Effective Realism and the Problem of Boltzmann Brains

Marco Maggiani

Department of Philosophy, University of Pittsburgh

Abstract

The Boltzmann brain problem threatens to undermine our use of cosmological evidence. If our best-fit cosmological model, ΛCDM, holds indefinitely into the future, most observers with our evidence would be random fluctuations rather than ordinarily evolved observers like ourselves. This seems to erode the connection between our data and ΛCDM itself. Existing responses to the problem, such as denying our typicality, appealing to externalist evidence, or assigning zero priors to cognitively unstable theories, all face serious difficulties. I propose a different solution: ΛCDM should be interpreted as an *effective theory* whose domain of applicability does not extend to the extreme time scales that lead to a numerical domination by Boltzmann brains over ordinary observers. This is analogous to the case of quantum field theory, which is highly successful within its own domain of applicability, but breaks down at sufficiently short length scales. On this interpretation, the skeptical issue is resolved not by new epistemic rules but by recognizing that ΛCDM, like other effective theories, must be realized by a more fundamental description. The existence of many possible such descriptions compatible with our evidence, and the way they respond to evidence under self-locating uncertainty, already dissolves the skeptical problem.

1 Introduction

The Boltzmann Brain problem is not just a speculative feature of cosmology but a threat to our use of cosmological evidence. If the ΛCDM model, our current best-fit model for cosmology, is taken to hold without qualification, then most observers with our evidence would be random

fluctuations rather than evolved beings. In that case, the evidential connection between our data and ΛCDM would be undermined, since a theory that implies our evidence is almost certainly misleading cannot be confirmed in the usual way by what we take ourselves to observe. The problem is therefore not only about the far future of the universe but about whether our present evidence can genuinely confirm our best cosmological theories.

Several responses have been proposed. Some deny that we are typical among observers with our evidence, while others appeal instead to externalist notions of evidence. A third proposal, due to Sean Carroll, holds that any cognitively unstable theory must be assigned a prior probability of zero. Each of these moves has its difficulties. Typicality-denying views require agents to assign different probabilities to their self-location, even though those locations cannot be distinguished either by evidence or by theoretical virtues. Externalism fails to preserve our knowledge of cosmology and of the deep past. Carroll's proposal, finally, introduces a supplementary Bayesian rule that faces serious objections in its own right.

In this paper, I defend a different response. I argue that the ΛCDM model should be treated as an effective theory, with a domain of applicability that does not extend to the extreme timescales at which Boltzmann Brains would dominate. This parallels the status of quantum field theory, which is well confirmed within its domain but known to break down at very short length scales. On this view, the Boltzmann Brain problem does not force skepticism, and it does not require new rules of scientific reasoning. It only shows that ΛCDM, like other effective theories, must ultimately be realized by a more fundamental description of the universe, what I will call a *completion*, one that avoids its theoretical problems. Integral to this response is a discussion of undermining drug cases, following Elga (2025), and of how different possible completions are confirmed under self-locating uncertainty.

The structure of the paper is as follows. Section 2 lays out the Boltzmann Brain problem in detail. Section 3 considers Carroll's proposal and the idea of cognitive instability. Section 4 discusses Elga's treatment of cognitive instability, including his Labelscramble analogy and its lessons. Sections 5 and 6 introduce effective realism in the context of quantum field theory and apply it to the ΛCDM model. Section 7 explores the implications for confirmation and self-locating evidence. Section 8 compares my proposal with Carroll's, and Section 9 concludes.

2 The Boltzmann Brain Problem

As noted by Carroll (2021), the Boltzmann brain problem arises if the universe meets two conditions: it is extraordinarily long-lived, or even eternal, and undergoes random fluctuations that can instantiate conscious observers. If the product of the rate of fluctuations and the lifetime of the universe is sufficiently large, then the most common observers will be fluctuations, not the "ordinary observers" we take ourselves to be.

A discussion of this situation is pressing because the current best-fit model for cosmology—the ΛCDM model (where Λ stands for the cosmological constant and CDM for "cold dark matter")—seems to predict that ordinary observers are outnumbered for this reason. According to ΛCDM, if Λ remains constant, the universe asymptotically approaches a de Sitter phase, which in some respects resembles a box of gas at equilibrium. De Sitter space continually expands; therefore, classically, we would expect its density to approach zero, leaving nothing from which conscious observers could be created. However, de Sitter space has, like a black hole, a horizon. Rather than observing that horizon from outside (as we would for a black hole), we observe it from inside, and Gibbons and Hawking (1977) have shown that it generates thermal radiation with a very low but non-zero temperature.[1]

As a result, the ΛCDM model describes a universe that is not only eternal but will also asymptotically reach a certain non-zero fixed temperature. Consequently, it predicts that Boltzmann brains will eventually dominate in number over ordinary observers. This remains true, assuming that subjective states supervene on physical states, even if we conditionalize on each possible subjective state that humans can experience.[2] Applying a "bland indifference principle" (Bostrom 2003) in this context should then persuade us that we are Boltzmann brains, not ordinary observers (Carroll 2021).

Let me mention two strategies for blocking the skeptical inference that we are Boltzmann brains before mostly setting them aside for the rest of this paper. First, we could deny that we are *typical* observers among those who share our subjective experiences. For instance, Hartle and Srednicki (2007, 2010) argue that, in addition to a cosmological theory T (which specifies, among other things, the kinds of observers in the universe and their numbers), we must also choose a *xerographic distribution* ξ. This distribution determines the probability that we are

[1] The exact Gibbons–Hawking temperature is $T = \sqrt{\Lambda/(12\pi^2)}$. The expression is given in natural units, obtained by setting $c = \hbar = k_B = 1$, where c is the speed of light, $\hbar$ the reduced Planck constant, and k_B Boltzmann's constant. For the observed value $\Lambda \approx 1.1 \times 10^{-52}\,\mathrm{m}^{-2}$, the resulting temperature is about $2 \times 10^{-30}\,\mathrm{K}$.

[2] See Carroll (2021) for a thorough exploration of the scientific details.

one of the observers who share our subjective state among those identified by T. It is natural to assume a uniform distribution across all such observers. Hartle and Srednicki, however, maintain that it is equally legitimate to adopt a distribution according to which we are atypical, since *empirical observations* do not privilege a uniform distribution. Other proposals that fall into this category include Dogramaci (2020), Kotzen (2021), and Dogramaci and Schoenfield (forthcoming), all of which reject the assumption that we are typical observers.

This class of proposals can be rejected if we accept a limited indifference principle called *Center Indifference* (Elga 2004). Stating this principle requires introducing a couple of definitions. Let a *centered* world be a possible world associated with a designated individual and a time (Lewis 1979), and call two centered worlds *similar* if they share the same possible world. Then Center Indifference can be stated as follows (Builes 2024):

Center Indifference: For any two similar centered worlds c_1 and c_2, if both c_1 and c_2 are compatible with your evidence, then it is rationally required to set $\text{Cr}\,(c_1 \mid c_1 \text{ or } c_2) = \frac{1}{2}$. [3]

Center Indifference is not only intuitive but is supported by powerful arguments. For instance, Builes (2024, pp. 780–781) argues that, for any two similar centers c_1 and c_2 that share the same evidence, the usual conditions for assigning a higher credence to one possibility over another do not apply. The reason is that c_1 and c_2, in addition to possessing the same evidence, share the same possible world, and therefore agree on all non-indexical facts. This makes it difficult to see how one of the centers could be preferable to the other with respect to theoretical virtues like simplicity and explanatory power. Moreover, even supposing we could make sense of the idea that one of the centers is theoretically favored, either one of the agents occupies it, or someone with the same evidence does. And why would the theory, according to which someone with the same evidence as the first agent occupies the preferred predicament, be more complex or less explanatory than the one according to which the first agent occupies it? In short, Center Indifference offers a well-motivated reason to resist attempts to deny our typicality.

Another possible strategy for solving the Boltzmann Brain problem is to appeal to some version of externalism about evidence. An externalist is likely to argue that Boltzmann brains, in virtue of having never interacted with their environments, possess extremely impoverished

[3] Elga (2004) defended the closely related principle that "similar centered worlds deserve equal credence," defining two centered worlds as similar if they share the same possible world and are subjectively indistinguishable. Builes' principle improves on Elga's original formulation because it is neutral between internalist and externalist conceptions of evidence.

evidence compared to ours. However, it is very difficult to make this strategy fully work. Universes in which Boltzmann brains dominate would also contain larger fluctuations: Boltzmann people, Boltzmann solar systems, and even Boltzmann galaxies (Carroll 2021, p. 10). Agents that live in such larger fluctuations ordinarily interact with their environments in the same way humans do. So, while externalist considerations might ameliorate the problem of being disembodied brains floating in the void, adopting externalism is unlikely to preserve our knowledge of cosmology and of the deep past, and I take the latter to be a requirement that any satisfying solution to the Boltzmann Brain problem should meet.[4]

3 Cognitive Instability and Carroll's Zero-Prior Proposal

Additional complications arise because any theory that predicts the dominance of Boltzmann brains seems to be self-undermining. In a randomly fluctuating universe, there is no reason for our knowledge to be correlated with the outside world. So, if we live in such a universe, it is overwhelmingly likely that our beliefs concerning physics, including our beliefs about our cosmological models, are a product of a random fluctuation. Therefore, we would have no reason to trust them.

The question, then, is how we should respond to the prospect of a theory that is self-undermining in this sense—or, in words that Carroll attributes to David Albert, *cognitively unstable*. Carroll (2021) has proposed that the best approach is to refuse to take such a possibility seriously, by setting our priors for cognitively unstable theories to zero (or so low that their posteriors remain negligible even after receiving confirming evidence). Carroll's prescription is then that "If we construct a model such as ΛCDM or a particular instantiation of the inflationary multiverse that seems to lead us into such a situation, our job as cosmologists is to modify it until this problem is solved, or search for a better theory" (Carroll 2021, p. 18).

Carroll's proposal is influential and constitutes the starting point for much of the philosophical discussion of the Boltzmann brain problem, but it faces at least two compelling objections. First, as Kotzen (2021, p. 26) notes, cognitive instability and low prior probability are logically distinct (even if sometimes correlated) properties of a hypothesis. Instability alone is not sufficient to drive a prior down to zero, or nearly so. A vivid counterexample comes from a certain class of "conspiracy theories." Suppose an agent or group systematically manipulates all of

[4]However, see Saad (forthcoming) for a radical externalist attempt at solving the Boltzmann Brain problem.

my evidence, including evidence about how they manipulate it. In that case, I could not form evidence-based reasons for believing the theory. Yet if the theory is independently reasonable and plausibly explains my evidence, it may still deserve a credence higher than "very low" (Kotzen 2021, p. 23). Moreover, Carroll's proposal introduces an entirely new rule that does not appear to follow from or fit naturally with standard Bayesian methodology (Page 2024).

Second, Elga (2024, pp. 4-5) highlights a problematic feature of Carroll's proposal.[5] Arguably, the principle can be applied not only to Boltzmann brain scenarios, but also to cases involving large numbers of radically misled agents, e.g., computer-simulated agents. If so, this fact could be used to manipulate the future. Suppose that humanity decides to set up a system where no misled agents are ever created unless a system of sensors detects that a nuclear war has started, in which case the sensors trigger the creation of a multitude of misled observers. Carroll's proposal rules out this possibility, because we must assign a very low prior to any hypothesis on which misled observers dominate numerically over accurate ones. Therefore, we can expect one of the following: (1) any attempt at setting up the described sensor arrangement will always be mysteriously prevented, which seems to impose a conspiratorial limit on our engineering powers, or (2) this strategy can, in fact, be implemented to ensure that there never is a nuclear war. Neither of these options seems plausible.

Beyond these objections, any adequate response must clarify the scientific status of the ΛCDM model itself. The Boltzmann brain problem thus raises not only epistemological issues but also the question of what sort of commitment we should have to ΛCDM. Ideally, the correct solution should address this question while also avoiding the two objections discussed above. In other words, it should clarify the standing of ΛCDM while remaining epistemologically plausible.

4 Undermining Drug Cases

In addition to the objections above, Elga (2025) shows that the Boltzmann brains scenario does not exhibit cognitive instability in the strict sense. To do so, he considers a thought experiment in which your evidence that you have taken a drug depends on your visual experience. More specifically, consider a case where, after swallowing a pill from a bottle in front of you (having swallowed no other pill), you read the bottle's label, and it reads:

[5] Elga (2024) is an earlier draft of Elga (2025), available on his website.

Labelscramble (50mg): causes hallucinations that replace the text of pill bottle labels with hallucinated random text.

Note how this case resembles the Boltzmann brain scenario with respect to cognitive instability. If you believe you have taken the drug, you shouldn't trust your visual experiences, and that seems to undermine your evidence of having taken the drug (given that your evidence consists in your having a visual impression of the bottle's label). On the other hand, if you believe you have not taken the drug, you should trust your visual impression of the label and conclude you have taken it.

However, Elga (drawing on Egan and Elga [2005, p. 81] and Talbott [2020, p. 295]) presents a simple Bayesian argument that shows that there is no instability involved in the Labelscramble case after all. Let P be the probability function you have before you read the label "Labelscramble…", E the evidence you gain when you read it, and L the proposition that you have taken a Labelscramble pill.

Now suppose, in a simplified version of the Labelscramble case, that you are sure no other hallucinogens are present, that you have ingested only one pill from an accurately labeled bottle, and that your visual perception is reliable unless you have taken a hallucinogen. Under these assumptions, you rule out scenarios in which you merely appear to see "Labelscramble…" without in fact having taken Labelscramble. In this version of the case, the evidence settles that you have taken the drug: $P(L \mid E)=1$, since $P(E\&\neg L)=0$.

Why isn't self-undermining present? The reason is that your visual impression of "Labelscramble…" is much more likely to come about if you took Labelscramble than if you did not. This is in no way grounded in your ability to read labels, but depends on whatever evidence you had that allowed you to rule out scenarios in which you merely seemed to see "Labelscramble…" without taking it.

Labelscramble is the only, in Elga's words, "fishy" hypothesis you take into consideration. In a more realistic scenario, you would instead give some small credence to a large number of "fishy" hypotheses: that the bottle contains Labelscramble, that it contains some other hallucinogen, that it was mislabeled, that your eyes are failing despite not having taken a hallucinogen, and so on.

In this case, what should you believe after you read the label? To simplify, assume that the fishy hypotheses F_1, F_2, …, F_k are mutually incompatible, and let F be the disjunction that some fishy hypothesis or other is true. Plausibly, your evidence strongly confirms F,

because E is far more to be expected given that something fishy was going on than given that nothing was. Even though each of the fishy scenarios is individually unlikely to produce the particular impression of seeing "Labelscramble…", a non-fishy scenario (one where you took no hallucinogen, the bottle is accurately labeled, your eyes aren't failing, and so on) was much less likely to do so. So seeing "Labelscramble…" should make you confident that something fishy is going on.

In sum, the visual experience of reading "Labelscramble…," rather than prompting the conclusion that you have taken Labelscramble and are hallucinating, should give rise to the conclusion that one of the "fishy" hypotheses obtains. Of course, this raises the further question of how your credences should be distributed among the many fishy hypotheses. If they each begin with roughly equal priors, then those under which "Labelscramble…" was most probable will end up with the greatest posterior weight. Formally, the posterior odds between two fishy hypotheses F_a and F_b are given by the ratio $\mathrm{P}(\mathrm{E} \mid F_a)/\mathrm{P}(\mathrm{E} \mid F_b)$, which in turn depends on the details of the Labelscramble case. I return to this issue in Section 7.

Elga devised the Labelscramble case as an analogue of the Boltzmann brain problem.[6] The point of the case is that instability considerations about theories that predict the numerical dominance of Boltzmann brains do not make it reasonable to be confident that we ourselves are Boltzmann brains. But they also do not license, by themselves, the opposite conclusion that we can keep trusting ordinary science. Rather, the lesson is that we should become confident that something "fishy" is going on without becoming particularly committed to any specific fishy hypothesis, including the hypothesis that we are Boltzmann brains. Some of the "fishy" possibilities might include being a Boltzmann brain, being a brain in a vat, living in a simulation, that there is a conspiracy producing misleading science, or that our universe does not in fact contain Boltzmann brains even though we receive cosmological evidence that (by ordinary standards) indicates that it does (the latter is the hypothesis that Elga ultimately favors).

However, the hypotheses you deem worthy of serious consideration will depend on your initial credence in each of them. In particular, other things being equal, hypotheses you consider

[6] As Elga (2024) observes, the framework of Section X must be generalized to apply to the Boltzmann brain problem. This is because the framework assumes simple conditionalization: agents start with a prior credence function $\mathrm{P}(\cdot)$, then learn about some evidence E, and end up with a posterior $\mathrm{P}(\cdot|\mathrm{E})$. However, the Boltzmann brain scenario involves distrust in one's own memories, and simple conditionalization cannot account for a situation in which our memories are in question (as shown, for instance, by Arntzenius [2003]). Elga's solution is to amend simple conditionalization by introducing what he calls a bracketing constraint: "one's credence in a claim should equal one's prior credence in that claim conditional not on the deliverances of the faculty, but rather the claim that the faculty produced those deliverances" (Elga 2024, p. 14). For further development and defense of this generalization, see Elga (2024), pp. 12–15.

mundane (and to which, therefore, you assign a high prior probability) will receive most of your confidence. My contention is that a mundane hypothesis, namely, that our attitude towards the ΛCDM model should be to regard it as an *effective* theory, is also available in the context of the Boltzmann brain problem.

5 Effective Realism and Quantum Field Theory

Recently, a position known as *effective realism* has been developed in the context of the philosophy of quantum field theory (QFT) (Wallace 2006, 2011; Fraser 2018, 2020; Williams 2019). Effective realism is a form of scientific realism related to *selective realism* (developed by, among others, Worrall 1989, Psillos 1999, and Kircher 1993). However, while selective realism prescribes that we should be committed only to certain *posits* among those our scientific theories make (usually, those that are used to explain and predict phenomena), effective realism restricts our commitment specifically to a certain *domain of applicability*.

It is useful to introduce effective realism in the context of QFT, both because it is there that it can be found in its most developed form, and because there are many interesting parallels between the effective interpretation of QFT and that of the ΛCDM model.

To understand why, in QFT, our commitments should be restricted to a certain domain of applicability, it is useful to consider condensed matter physics and its employment of QFT methods. Condensed matter physics studies the properties of matter in its solid and liquid phases. At large scales, solid and liquid bodies look like continuous systems and can therefore be treated as fields.

For this reason, they are suitable for the application of methods taken from "particle physics" QFT. However, it is well known that these methods lead to divergent integrals at short length scales. In the case of condensed matter physics, this is unsurprising: we know that at short scales, solids have a discrete structure, so we should expect their continuous description to break down at those scales. Hence, it is reasonable to refrain from calculating the divergent integrals all the way down to zero length scales; instead, they should be cut off around the atomic length scale (i.e., around 0.1 nm).

Naturally, the failure of the continuum description of solids at atomic length scales is more complex than any simple "cutoff" can account for. As a result, there will be various methods for implementing the cutoff, and none of them will be exact.

Because of the strength of the interactions at short length scales, we would reasonably expect this to have an impact at larger scales. And it does—but, surprisingly, modern renormalization theory reveals that the only effects at large scales are changes in a finite number of interaction terms, known as renormalizable interactions. The renormalizable interactions cannot be derived from our theories but can be measured empirically.

Since this story is perfectly reasonable for condensed matter physics, there is no reason we cannot tell the same story for QFT. Of course, an explanation must exist for why the degrees of freedom on length scales far below those accessible experimentally can be frozen out: this could be accomplished by a deeper theory that avoids the divergences of QFT, such as string theory, loop quantum gravity, or other proposed or as-yet unimagined theories. Call these deeper theories the **completions** of our quantum field theories.

This discussion should give a sense of what effective realism means in the case of QFT. A realist attitude towards empirically successful QFTs can be maintained; however, they have an inherently limited range of applicability. By their very nature, they do not provide a complete description of reality across all length scales.

6 Effective Realism and the ΛCDM Model

Now, there is a question concerning whether scientific theories *in general* should be regarded as effective, but that is beyond our present scope. For our purposes, what matters is that the ΛCDM model should be regarded as effective. That is the case for the same reasons that led us to believe that QFT is effective. In the case of QFT, it follows from the discussion above that the following three conditions hold and speak in favor (especially the first and the second) of an effective interpretation of empirically successful QFTs:

i. The theory breaks down at arbitrarily short length scales.

ii. The scales at which the theory breaks down are well outside our empirical reach.

iii. Candidate theories that provide a completion of QFT without breaking down at short length scales—such as string theory and loop quantum gravity—have been developed.

Analogous considerations apply to the ΛCDM model.

Breakdown. When the ΛCDM model is extrapolated to extremely long timescales, so that Boltzmann brains vastly outnumber ordinary observers, it appears to run into epistemological

difficulties, a kind of *epistemological breakdown*. In such a scenario, a large part of its apparent empirical support is undermined by the fact that most observers with our evidence would be predicted to be Boltzmann brains. If we make the plausible assumption—one I will return to below—that an unconfirmed scientific model begins with a low but non-negligible prior and that ordinary forms of confirmation are needed to raise its posterior significantly, the model's posterior remains low. This remains the case even if we do not adopt Carroll's proposal to assign a prior of zero, or so low that no confirmation could make the posterior significant. The posterior remains low not because the prior is especially low but because, under standard Bayesian updating, the model cannot be confirmed in the ordinary way.

Remoteness from Experience. According to the ΛCDM model, the time scales involved in producing a significant number of Boltzmann brains are well outside our empirical reach. For instance, Carroll (2021) estimates an average of $10^{10^{66}}$ years for the production of a single Boltzmann brain, while Davenport and Olum (2010) estimate $10^{10^{69}}$ years. There is no obvious reason to believe *a priori* that the ΛCDM model must hold at those scales.

Completions. The existence of serious scientific completions that prevent the breakdown of a model is not strictly necessary to argue that it is effective (in the sense delineated above), but it certainly strengthens the case. I will present two such possible completions of the ΛCDM model. My aim here is simply to establish that such completions exist, not to provide an exhaustive survey of all possible completions, either under present discussion by the scientific community or merely hypothetical.

The first concerns the possibility of a *false vacuum collapse*. A straightforward model that avoids the Boltzmann brain problem involves vacuum transitions: if the de Sitter vacuum decays to a lower energy state quickly enough, Boltzmann brains would never have the time to form. Specifically, the decay could occur through bubble nucleation, where a small region of space transitions to a lower-energy vacuum via quantum tunnelling and forms a "bubble" that expands at a speed close to that of light. If produced in sufficient numbers, these bubbles eventually fill space, and the phase transition is said to have *percolated*. A universe that has undergone percolation might still last for a long time (or even forever), but what matters is that it does not remain forever in a de Sitter phase where Boltzmann brains can emerge and ultimately outnumber ordinary observers (Carroll 2017).[7]

[7]Achieving percolation is not trivial. Since the space between the bubbles expands exponentially, the phase transition can be completed only if the bubble formation rate is at least one per Hubble volume (the region of spacetime surrounding an observer, extending roughly 10^{31} light years, beyond which objects move away from

Another promising possibility involves various dynamical dark energy models, particularly those based on *quintessence*. Unlike the constant dark energy postulated by the ΛCDM model, quintessence is a hypothetical form of dark energy described by a dynamical scalar field that evolves over time. If dark energy were a form of quintessence, this would affect how the universe expands (or contracts) in the long term.

The rate at which the universe expands depends on the interplay between gravity and dark energy. In the early universe, gravity was dominant, and as an attractive force, it caused the universe's expansion to slow down over time. However, as matter became more diluted, the effects of dark energy took over, causing the expansion to speed up.

In the ΛCDM model, the density of dark energy is assumed to be constant; therefore, once the effects of dark energy overtake the gravitational ones, they remain dominant forever. The universe eventually approaches a state of steady exponential expansion that continues indefinitely (with distances between galaxies growing at an exponential rate, meaning that for each time interval, the proportional increase in distances is the same as in the previous interval). This is precisely the everlasting de Sitter phase that gives rise to the Boltzmann brain problem.

Quintessence models can avoid this situation in at least two different ways. The first is through *phantom quintessence* models, which postulate that the density of dark energy increases over time rather than remaining constant. This would cause the universe to expand at more than a steady exponential rate: its expansion would accelerate more and more as the effect of dark energy grows. Eventually, this leads to a catastrophic "*Big Rip*" singularity at a finite future time (Caldwell 2002; Caldwell et al. 2003). A Big Rip scenario entails that all bound structures (galaxies, planets, molecules, atoms, and so on) and, finally, spacetime itself are disintegrated (see Caldwell et al. 2003 for a timeline of the events leading to the eventual spacetime disintegration). Thus, unlike the standard ΛCDM model, phantom quintessence models do not predict a long-lasting de Sitter-like future where Boltzmann brains can come into existence.

The second is through quintessence models where dark energy *decreases* in value over time instead. After the effects of dark energy weaken enough, gravity becomes dominant again, and the expansion slows down. If there is enough mass in the universe, gravity can by itself reverse the expansion, leading to a collapse of the universe in an eventual *Big Crunch*. A Big Crunch

the observer faster than the speed of light, as a result of the expansion of the universe). For this reason, Page (2008) has estimated that, if a false vacuum collapse is to solve the Boltzmann brain problem, our universe will decay within 20 billion years.

is also possible if there isn't enough mass for gravity to cause it on its own, but dark energy changes its behavior and switches from repulsive to attractive. Clearly, a Big Crunch would be as effective as a Big Rip in preventing an eternal de Sitter-like future.

The discussion above illustrates how effective realism is needed for interpreting both QFT and the ΛCDM model, and can also address the Boltzmann brain problem. For both theories, there are strong reasons to consider them effective rather than exact descriptions, given that they break down at extreme scales, we do not have empirical evidence for them at those scales, and have plausible (albeit speculative) completions that do not suffer a breakdown.

If the ΛCDM model is merely effective rather than an exact description, there is no reason to believe it will remain valid at the extremely long time scales where Boltzmann brains would numerically dominate. Instead, we expect it to be replaced by a deeper model that does not predict that the universe will end up in a cognitively unstable de Sitter-like phase. This model could be one that includes a vacuum transition that destroys the de Sitter-like phase, a quintessence model that leads to cosmic collapse, or something we have not yet conceived. Therefore, recognizing that the ΛCDM model has a limited domain of applicability offers a principled and relatively "mundane" way of resolving the philosophical challenges it poses.

7 Epistemological Breakdown, Confirmation, and Self-Location

As mentioned in Section 4, our initial credence for each hypothesis in the Labelscramble case is not the only factor that determines our posterior credence. For any "fishy" hypothesis F_k, the likelihood of the evidence E given F_k, P(E | F_k) also matters.

In particular, as Elga notes in the Labelscramble case, this means the hypothesis that you took Labelscramble is strongly disconfirmed. Assuming that Labelscramble causes one to hallucinate characters uniformly at random, the hallucinated text will almost always be nonsense.

So, letting L be the hypothesis that you took Labelscramble, P(E | L) is very small. Indeed, because it is so small, the odds P(E | L)/P(E | ¬L) are also small. This provides strong evidence against having taken Labelscramble and suggests that the evidence is explained by some alternative "fishy" hypothesis, for example, the hypothesis that pill bottles are randomly labeled while still matching a real drug.

Could the same kind of reasoning be applied to the Boltzmann brain problem? Specifically, might it help us to characterize the "breakdown" of the ΛCDM model when its validity is

extrapolated to very long timescales?

Again, let P be your probability function and E your evidence. Page (2024, p. 62) and Dogramaci and Schoenfield (forthcoming) argue that $P(E \mid T_1)$, where T_1 is a theory that predicts that Boltzmann brains dominate, is generally much smaller than $P(E \mid T_2)$, where T_2 is a theory compatible with E but does not contain a large number of Boltzmann brains. The reason is that Boltzmann brains, which are formed in largely random configurations, possess a range of possible experiences far broader than the range available to humans.

Given these considerations, we might be tempted to apply the moral of the Labelscramble case and conclude that the evidence disconfirms the "exact" completion of ΛCDM. This would offer a neat way of understanding the breakdown of ΛCDM at large timescales: extrapolated far enough, the model would be massively disconfirmed, so that, as an effective theory, ΛCDM would almost surely be realized by a different completion.

However, there is an important difference between the Labelscramble case and the one involving ΛCDM and Boltzmann brains, because self-location is a crucial feature of the latter, and it remains a vexed question how one ought to update self-locating probabilities.

In particular, the two cases work in parallel with respect to how hypotheses are disconfirmed only if we assume something like the "Self-Sampling Assumption" (SSA) (Bostrom 2002, p. 57). In plain terms, SSA holds that an observer should reason as if they were a random draw from the set of all actual observers a theory postulates. Under SSA, observers give higher weight to worlds where the proportion of observers whose total evidence matches theirs is higher.

However, SSA faces a number of well-known difficulties. Most prominently, the *Doomsday Argument* (Bostrom 2002, p. 89 ff., and references therein) seems to show that the rule predicts implausibly high odds for the extinction of the human race in the near future. Moreover, its recommendations are sensitive to what one takes to count as an observer (the *Reference Class Problem*; Bostrom 2002, chs. 10-11; Arntzenius and Dorr, ms.).

Two alternatives to SSA that have been explored in the literature are Compartmentalized Conditionalization (CC) and the Self-Indication Assumption (SIA). CC handles self-locating evidence as follows: whenever a hypothesis entails that there is at least one observer with your evidence, it keeps the hypothesis's original probability and divides it evenly among those observers. SIA instead boosts a hypothesis's probability in proportion to the number of observers whose evidence matches yours, so that hypotheses with more of those observers receive higher

posterior probability. Both rules have a built-in tendency to massively confirm hypotheses with a large number of observers (Isaacs et al. 2022)—a feature that has long been recognized as constituting an objection to SIA. Isaacs et al. (2022) show that even ordinary evidence produces this effect, yielding confirmation for large multiverses; for the same reasons, the two rules also confirm hypotheses according to which the universe contains a large number of Boltzmann brains.[8]

Again, the effect does not require any fancy sort of evidence. A peasant living in the Middle Ages, looking at a cloud with a specific shape, would, under CC and SIA, massively confirm the Boltzmann brain hypothesis. In other words, the deep skeptical implications of CC and SIA have little to do with cosmology. For that reason, I will set them aside in this paper. I assume instead that ΛCDM is either disconfirmed by the evidence, in the spirit of SSA, or, for generality, that it is not disconfirmed but also not massively confirmed by the evidence (this latter case could hold if some entirely new rule for updating self-locating probabilities, or an amended version of CC or SIA that avoids their most troubling features, turned out to be correct).

As noted above, the first case (where updating broadly follows SSA) is straightforward. In the second case, where ΛCDM in its exact form merely does not receive massive confirmation, there is still a sense in which it undergoes an epistemological breakdown. Its own confirmation is modest, while various alternative completions, for which we receive ordinary cosmological evidence, will typically receive much greater confirmation. Assuming that the prior in the "exact" ΛCDM is low (since it is a specific scientific theory and so we should not have too large a credence in it *a priori*, before obtaining any empirical evidence), its posterior will also be quite low.

This does not mean, however, that the same holds for the disjunction of all theories that predict the dominance of Boltzmann brains. Even if such theories individually undercut their own evidence, the combined prior (and therefore the combined posterior) for the whole disjunction need not itself be very low. Therefore, having a high credence that we are not Boltzmann brains requires an anti-skeptical assumption that assigns a low prior probability to being in any universe that generates Boltzmann brains. And while it is not obvious how to justify such an

[8]Theories that contain Boltzmann brains would also, under CC and SIA, be confirmed much more strongly than multiverse theories without Boltzmann brains dominance. The reason is that many potential multiverse theories are disconfirmed by the evidence, while no Boltzmann brain theories are (since, for any human experience, those theories contain some Boltzmann brains that experience it). I reserve a full discussion of this effect for further work.

assumption, this enterprise belongs more to the traditional anti-skeptical project of defending low priors for skeptical hypotheses than to the supposed empirical skeptical problem posed by cosmology.

8 Zero-Priors vs. Effective Realism

To summarize the discussion so far, our leading cosmological model appears to have a troubling implication: if we live in the kind of universe it describes, we could not have gained scientific evidence of its truth. To address this, Carroll has proposed assigning a probability of zero to any theory predicting such a situation. However, philosophers have raised objections to Carroll's proposal, and if we set it aside, it is desirable to aim for a solution that clarifies how our existing theories can remain evidentially well-founded while still avoiding the skeptical outcome.

I proposed the following resolution to this problem. Elga's analysis of the Labelscramble case, presented as an analogue of the Boltzmann brain problem, suggests that instability considerations do not make it reasonable to be confident that we are Boltzmann brains. At the same time, they also do not license the opposite conclusion that we can straightforwardly keep trusting ordinary science. Instead, Elga's analysis supports becoming confident that something "fishy" is going on without becoming particularly committed to any specific fishy hypothesis, for instance, that we live in a simulation, that we are brains in a vat, that there is a conspiracy to produce misleading science, or that our universe does not actually contain Boltzmann brains even though we receive cosmological evidence that, by ordinary standards, indicates that it does.

However, which hypothesis you adopt, among those that can explain our scientific evidence, will ultimately depend on how high your priors are for each of them. Relatively mundane hypotheses, therefore, deserve precedence over more 'fishy' ones. In this paper, my aim was to argue for the existence of such a mundane hypothesis: one that could explain why our scientific evidence seems to entail that Boltzmann brains numerically dominate over ordinary observers.

The mundane hypothesis I argued for is that the ΛCDM model should be considered an effective model, with a domain of applicability that does not extend to the timescales needed for producing a significant number of Boltzmann brains. On this view, the model is realized by a completion that does not imply Boltzmann brain domination. There is nothing "fishy" about this hypothesis; after all, the timescales in question are well outside our empirical reach, and

there are many candidate (albeit speculative) completions of the ΛCDM model according to which the number of Boltzmann brains is negligible.

Moreover, on the view developed here, the cognitive instability of the "exact" ΛCDM model places a limit on how much empirical support it can receive. Our posterior credence that the model holds at all time scales can never significantly exceed the prior credence we assign to that possibility before acquiring any cosmological evidence. By contrast, interpreting ΛCDM as an effective model allows it to accumulate ordinary confirmation within its intended domain, while skeptical scenarios tied to its unbounded extension continue to carry only low posterior probabilities.

My proposal can also handle the objections that have been raised against Carroll's solution. It resists Kotzen's objection because it does not require a logical connection between cognitive instability and a low prior probability. Rather, if the ΛCDM model is taken to be valid at all timescales, it is assigned a low prior probability (but not zero and not so low that its posterior could not become substantive if we obtained strong confirming evidence), because it is a specific scientific model that cannot be discovered a priori. Instead of being logically related to a very low prior, cognitive instability is related to the fact that, as mentioned above, for that model, its posterior probability is either reduced or at least remains close to its prior. This is because, in a universe with many Boltzmann brains, the physical beliefs we take ourselves to have good reasons for are far more likely to be the product of a random fluctuation.

Elga's objection also does not pose a problem for my proposal. Since the proposal does not postulate any logical connection between a theory's cognitive instability and its low prior probability, it does not imply that something must always prevent the construction of a setup like the one described by Elga (as presented above). Nor does it imply that, if such a setup could be constructed, we must assign a very low probability to the outbreak of a nuclear war.

The view also sheds some light on why Carroll's zero-prior proposal could seem intuitive at first sight and, in some respects, was not entirely misguided. Carroll thought that the ΛCDM model must be assigned a very low probability, and my view agrees with that, but only if the ΛCDM model is taken to hold at all timescales. And while I do not agree that the ΛCDM model must be rejected or that cosmologists must modify it until the Boltzmann Brain problem is solved (since effective theories are perfectly valid), the completion that in fact instantiates the ΛCDM model must not itself give rise to a Boltzmann brain problem.

This view also has the advantage that, if we reject Carroll's proposal, it does not mandate

discarding the ΛCDM model in favor of an alternative theory, while also clarifying its scientific status. When understood as an effective theory, the ΛCDM model has the standing of a fully legitimate scientific theory, much like quantum field theory. While discovering its true completion would be an important achievement, this does not imply that we must abandon the ΛCDM model in the meantime.

Finally, consider the following question. Could the possibility that our theories are merely effective act as a "mundane" solution against the Boltzmann brain problem in all possible scenarios? No, because in certain cases the putative Boltzmann brains would not be remote or existing outside any plausible domain of applicability of our theories. For instance, we could have direct evidence for the existence of Boltzmann brains. An example would be the scenario imagined by Dogramaci and Schoenfield (forthcoming), where I sample many brains that have exactly my mental state, and they all seem to be Boltzmann brains.

However, that is not a defect of my view. A case like the one above is closely related to more traditional skeptical scenarios. For instance, Elga (2004) recounts a story he attributes to Hartry Field, where a maintenance man, while dusting off some brains in vats, notices that they are being stimulated through a program labeled MAINTENANCE MAN DUSTS BRAINS IN VATS. For Field, this story illustrates a case where it is reasonable to suspect that one is a brain in a vat, and Elga concurs. In fact, Elga thinks (if the maintenance man has reasons to be convinced that some of the brains are in states that feel subjectively identical to his own) that Center Indifference vindicates that intuition. Similarly, in a case like the one described by Dogramaci and Schoenfield, it seems reasonable to suspect that something extremely "fishy" is going on.

9 Conclusion

In this paper, I argued that the Boltzmann brain problem does not force either skepticism about science or the introduction of new rules for confirmation. Drawing on Elga's Labelscramble analogy, I proposed a "mundane" hypothesis: ΛCDM should be treated as an effective rather than exact theory, with a domain of applicability that does not extend to the timescales that would lead to the numerical domination of Boltzmann brains over ordinary observers. On this reading, the apparent cognitive instability of ΛCDM only signals the need for a completion that does not produce Boltzmann brains at the relevant timescales.

The details depend on the correct way of handling self-locating evidence. If the Self-Sampling Assumption is broadly correct, our evidence massively disconfirms Boltzmann-brain-dominated completions, and no anti-skeptical priors are needed. If instead CC or SIA are correct, then Boltzmann-brain hypotheses receive massive confirmation from ordinary evidence, and we face skeptical consequences for reasons that have little to do with cosmology itself; for that reason, such issues are beyond the scope of a paper in the philosophy of cosmology like this one.

The remaining case is one in which the "exact" ΛCDM is neither massively confirmed nor massively disconfirmed. Even in this residual case, treating ΛCDM as an effective theory still avoids the conclusion that our evidence is overwhelmingly likely to come from Boltzmann brains. Ruling out skeptical hypotheses altogether, however, requires assigning low, though not Carroll-style zero, prior credence to the hypothesis that we live in a Boltzmann-brain universe.

Justifying such priors would require explaining how, even in this residual case, we can have a high *a priori* credence that the universe is not a Boltzmann-brain universe. I do not claim to solve that problem here, but it is philosophically interesting because it concerns a striking kind of knowledge: *a priori* knowledge about the physical constitution of the universe.

Acknowledgment. I thank Porter Williams, Robert Batterman, the audience at the BLOC Philosophy of Physics Workshop, and especially David Wallace, for helpful conversations and feedback.

References

Arntzenius, F. (2003). "Some Problems for Conditionalization and Reflection." *Journal of Philosophy* 100, 356–371.

Arntzenius, F., Dorr, C., ms. "What to Expect in an Infinite World."

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, New York.

Bostrom, N. (2003). "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53, 243–255.

Builes, D. (2024). "Center Indifference and Skepticism." *Nous* 58, 778–798.

Caldwell, R. (2002). "A Phantom Menace? Cosmological Consequences of a Dark Energy

Component with Super-Negative Equation of State." *Physics Letters B* 545, 23–29.

Caldwell, R., Kamionkowski, M., Weinberg, N. (2003). "Phantom Energy and Cosmic Doomsday." *Physical Review Letters* 91, 071301.

Carroll, S. (2017). "Why Boltzmann Brains Are Bad." arXiv:1702.00850 [hep-th].

Carroll, S. (2021). "Why Boltzmann Brains Are Bad." In Dasgupta, S., Weslake, B., Ravit, D. (eds.), *Current Controversies in Philosophy of Science*, 7–20. New York: Routledge.

Davenport, M., Olum, K. (2010). "Are There Boltzmann Brains in the Vacuum?" arXiv:1008.0808 [hep-th].

Dogramaci, S. (2020). "Does My Total Evidence Support that I'm a Boltzmann Brain?" *Philosophical Studies* 177, 3717–3723.

Dogramaci, S., Schoenfield, M. (forthcoming). "Why I Am Not a Boltzmann Brain." *Philosophical Review*.

Egan, A., Elga, A. (2005). "I Can't Believe I'm Stupid." *Philosophical Perspectives* 19, 77–93.

Elga, A., (2004). "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69(2), 383–396.

Elga, A. (2024). "Bracketing and Boltzmann Brains." Manuscript. Available at https://www.princeton.edu/~adame/bracketing-and-boltzmann-brains-for-rec/2024-04-15.9-bracketing-and-boltzmann-brains.pdf.

Elga, A. (forthcoming). "Boltzmann Brains and Cognitive Instability." *Philosophy and Phenomenological Research* 111(1), 127–136.

Fraser, J. D. (2018). "Renormalization and the Formulation of Scientific Realism." *Philosophy of Science* 85, 1164–1175.

Fraser, J. D. (2020). "Towards a Realist View of Quantum Field Theory." In French, S., Saatsi, J. (eds.), *Scientific Realism and the Quantum*, 276–292. Oxford: Oxford University Press.

Gibbons, G., Hawking, S. (1977). "Action Integrals and Partition Functions in Quantum Gravity." *Physical Review D* 15, 2752–2756.

Hartle, J., Srednicki, M. (2007). "Are We Typical?" *Physical Review D* 75.

Hartle, J., Srednicki, M. (2010). "Science in a Very Large Universe." *Physical Review D* 81.

Kitcher, P. (1993). *The Advancement of Science: Science Without Legend, Objectivity without Illusions*. Oxford: Oxford University Press.

Isaacs, Y., Hawthorne, J., Russell, J. S. “Multiple Universes and Self-Locating Evidence.” *Philosophical Review* 131(3), 241–294.

Kotzen, M. (2021). “What Follows from the Possibility of Boltzmann Brains?” In Dasgupta, S., Weslake, B., Ravit, D. (eds.), *Current Controversies in Philosophy of Science*, 21–34. New York: Routledge.

Lewis, D. (1979). “Attitudes De Dicto and De Se.” *Philosophical Review* 88, 513–543.

Page, D. (2008). “Is Our Universe Likely to Decay within 20 Billion Years?” *Physical Review D*, 063535.

Page, D. (2024). “Bayes Keeps Boltzmann Brains at Bay.” *Foundations of Physics* 54(5), 1–5.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Saad, B. (forthcoming). “Lessons from the Void: What Boltzmann Brains Teach.” *Analytic Philosophy*.

Talbott, W. (2020). “Is Epistemic Circularity a Fallacy?” *Philosophical Studies* 177, 2277–2298.

Wallace, D. (2006). “In Defence of Naiveté: The Conceptual Status of Lagrangian Quantum Field Theory.” *Synthese* 151, 33–80.

Wallace, D. (2011). “Taking Particle Physics Seriously: A Critique of the Algebraic Approach to Quantum Field Theory.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42, 116–125.

Williams, P. (2019). “Scientific Realism Made Effective.” *British Journal for the Philosophy of Science* 70, 209–237.

Worrall, J. (1989). “Structural Realism: The Best of Both Worlds?” *Dialectica* 43, 99–124.