the c-dependence; and although there is no explicit D_0 dependence, recall that $T_{v_\beta} = 1$ iff domain v_β is red, which, when $\sigma_{v_\beta} = 1$, corresponds to the existence of our data D_0 in domain v_β.

We can now calculate a closed-form expression for $P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typO}})$, by directly adapting Eq. (1) for the particular case considered here:

$$
\begin{aligned}
P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typO}}) &= \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})} P(\vec{\sigma}) \sum_{l' \in L_{D_0}(c, \vec{\sigma}, \mathcal{T})} \xi_{l'} \\
&= \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})} \prod_{i=1}^{N} p_i^{\sigma_i} (1 - p_i)^{1 - \sigma_i} \sum_{l' \in L_{D_0}(c, \vec{\sigma}, \mathcal{T})} \xi_{l'} \\
&= \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})} \prod_{i=1}^{N} p_i^{\sigma_i} (1 - p_i)^{1 - \sigma_i} \sum_{l' \in L_{D_0}(c, \vec{\sigma}, \mathcal{T})} \frac{1}{\sum_{\alpha=1}^{M} \sigma_{v_\alpha}} \\
&= \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})} \prod_{i=1}^{N} p_i^{\sigma_i} (1 - p_i)^{1 - \sigma_i} \frac{1}{\sum_{\alpha=1}^{M} \sigma_{v_\alpha}} |L_{D_0}(c, \vec{\sigma}, \mathcal{T})| \\
&= \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})} \prod_{i=1}^{N} p_i^{\sigma_i} (1 - p_i)^{1 - \sigma_i} \left(\frac{\sum_{\beta=1}^{M} \sigma_{v_\beta} T_{v_\beta}}{\sum_{\alpha=1}^{M} \sigma_{v_\alpha}} \right);
\end{aligned} \tag{5}
$$

where we have used Eq. (4) to obtain the second line, and the fact that the value of the xerographic distribution is the same at each location $l' \in L_{D_0}(c, \vec{\sigma}, \mathcal{T})$ to obtain the fourth line.

C. Construction of $P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typD}})$

We turn next to **C1**, and the construction of the first-person likelihood for our data D_0, assuming we are typical among instances of our data, as encoded by ξ_c^{typD}. Having completed the above construction, it is straightforward to compute this. For a configuration $\vec{\sigma} \in$

$\mathcal{K}_{D_0}(c, \mathcal{T})$, since the total number of observers who see our data within c is $\sum_{\alpha=1}^{M} \sigma_{v_\alpha} T_{v_\alpha}$, the xerographic distribution corresponding to ξ_c^{typD} is $\xi_l = 1/(\sum_{\alpha=1}^{M} \sigma_{v_\alpha} T_{v_\alpha})$ for each location x_l of our data in c. The term that imposes the xerographic distribution in the computation of the first-person likelihood, namely $\sum_{l' \in L_{D_0}(c, \vec{\sigma}, \mathcal{T})} \xi_{l'}$, is then simply unity: $\sum_{l' \in L_{D_0}(c, \vec{\sigma}, \mathcal{T})} \xi_{l'} = 1$. Hence, we find

$$
P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typD}}) = \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})} \prod_{i=1}^{N} p_i^{\sigma_i} (1 - p_i)^{1 - \sigma_i}. \tag{6}
$$

Note that we recover the analog of the claim in the last paragraph of section II A: that for **C1**, the first-person likelihood $P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\mathrm{typD}})$, is equal to the appropriate third-person likelihood—which expresses the likelihood that there is at least one observing system that witnesses our data D_0, among the multiverse domains specified by c.

Equations (5) and (6) are the appropriate generalizations of the likelihoods calculated for the cyclic cosmological model of Srednicki and Hartle [19]. In order to make this connection more precise, and to set the stage for some of the computations described in section IV, we now show that in the case where our multiverse model reduces to the cosmological model of Srednicki and Hartle [19], Eqs. (5) and (6) indeed reduce to their likelihoods.

D. A multiverse of cycles

Consider, then, the case where the N domains of the multiverse model introduced in section III A are stretched out in time, so that the index i that labels a domain, corresponds to the order in which the domain occurs in what we will refer to as a 'multiverse of cycles'. Assume further that the xerographic distribution can be nonzero for each of these domains, so that $c = \mathcal{V}$. Furthermore, assume, as in [19], that the probability of observers existing in each of these cycles is the same (as well as independent of whether they exist in other cycles), and is given by $p_i = p$. We wish to first calculate the likelihood of our data under the assumption that we are typical among all instances of observers, i.e., $P^{(1p)}(D_0|\mathcal{T}, \xi_{\mathcal{V}}^{\mathrm{typO}})$. Under these assumptions, Eq. (5) reduces to

$$P^{(1p)}(D_0|\mathcal{T}, \xi_{\mathcal{V}}^{\mathrm{typO}}) = \sum_{\vec{\sigma} \in \mathcal{K}_{D_0}(\mathcal{V}, \mathcal{T})} p^{\sum_{i=1}^{N} \sigma_i} (1-p)^{N - \sum_{j=1}^{N} \sigma_j} \left(\frac{\sum_{m=1}^{N} \sigma_m T_m}{\sum_{n=1}^{N} \sigma_n} \right). \tag{7}$$

The large amount of symmetry in this expression allows us to organize the outermost sum by separately considering configurations $\vec{\sigma}$ according to the total number of observers $\sum_{i=1}^{N} \sigma_i$ in each configuration. We will call the total number of observers in each configuration n_O, and note that this will range from 1 to N for consistency with the fact that there exists at least one cycle with observers, that is, for consistency with our data (recall that we are not counting how many individuals are in each domain, just whether domains house observers or not). Equation (7) can then be thought of

as $P^{(1p)}(D_0|\mathcal{T}, \xi_{\mathcal{V}}^{\mathrm{typO}}) = \sum_{n_O=1}^{N} F(p, n_O, N_R)$, for some function F we will compute below, where N_R denotes the total number of red cycles predicted by the theory $\mathcal{T}$. There are two obvious cases to consider in this sum: namely, $1 \leq n_O \leq N_R$ and $N_R < n_O \leq N$, and we formally treat these two cases separately.

For $1 \leq n_O \leq N_R$, one can generate an expression for $F(p, n_O, N_R)$ by sequentially considering all possible numbers of observers (out of a maximum of n_O observers) placed in N_R red cycles. In general, one finds

$$\begin{aligned} F(p, n_O, N_R) &= p^{n_O}(1-p)^{N-n_O} \frac{1}{n_O} \left[\binom{N_R}{n_O} n_O + \binom{N_R}{n_O-1}\binom{N-N_R}{1}(n_O-1) + \cdots + \binom{N_R}{1}\binom{N-N_R}{n_O-1} \right] \\ &= p^{n_O}(1-p)^{N-n_O} \frac{1}{n_O} \sum_{k=0}^{n_O-1} \binom{N_R}{n_O-k}\binom{N-N_R}{k}(n_O-k) \\ &= p^{n_O}(1-p)^{N-n_O} \frac{1}{n_O} N_R \sum_{k=0}^{n_O-1} \binom{N-N_R}{k}\binom{N_R-1}{n_O-1-k} \\ &= p^{n_O}(1-p)^{N-n_O} \frac{1}{n_O} N_R \binom{N-1}{n_O-1} \\ &= p^{n_O}(1-p)^{N-n_O} N_R \frac{1}{N} \binom{N}{n_O}, \end{aligned} \tag{8}$$

where we have used a standard binomial identity to obtain the third and fifth lines, together with Vandermonde's identity to obtain the fourth one.

In the case where $N_R < n_O \leq N$, it is not difficult to show that we obtain precisely the same final result using a similar sequence of steps. Putting this all together,

gives

$$P^{(1p)}(D_0|\mathcal{T}, \xi_{\mathcal{V}}^{\text{typO}}) = \frac{N_R}{N} \sum_{n_O=1}^{N} \binom{N}{n_O} p^{n_O}(1-p)^{N-n_O}$$

$$= \frac{N_R}{N} \left[1 - (1-p)^N\right], \qquad (9)$$

agreeing with the corresponding expression (Eqs. (5.8) and (B5)) in [19].

A similar calculation (presented in the appendix) gives the correct formula for $P^{(1p)}(D_0|\mathcal{T}, \xi^{\text{PM}})$ (recall that $\xi^{\text{PM}} := \xi_{\mathcal{V}}^{\text{typD}}$), where

$$P^{(1p)}(D_0|\mathcal{T}, \xi^{\text{PM}}) = 1 - (1-p)^{N_R}, \qquad (10)$$

corresponding to Eq. (5.5) in [19]. In this way, the generalized multiverse model introduced in section III A reduces to the cyclic model of Srednicki and Hartle [19] under the appropriate simplifying assumptions.

IV. EVALUATING THE PRINCIPLE OF MEDIOCRITY

With the results of the last section in hand, we can now address the central conceptual task of this paper: namely, to assess the predictive power (understood in Bayesian terms) of the principle of mediocrity, under the scheme introduced in section II. We will develop the answers to three central questions in turn: (A) What framework produces the highest likelihoods? (B) For a fixed theory, what xerographic distribution gives rise to the framework with the highest likelihoods? (C) Does the principle of mediocrity generally provide the framework with the highest likelihoods?

A. The best performing framework

To begin our analysis, we show that the framework whose likelihood attains the supremum of the likelihoods for all frameworks considered, and for all assignments of probabilities $\{p_i\}_{i=1}^{N}$ to multiverse domains, is $(\mathcal{T}_{\text{all red}}, \xi^{\text{PM}})$: that is, the theory that predicts each of the domains is red ($\mathcal{T} = \mathcal{T}_{\text{all red}}$), together with the xerographic distribution corresponding to the principle of mediocrity ($\xi = \xi^{\text{PM}} := \xi_{\mathcal{V}}^{\text{typD}}$). For this framework, the likelihood attained is the same as that with a xerographic distribution corresponding to typicality across all observers for $c = \mathcal{V}$, that is, for $\xi = \xi_{\mathcal{V}}^{\text{typO}}$.

To show this, let the p_i's be arbitrarily chosen but fixed. Note first that for arbitrary c and $\mathcal{T}$, Eqs. (5) and (6) imply $P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typO}}) \leq P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typD}})$: as the same configurations $\vec{\sigma} \in \mathcal{K}_{D_0}(c, \mathcal{T})$ contribute to the sums in Eqs. (5) and (6), but in the case of Eq. (5), each contributing term is multiplied by a factor that is less than or equal to 1. Now, the *maximal* value of $P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typD}})$ is attained when we choose c and

$\mathcal{T}$ such that $\mathcal{K}_{D_0}(c, \mathcal{T})$ includes the maximal number of configurations in the sum over (manifestly non-negative) probabilities of configurations. This will occur for the case where all multiverse domains are included ($c = \mathcal{V}$) and for the theory that predicts that all domains are red. Note finally that for $\mathcal{T} = \mathcal{T}_{\text{all red}} \equiv (1, 1, \ldots, 1)$, $\sum_{\beta=1}^{M} \sigma_{v_\beta} T_{v_\beta} = \sum_{\alpha=1}^{M} \sigma_{v_\alpha}$ (regardless of c), and so the final term in brackets in Eq. (5) is unity; hence the likelihood for the framework $(\mathcal{T}_{\text{all red}}, \xi^{\text{PM}})$, coincides with the likelihood for $(\mathcal{T}_{\text{all red}}, \xi_{\mathcal{V}}^{\text{typO}})$.

To see how these conclusions manifest in a particular setting, consider the multiverse of cycles introduced in section III D, which assumes that the probability of observers existing in any cycle is the same for each cycle and is given by p. We will consider the case where the xerographic distribution can be nonzero only on an initial segment of cycles starting with the first cycle and ending with some terminal cycle (after which the xerographic distribution is assumed to be zero). So for any number of cycles N, there are a total of N such possible 'cut-off schemes'. Figure 1 shows likelihoods as a function of p for the case where $N = 3$. We see that the highest likelihoods indeed occur for the framework with $(\mathcal{T}_{\text{all red}}, \xi^{\text{PM}})$ (uppermost trace in Fig. 1 (left)), and that these likelihoods coincide with the likelihoods for $(\mathcal{T}_{\text{all red}}, \xi_{\mathcal{V}}^{\text{typO}})$ (uppermost trace in Fig. 1 (right)).

B. The best xerographic distribution for a fixed theory

It is an encouraging check of our intuition that the framework with the highest likelihood (and so also the highest posterior probability under the assumptions of section II B) involves the theory that predicts all domains are red. A natural question that arises is: for a *fixed theory*, what xerographic distribution leads to the best performing framework? It does not take much more work to show that under these circumstances, the framework whose likelihood attains the supremum of the likelihoods over all xerographic distributions considered, for all assignments of probabilities p_i to the multiverse domains, is the one whose xerographic distribution corresponds to **PM**: $\xi = \xi^{\text{PM}} := \xi_{\mathcal{V}}^{\text{typD}}$.

To prove this, fix the theory $\mathcal{T}$ and let the p_i's be arbitrarily chosen but fixed. For an arbitrary c, we showed in section IV A that $P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typO}}) \leq P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typD}})$. Now, as in the proof there, the choice of c that will maximize $P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\text{typD}})$ is the one that will include the most number of non-negative terms in the sum in Eq. (6). This is just $c = \mathcal{V}$, and hence $(\mathcal{T}, \xi^{\text{PM}})$ will in general give the highest likelihood.

We see how this claim manifests in a simple case, by again considering the multiverse of cycles introduced in section III D, with xerographic distributions implementing a cut-off in time, as described in the last paragraph of section IV A. Figure 2 displays plots of the differ-
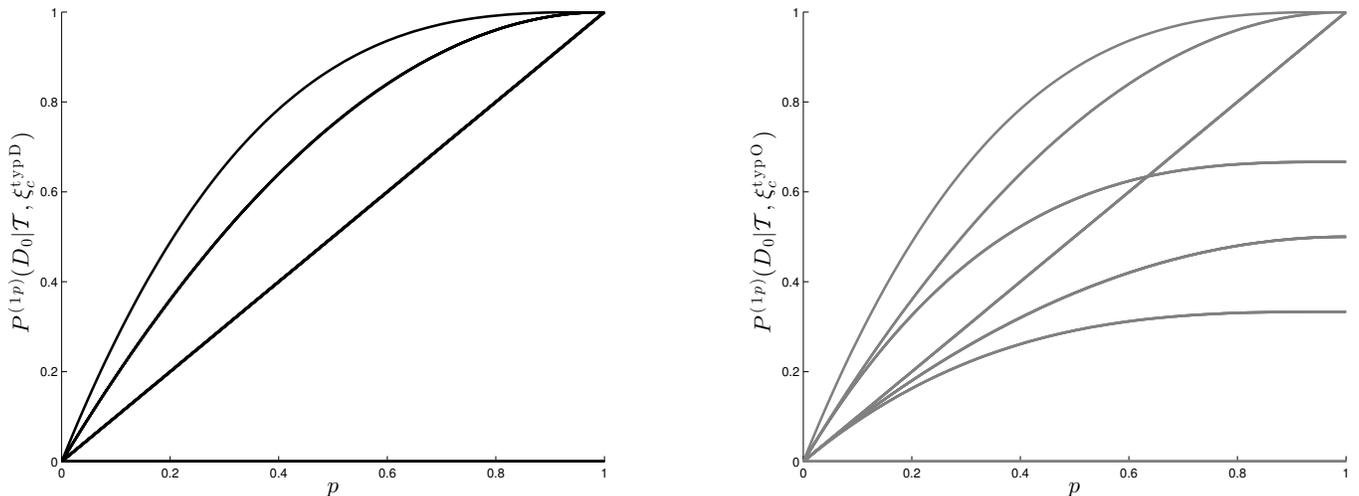
FIG. 1. (Left): First-person likelihoods for our data D_0, given a theory $\mathcal{T}$ and a xerographic distribution ξ_c^{typD}, for a multiverse of cycles with a total of 3 cycles (see section III D). The theories considered include all possible theories assigning the observable red/blue to each cycle (for a total of 2^3 theories), and the xerographic distributions can be nonzero only up to (and including) some terminal cycle (corresponding to the final element of c). We implement the assumption that we are typical with respect to other instances of our data up to (and including) that terminal cycle (for a total of 3 possible xerographic distributions; see **C1** of section III A). Likelihoods are shown as a function of p, the probability of the existence of observers in each cycle, which is here assumed to be the same for each cycle. Likelihoods for all possible frameworks (found by varying $\mathcal{T}$ and c—a total of 24 frameworks) have been superimposed on the plot. In this case, the existence of equivalent functional forms for likelihoods for different frameworks, gives rise to 4 distinct traces (including likelihoods which are zero for all values of p). The uppermost trace corresponds *only* to the framework $(\mathcal{T}_{\text{all red}}, \xi^{\text{PM}} := \xi_{\mathcal{V}}^{\text{typD}})$, and is the supremum of the likelihoods for all possible values of p. (Right): This plot displays similar information as in the left panel, but for the xerographic distribution that assumes we are typical of all observers, regardless of what data they see (**C2** of section III A). Here, the existence of equivalent functional forms for likelihoods for different frameworks, gives rise to 7 distinct traces (again, including likelihoods which are zero for all values of p). The uppermost trace is the same as the uppermost trace for the left panel, and corresponds only to the framework $(\mathcal{T}_{\text{all red}}, \xi_{\mathcal{V}}^{\text{typO}})$.

ence between the likelihood for the framework with some theory $\mathcal{T}$ and xerographic distribution corresponding to the principle of mediocrity $\xi = \xi^{\text{PM}} := \xi_{\mathcal{V}}^{\text{typD}}$, and the likelihood for the framework with the same theory and all other xerographic distributions (the probability p of observers in domains, labels the x-axis of each subplot). The case for $N = 3$ cycles is displayed. We see that for each row of the plot (corresponding to a fixed theory), the difference is non-negative in each case, confirming the general claim advanced in this section.

C. Does the principle of mediocrity generally give rise to the most predictive frameworks?

In light of the results in the last two subsections it is natural to ask whether there are cases where **PM** does not provide the highest likelihoods. We will show by construction that indeed it does *not*, when one is allowed to vary both the underlying theory and the xerographic distribution. For the sake of simplicity, we will restrict attention to the multiverse of cycles introduced in section III D, where again, the probability of observers in any cycle is the same for each cycle and is given by p.

There is a strong motivation for embarking on this

search, since some well-motivated theories suggest xerographic distributions which do not express the principle of mediocrity. We will further explicate such scenarios in the next section; but for now, we aim to figure out whether in such instances, likelihoods are generally highest for those frameworks that involve the principle of mediocrity.

To be explicit about the terms for our search: we are interested in whether for a given framework characterized by **PM**, that is, for some 'reference framework' $(\mathcal{T}, \xi^{\text{PM}} := \xi_{\mathcal{V}}^{\text{typD}})$, there could exist another framework $(\mathcal{T}^\star, \xi^\star)$ with a higher likelihood for at least some values of p. To steer the discussion away from more trivial cases, we will only consider situations where:

(i) $\mathcal{T} \neq \mathcal{T}_{\text{all red}}$: since otherwise, by the results of section IV A, we will not be able to find the required framework $(\mathcal{T}^\star, \xi^\star)$.

(ii) $\mathcal{T} \neq \mathcal{T}_{\text{no red}}$: where $\mathcal{T}_{\text{no red}} = (0, 0, \ldots, 0)$ is the theory that predicts no red cycles; since then the likelihood of our data D_0, given the framework $(\mathcal{T}_{\text{no red}}, \xi^{\text{PM}})$, vanishes, and we will be able to find the required frameworks $(\mathcal{T}^\star, \xi^\star)$ rather trivially.

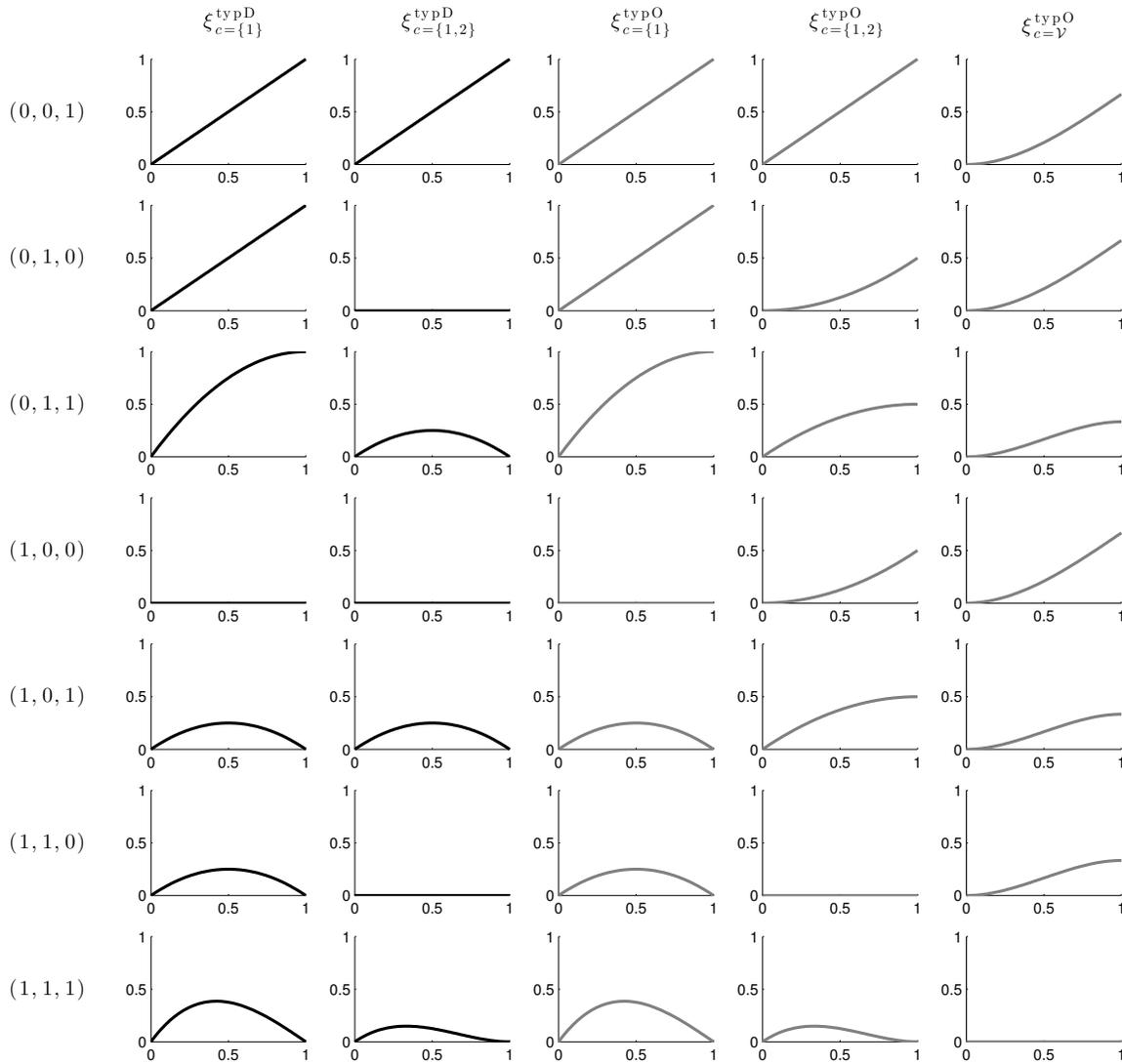(iii) $\xi^\star \neq \xi^{\text{PM}}$: since we already know from section IV A

FIG. 2. The multiverse of cycles introduced in section III D with $N = 3$ cycles. Each subplot shows $[P^{(1p)}(D_0|\mathcal{T}, \xi^{\mathrm{PM}}) - P^{(1p)}(D_0|\mathcal{T}, \xi_c^{\mathrm{typX}})]$ vs p (where X is either D or O). The term in square brackets is the difference between the first-person likelihood for the framework that implements the principle of mediocrity, namely $(\mathcal{T}, \xi^{\mathrm{PM}} := \xi_{\mathcal{V}}^{\mathrm{typD}})$, and the likelihood for another framework specified by varying only the xerographic distribution in accord with the cut-off procedure described in the last paragraph of section IV A. The variable p (labeling the x-axis of each subplot) is the probability of the occurrence of observers in any cycle (and is assumed to be the same in each cycle). The theory is shown on the left of each row, with 1 or 0 respectively denoting the prediction of red or blue in the corresponding cycle. The xerographic distribution against which we are comparing the principle of mediocrity is shown on the top of each column (traces referring to ξ_c^{typO} are shown in grey to guide the eye). We see that for each row, that is, for a fixed theory, the difference is non-negative for all possible values of p. It is in this sense that we claim the best xerographic distribution for a fixed theory is that corresponding to the principle of mediocrity.

that for $\xi^\star = \xi^{\mathrm{PM}}$, $\mathcal{T}^\star = \mathcal{T}_{\mathrm{all\ red}}$ will provide the supremum of the likelihoods in this case—and moreover, we are interested in finding frameworks where **PM** is not integral to constructing higher

likelihoods.

(iv) $(\mathcal{T}^\star, \xi^\star) \neq (\mathcal{T}_{\mathrm{all\ red}}, \xi_{\mathcal{V}}^{\mathrm{typO}})$: since again, this corresponds to the supremum as mentioned in (iii), following the results of section IV A.

So consider the case $N = 3$. Set $\mathcal{T}$ in our reference framework $(\mathcal{T}, \xi^{\mathrm{PM}})$, to be the theory that predicts a total of one of the three cycles is red: $\mathcal{T} = \mathcal{T}_{\mathrm{one\ red}}$. Then, from Eq. (10), we have $P^{(1p)}(D_0|\mathcal{T}_{\mathrm{one\ red}}, \xi^{\mathrm{PM}}) = p$. It is straightforward to show from Eq. (5) that for the theory that predicts only the first two cycles are red, which we will denote by $\mathcal{T}^\star = (1,1,0) \equiv \mathcal{T}_{RRB}$, the xerographic distribution given by $\xi^{\mathrm{typO}}_{c=\{1,2\}}$, which (clearly) *does not* correspond to the principle of mediocrity, implies: $P^{(1p)}(D_0|\mathcal{T}_{RRB}, \xi^{\mathrm{typO}}_{c=\{1,2\}}) = p(2-p)$. So the framework $(\mathcal{T}_{RRB}, \xi^{\mathrm{typO}}_{c=\{1,2\}})$ has a likelihood higher than the framework involving **PM**, since $p(2-p) > p$ (for all nontrivial values of p, i.e. $p \in (0,1)$).

More interesting behavior can be exhibited for the case where we take $\mathcal{T}^\star$ to be the theory that predicts a total of two of the cycles are red and the xerographic distribution to be $\xi^{\mathrm{typO}}_{\mathcal{V}} \equiv \xi^{\mathrm{typO}}_{c=\{1,2,3\}}$. In this case, as seen from Eq. (9), $P^{(1p)}(D_0|\mathcal{T}_{\mathrm{two\ red}}, \xi^{\mathrm{typO}}_{\mathcal{V}}) = \frac{2}{3}p(p^2 - 3p + 3)$. This likelihood displays the behavior that it does better than our reference framework $(\mathcal{T}_{\mathrm{one\ red}}, \xi^{\mathrm{PM}})$, depending on the value of p. In particular, for $p > \sim 0.63$, $(\mathcal{T}_{\mathrm{one\ red}}, \xi^{\mathrm{PM}})$ has a higher likelihood than $(\mathcal{T}_{\mathrm{two\ red}}, \xi^{\mathrm{typO}}_{\mathcal{V}})$, whereas the situation is reversed otherwise. Such behavior thereby exhibits parametric dependence of the success of **PM**. Both this situation and the one discussed in the last paragraph are displayed graphically in Fig. 3 (left).

Similar results can be obtained in the case where $N = 4$ (see Fig. 3 (right)). In this case, there exists a framework implementing **PM** with $P^{(1p)}(D_0|\mathcal{T}_{\mathrm{two\ red}}, \xi^{\mathrm{PM}}) = p(2 - p)$. This is less than or equal to the likelihood $P^{(1p)}(D_0|\mathcal{T}_{RRRB}, \xi^{\mathrm{typO}}_{c=\{1,2,3\}}) = p(p^2 - 3p + 3)$, for all p, which assumes a theory where only the first three cycles are red ($\mathcal{T}^\star = (1,1,1,0) \equiv \mathcal{T}_{RRRB}$). In addition, our reference framework's likelihood $P^{(1p)}(D_0|\mathcal{T}_{\mathrm{two\ red}}, \xi^{\mathrm{PM}})$ displays only parametric dominance over the likelihood associated with another framework not implementing **PM**, namely $(\mathcal{T}_{\mathrm{three\ red}}, \xi^{\mathrm{typO}}_{\mathcal{V}})$, which takes the value $P^{(1p)}(D_0|\mathcal{T}_{\mathrm{three\ red}}, \xi^{\mathrm{typO}}_{\mathcal{V}}) = \frac{3}{4}p(4 - 6p + 4p^2 - p^3)$. Again, interestingly, **PM** does not universally give rise to the highest likelihoods and in certain cases exhibits only a parameter-dependent dominance.

V. DISCUSSION

The principle of mediocrity is a controversial issue in multiverse cosmology. According to one way of thinking, it articulates our intuitions about how one should reason from an appropriately defined, peaked probability distribution, to a prediction of a possible observation. But crucially, this intuition has been developed either in controlled laboratory settings, or more generally, in cases where we understand, and have some (at least theoretical) control over, the conditions that obtain in systems of interest. The multiverse, however, is a different story.

It is plausible that we are *more* typical of a set of appropriately restricted multiverse domains, but whether we can positively assert typicality heavily depends on who or what is predicted by the theories and the conditionalization schemes which we consider in multiverse cosmology. These latter schemes, at best, set down conditions that are necessary for the presence of 'observers', but there is an ambiguity in defining precisely who or what these observers are. In addition we do not know precisely what parameters or conditions need to be fixed within the confines of any theory in order to unambiguously describe these observers. As a result, it is not clear that typicality is justified, even if we conditionalize in accord with the 'ideal reference class' of Garriga and Vilenkin [17]. Of course, we *may* be typical, but following this line of thinking, we do not have good reason to assert that we are.

The formalism of Srednicki and Hartle [19] allows one to neatly address this multi-faceted concern, which affects our ability to reason in multiverse scenarios [20]. Through their formalism, one can formulate a set of assumptions regarding typicality, and from this set, one can calculate relevant likelihoods for possible observations, to then see how well different assumptions do in terms of describing our observations. This is implemented in a Bayesian framework, so that *if*, as we have assumed in this paper (see section II B), we can assign equal amounts of prior credence to various candidate frameworks, then higher likelihoods translate to greater support for those frameworks given relevant experimental data. From the framework with the highest posterior probability then, one can *infer* how typical we are.

How well does typicality do? As we have discovered within the admittedly simplified multiverse setting of section III: for a fixed theory, the principle of mediocrity yields likelihoods for our data that attain the supremum of all likelihoods considered, for all values of the probabilities of observers in domains.

But an important caveat is that this result is *not* universal. Namely, if one is allowed to vary both the theory and the xerographic distribution implementing assumptions regarding typicality, the principle of mediocrity does not always provide the highest likelihoods. This is particularly pertinent when the set of candidate frameworks that constitute plausible alternatives for the description of a physical situation, do not always include xerographic distributions implementing the principle of mediocrity.

One example where this occurs is in the situation where 'Boltzmann brains' exist and out-number ordinary observers, but both sets of observers record the same data (see the discussion in [19], as well as [21–23]). In this case, the first-person likelihood of our observations might be higher under the principle of mediocrity; but an unwanted consequence of favoring the corresponding framework, is the high likelihood of us *being* Boltzmann brains. That is, this framework would also predict that our future measurements will be disordered, that is, uncorrelated with past measurements (as is presumed for Boltz-
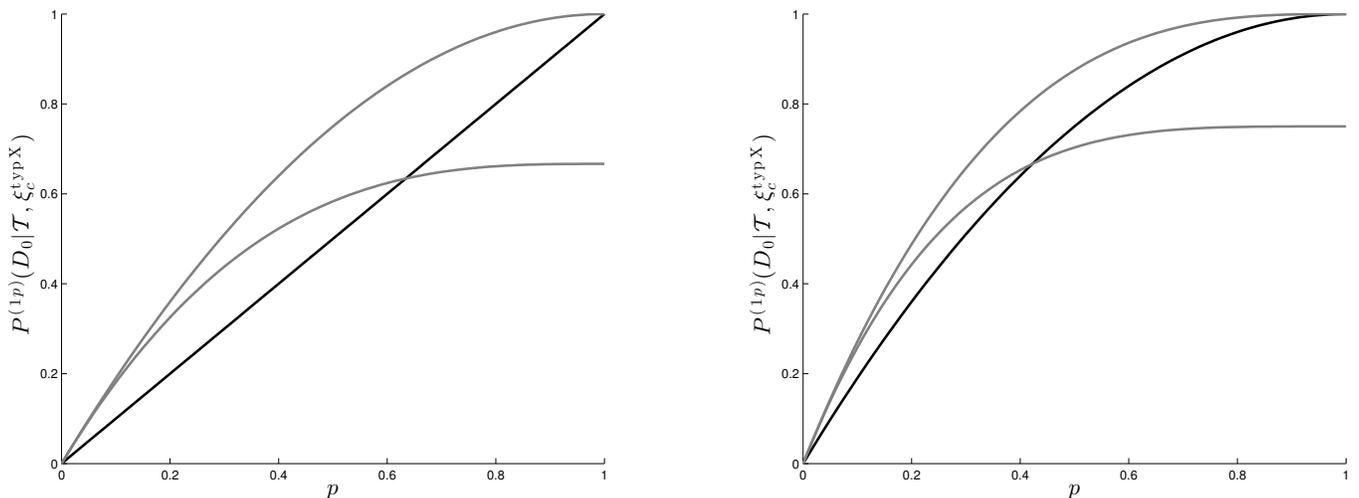
FIG. 3. (Left): First-person likelihoods for our data D_0, given a particular theory $\mathcal{T}$ and xerographic distributions ξ_c^{typX} (where X is either D or O), vs p, for the multiverse of cycles introduced in section III D for $N = 3$. The likelihoods for three separate frameworks are shown. The black line shows the likelihood for the framework $(\mathcal{T}_{\text{one red}}, \xi^{\text{PM}} := \xi_{\mathcal{V}}^{\text{typD}})$, where the theory predicts a single red cycle and the xerographic distribution corresponds to the principle of mediocrity $[P^{(1p)}(D_0|\mathcal{T}_{\text{one red}}, \xi^{\text{PM}}) = p]$. The uppermost grey trace corresponds to the framework $(\mathcal{T}_{RRB}, \xi_{c=\{1,2\}}^{\text{typO}})$, where the theory predicts only the first two cycles are red $(\mathcal{T}_{RRB} \equiv (1,1,0))$, and the xerographic distribution corresponds to typicality with respect to observers considered among only the first two cycles $[P^{(1p)}(D_0|\mathcal{T}_{RRB}, \xi_{c=\{1,2\}}^{\text{typO}}) = p(2-p)]$. We note that this framework *does better* than that containing a xerographic distribution implementing the principle of mediocrity. The lower grey trace corresponds to the likelihood for the framework $(\mathcal{T}_{\text{two red}}, \xi_{\mathcal{V}}^{\text{typO}})$, whose theory predicts that some two of the cycles are red, and that we are typical with respect to observers considered over the entire multiverse $[P^{(1p)}(D_0|\mathcal{T}_{\text{two red}}, \xi_{\mathcal{V}}^{\text{typO}}) = \frac{2}{3}p(p^2 - 3p + 3)]$. This trace crosses the likelihood for $(\mathcal{T}_{\text{one red}}, \xi^{\text{PM}})$, implying a parametric dependence of the dominance of the framework implementing the principle of mediocrity. The cross-over point of this dependence occurs for $p \sim 0.63$. (Right): Similar results for the case $N = 4$. Here, the black trace shows the likelihood for the framework implementing the principle of mediocrity with $P^{(1p)}(D_0|\mathcal{T}_{\text{two red}}, \xi^{\text{PM}}) = p(2-p)$. The uppermost trace corresponds to $P^{(1p)}(D_0|\mathcal{T}_{RRRB}, \xi_{c=\{1,2,3\}}^{\text{typO}}) = p(p^2 - 3p + 3)$ (where $\mathcal{T}_{RRRB} \equiv (1,1,1,0)$). The lower grey trace corresponds to the likelihood $P^{(1p)}(D_0|\mathcal{T}_{\text{three red}}, \xi_{\mathcal{V}}^{\text{typO}}) = 3p(4 - 6p + 4p^2 - p^3)/4$. In this latter case, the principle of mediocrity is only dominant for $p > \sim 0.42$.

mann brains). One way to avoid having to accept this consequence is by restricting the xerographic distribution accordingly—for example, by focussing attention on only a proper subset of appropriately chosen domains. This type of restriction has been actively developed in this paper, and is equivalent to an assumption of non-mediocrity under the scheme described in section II.[4]

To sum up: if some of the frameworks one considers have xerographic distributions that do not implement

the principle of mediocrity for relevant theories, then demanding that we favor a framework that includes the principle of mediocrity is hazardous. For as shown in section IV C, we cannot guarantee it will produce the framework with the highest likelihood.

It is important to note that we have selected a particular reference class within which to implement the principle of mediocrity—namely, observers who witness our experimental data. The motivation for this choice was to test a limiting case of a 'top-down' approach to conditionalization (see [14, 17, 18, 24] for further details on top-down approaches). This limiting case requires consideration of the maximally specific reference class in the setup at hand. The assumption of typicality with respect to this reference class (encoded in $\xi_{\mathcal{V}}^{\text{typD}}$) then corresponds to the principle of mediocrity. A key result in this paper is that typicality with respect to this reference class does not necessarily give rise to the highest likelihoods for our data D_0, if one is allowed to vary both the theory and the xerographic distribution under consideration (as explained in section IV C).

It is also important to note that we have evaluated frameworks based on first-person likelihoods generated

---

[4] One might wonder why we do not simply reject a theory that when partnered with the principle of mediocrity, does not constitute a plausible candidate framework. A response in the spirit of this paper, is that when the status of the principle of mediocrity is uncertain, it makes sense to examine the predictions of frameworks that constrain otherwise well-motivated theories by assumptions of non-mediocrity. A more pointed response is that one is simply not justified in rejecting a theory just because we would not be typical according to that theory. Hartle and Srednicki [16] raise this objection in discussing a (hypothetical) theory that predicts the existence of many more observers on Jupiter than on Earth. They claim it is unreasonable to reject such a theory, when we notice we are human and not Jovian, just because we would not be typical according to the theory.

for our data D_0. These computations are in accordance with the approach adopted by Srednicki and Hartle [19, §IV], who invoke such likelihoods in the evaluation of frameworks (where they also use xerographic distributions that correspond to each of $\xi_{\mathcal{V}}^{\mathrm{typO}}$ and $\xi_{\mathcal{V}}^{\mathrm{typD}}$).

Another significant question to address, that manifestly goes beyond the cosmological settings studied in this paper, is: how should one evaluate the predictive power of frameworks (and in particular, the principle of mediocrity) in a case where one aims to predict the value of an observable, say, that is not explicitly included in the conditionalization scheme adopted? For realistic cosmological calculations, securing the required separation of the observable from the conditionalization scheme is non-trivial. One naturally requires that the observable being predicted is (i) correlated with the conditionalization scheme (otherwise the conditionalization scheme will play no role in the predictive framework), but (ii) is not perfectly correlated with the conditionalization scheme (otherwise one is open to the charge of circularity). Thus when it is not clear exactly how observables are correlated with the defining features of a conditionalization scheme, the need to strike a balance between (i) and (ii) gives rise to a difficult problem—namely, how to distinguish the observable to be predicted, from the defining features of the conditionalization scheme (see Garriga and Vilenkin [17, §III] who mention such concerns).

Assuming that one has a solution to this problem, a formalism that can handle this type of predictive setting is described by Srednicki and Hartle [19, §VI]. For the computation of first-person likelihoods, the appropriate way to proceed, as detailed by Garriga and Vilenkin [17] and Hartle and Hertog [25, 26], is to explicitly leave out that part of our data that involves the observable we aim to predict, in the specification of our conditionalization scheme. The assumption of typicality with respect to the reference class implicit in this specification, would then be the appropriate implementation of the principle of mediocrity [17].

It remains to apply the methods introduced by Srednicki and Hartle [19], and advanced in this paper, to more realistic cosmological settings, in order to more fully assess the extent of the errors that may arise from universally imposing the principle of mediocrity (see [25, 26] for recent work in this direction). For now, our conclusion must be that the principle of mediocrity (in the style of [10–13, 17]) is more questionable than has been claimed.

Appendix: Calculation of $P^{(1p)}(D_0|\mathcal{T}, \xi^{\mathbf{PM}})$ for the multiverse of cycles

For the sake of completeness, we present the calculation for the first-person likelihood of our data D_0, given a theory $\mathcal{T}$, under the assumption of a xerographic distribution that imposes the principle of mediocrity, $\xi = \xi^{\mathrm{PM}} := \xi_{\mathcal{V}}^{\mathrm{typD}}$. We work within the cyclic cosmological model of Srednicki and Hartle [19] under the assumptions of section III D; where we have N domains stretched out in time, each with a probability p of housing observers. From these assumptions, Eq. (6) reduces to

$$P^{(1p)}(D_0|\mathcal{T}, \xi^{\mathrm{PM}}) = \sum_{\vec{\sigma}\in\mathcal{K}_{D_0}(\mathcal{V},\mathcal{T})} p^{\sum_{i=1}^N \sigma_i}(1-p)^{N-\sum_{j=1}^N \sigma_j}. \tag{A.1}$$

We organize the sum by separately considering configurations according to the total number of observers $n_O = \sum_{i=1}^N \sigma_i$ in each configuration $\vec{\sigma}$. Equation (A.1) can then be thought of as $P^{(1p)}(D_0|\mathcal{T}, \xi^{\mathrm{PM}}) = \sum_{n_O=1}^N G(p, n_O, N_R)$, for some function G we will compute shortly. The total number of red cycles N_R depends on the theory $\mathcal{T}$, and we will separately consider the two cases into which the sum naturally partitions, namely, $1 \le n_O \le N_R$ and $N_R < n_O \le N$.

For $1 \le n_O \le N_R$, we generate an expression for $G(p, n_O, N_R)$ by sequentially placing all possible numbers of observers (out of a maximum of n_O observers) in N_R red cycles. In general, we find

$$G(p, n_O, N_R) = p^{n_O}(1-p)^{N-n_O} \left[\binom{N_R}{n_O} + \binom{N_R}{n_O-1}\binom{N-N_R}{1} + \cdots + \binom{N_R}{1}\binom{N-N_R}{n_O-1} \right]$$

$$= p^{n_O}(1-p)^{N-n_O} \sum_{k=0}^{n_O-1} \binom{N_R}{n_O-k}\binom{N-N_R}{k}$$

$$= p^{n_O}(1-p)^{N-n_O} \left[\binom{N}{n_O} - \binom{N-N_R}{n_O} \right]$$

$$= p^{n_O}(1-p)^{N-n_O} \sum_{k=1}^{N_R} \binom{N-k}{n_O-1},$$

where we have used Vandermonde's identity in obtaining the third line, and have iterated using Pascal's formula in obtaining the fourth line.

For $N_R < n_O \leq N$, we obtain the same final result using a similar sequence of steps. Putting this all together, gives

$$P^{(1p)}(D_0|\mathcal{T}, \xi^{\mathrm{PM}}) = \sum_{n_O=1}^{N} p^{n_O}(1-p)^{N-n_O} \sum_{k=1}^{N_R} \binom{N-k}{n_O-1}$$

$$= \sum_{k=1}^{N_R} \sum_{n_O=1}^{N-k+1} \binom{N-k}{n_O-1} p^{n_O}(1-p)^{N-n_O}$$

$$= \sum_{k=1}^{N_R} p(1-p)^{k-1} \sum_{m=0}^{N-k} \binom{N-k}{m} p^m (1-p)^{N-k-m}$$

$$= p \sum_{k=1}^{N_R} (1-p)^{k-1}$$

$$= 1 - (1-p)^{N_R}, \tag{A.2}$$

where we have used the binomial theorem to obtain the fourth line. This agrees with the corresponding expression (Eq. (5.5)) in [19], and Eq. (10) in the text above.

[1] A. Vilenkin, Phys. Rev. D **27**, 2848 (1983).
[2] A. D. Linde, Phys. Lett. **129B**, 177 (1983).
[3] A. D. Linde, Phys. Lett. B **175**, 395 (1986).
[4] R. Bousso and J. Polchinski, J. High Energy Phys. 06 (2000) 006.
[5] S. Kachru, R. Kallosh, A. Linde, and S. P. Trivedi, Phys. Rev. D **68**, 046005 (2003).
[6] B. Freivogel, M. Kleban, M. Rodríguez Martínez, and L. Susskind, J. High Energy Phys. 03 (2006) 039.
[7] L. Susskind, in *Universe or Multiverse?*, edited by B. Carr (Cambridge University Press, Cambridge, 2007).
[8] B. Carter, in *Confrontation of Cosmological Theories with Observational Data. IAU Symposium No. 63*, edited by M. S. Longair (D. Reidel Publishing Company, Dordrecht + Boston, 1974) pp. 291–298.
[9] J. B. Hartle, in *Universe or Multiverse?*, edited by B. Carr (Cambridge University Press, Cambridge, 2007).
[10] A. Vilenkin, Phys. Rev. Lett. **74**, 846 (1995).
[11] J. R. Gott III, Nature **363**, 315 (1993).
[12] D. N. Page, Int. J. Mod. Phys. D **05**, 583 (1996).
[13] N. Bostrom, *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge, New York & London, 2002).
[14] S. Weinstein, Class. Quantum Grav. **23**, 4231 (2006).
[15] L. Smolin, in *Universe or Multiverse?*, edited by B. Carr (Cambridge University Press, Cambridge, 2007).
[16] J. B. Hartle and M. Srednicki, Phys. Rev. D **75**, 123523 (2007).
[17] J. Garriga and A. Vilenkin, Phys. Rev. D **77**, 043526 (2008).
[18] F. Azhar, Class. Quantum Grav. **31**, 035005 (2014).
[19] M. Srednicki and J. Hartle, Phys. Rev. D **81**, 123524 (2010).
[20] C. Smeenk, Stud. Hist. Phil. Mod. Phys. **46**, 122 (2014).

[21] A. Albrecht and L. Sorbo, Phys. Rev. D **70**, 063528 (2004).

[22] D. N. Page, J. Cosmol. Astropart. Phys. 10 (2008) 025.

[23] J. R. Gott III, arXiv:0802.0233.

[24] A. Aguirre and M. Tegmark, J. Cosmol. Astropart. Phys. 01 (2005) 003.

[25] J. Hartle and T. Hertog, Phys. Rev. D **88**, 123516 (2013).

[26] J. Hartle and T. Hertog, arXiv:1503.07205.