DOI: 10.1111/phpr.70014

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Boltzmann brains and cognitive instability

Adam Elga

Princeton University, NJ, USA

Correspondence
Adam Elga, 1879 Hall Phil Dept., Princeton University, Princeton, NJ 08544, USA.
Email: adame@princeton.edu

Abstract

A *Boltzmann brain* is a randomly-formed configuration of matter that is conscious. According to some theories that cosmologists take seriously, the universe is so spatiotemporally large that it contains a great many Boltzmann brains that are duplicates of you. In the light of this it seems to follow that you should have significant confidence that you are a Boltzmann brain. What's worse, your situation seems to be "cognitively unstable": It seems unstable to end up confident that you are a Boltzmann brain because you should then think that your apparent cosmological evidence was randomly generated and hence that your confidence was unwarranted. But it also seems unstable to end up confident that you are not a Boltzmann brain because then you should follow your cosmological evidence to the conclusion that many Boltzmann brain duplicates of you exist, and hence that you are probably a Boltzmann brain. A case involving unreliable vision exhibits a similar threat of instability. A simple Bayesian model of that case, however, shows that the threat is an illusion. And a corresponding model suggests that the same goes for the threat of instability associated with Boltzmann brains.

KEYWORDS

Boltzmann brain, centered indifference, cognitive instability, principle of indifference, undermining

1 | INTRODUCTION

According to some theories that cosmologists take seriously, the universe is so spatiotemporally large that just about any finite configuration of matter will repeatedly form, simply due to random fluctuations (Carroll, 2020, p. 7). Call a universe like this a "large" universe. A *Boltzmann brain* is a randomly-formed configuration of matter that is conscious (at least for a little while). Since Boltzmann brains in a large universe are so numerous and varied, if our universe is large then countless Boltzmann brains will be duplicates or near-duplicates of you. You might be tempted to conclude that if the universe is large, you are likely a Boltzmann brain. For in a given large universe, consider the *centers* (specifications of an individual and a time) that might—for all your evidence goes—represent who you are and what time it is. It seems that the vast majority of these centers are associated with Boltzmann brains—for short, it seems that in large universes "Boltzmann brains dominate". And given that Boltzmann brains dominate it seems that you should be confident that you are a Boltzmann brain. So overall it seems that you should have significant confidence that you are a Boltzmann brain—about as much confidence as you have that the universe is large.

It seems crazy to have significant confidence that you are a randomly-formed configuration of matter. That is a problem. But in the light of the above considerations some have worried about an additional problem: that our cosmological evidence threatens to put us in a cognitively unstable state. My aim is to clarify the threat of instability and show that it is an illusion.

2 | THE ARGUMENT THAT YOU ARE LIKELY TO BE A BOLTZMANN BRAIN

How exactly is the argument that you are likely to be a Boltzmann brain supposed to go? Dogramaci (2020) usefully divides the argument into two steps. First, ordinary scientific evidence makes it reasonable to have significant credence that Boltzmann brains dominate.[1] Second, a statistical rule or a principle of indifference[2] entails that conditional on Boltzmann brains dominating, one should be confident that one is a Boltzmann brain.

There are many places to resist this argument.

At the first step theorists might deny that Boltzmann brains dominate on the grounds that your evidence rules out that you are a Boltzmann brain. Some may doubt that randomly-generated entities are conscious at all.[3] Others may invoke an externalist conception of evidence according to which humans have plenty of strong evidence about their environments, grounded in their interactions and memories of interactions with cellphones, tables, and so on—evidence that rules out their being Boltzmann brains (Williamson, 2000).[4]

[1] Here I follow the usage from the previous section, according to which "Boltzmann brains dominate" means: Of the centers in the actual universe compatible with your evidence, the vast majority are associated with Boltzmann brains. This formulation was inspired by Builes (2024, §7).

[2] For a relevant principle of indifference see Elga (2004) or the improved formulation in Builes (2024). Compare also the assumption "typicality" from Avni (2022, p. 961).

[3] One might doubt this on the ground that consciousness requires an appropriate evolutionary past, for example. Though one might still run into trouble concerning the hypothesis that one is part of a long-lived but nevertheless still randomly formed Boltzmann "bubble" (randomly formed mini-universe) of an intermediate size (Saad, forthcoming).

[4] Here, too, there is room for a back-and-forth about Boltzmann bubbles: see Saad (forthcoming).

At the second step theorists may wish to deny the relevant statistical rule or principle of indifference, and so deny the inference from "Boltzmann brains dominate" to "I'm likely a Boltzmann brain".

Having flagged these lines of resistance, I would like to set them aside for the moment. This reflects no prejudice against them—one of them may well be correct. It is rather to focus attention on an apparent instability.

3 | COGNITIVE INSTABILITY?

Several theorists have pointed out that following the above argument to the conclusion that you are a Boltzmann brain seems to leave you in a "cognitively unstable" or self-undermining state.[5] Here is how the instability might be thought to arise:

> On the one hand, you are confident that you are a Boltzmann brain on the basis of (apparent) cosmological evidence that the universe is large. On the other hand, you realize that Boltzmann brains have memories that were randomly generated and so are not to be trusted. Therefore, confidence that you are a Boltzmann brain rationally requires confidence that you have no reason to think that you are a Boltzmann brain. And this combination of attitudes is unreasonable. It is as unreasonable as being confident that the time is 8:17am solely on the basis of seeing a clock that reads 8:17am—even after learning that the clock was malfunctioning and displaying a randomly-determined time.[6]

If this is right, it is unreasonable to react to the cosmological evidence with confidence that you are Boltzmann brain. But it also seems unreasonable to remain confident that you are *not* a Boltzmann brain. For if you are not a Boltzmann brain, it seems that your memories and the cosmological evidence *are* to be trusted—and you should conclude that Boltzmann brains dominate. But then the statistical rule or principle of indifference applies, and you should think that you are a Boltzmann brain after all.

There seems to be no reasonable or stable resting place. What is going on?

4 | NO INSTABILITY IN A SIMPLE TEST CASE

To better understand the above threat of instability, here is a simple case that seems to threaten a similar instability:

[5] See for example Carroll (2020, p. 16), Winsberg (2012, p. 406), Myrvold (2016, p. 584), Dogramaci (2020, §3), Avni (2022, p. 960), Chalmers (2022, p. 658). The term "cognitively unstable" is from Carroll (2020), which cites discussions of similar self-undermining phenomena in Albert (2000, p. 116).

[6] Here I restate and expand a formulation from Carroll (2020, p. 16). Note that the argument does not require the general principle that confidence in a claim can *never* be rationally combined with confidence that your evidence offers little support for that claim. Nor does it require a general "level-bridging" or "rational reflection" principle (Christensen, 2010) of the sort put under pressure by critics of such principles (Christensen, 2024; Lasonen-Aarnio, 2014; Williamson, 2014). Instead it is enough for the argument to assume that in this particular case, confidence in claims based on one's apparent memories cannot be rationally combined with confidence that one's apparent memories were produced entirely at random. (Thanks here to an anonymous referee.).

You are fairly certain that the bottle in front of you contains only aspirin. So you swallow a pill from the bottle without reading the bottle's label. You've taken no other pills. When you look at the label you are surprised that it reads:

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

Upon looking at the label, should you become confident that you have taken Labelscramble? At first glance the case seems to threaten instability: On the one hand, if you believe that you have not taken Labelscramble, you have no reason to doubt your visual perception—so you should trust what you read and conclude that you have taken Labelscramble. But on the other hand, if you believe that you *have* taken Labelscramble, you would seem to have no basis for so believing (since your only basis for so believing seems to depend on trusting your visual impression of a label).

At second glance, a Bayesian model shows that the above reasoning is mistaken.[7] The case need not involve any instability at all.

To see why, let P be the probability function you have before you read the label, let E be the evidence you get when you read it, and let L be that you take Labelscramble. Assume that you are rational and that you update by conditionalization. In the light of E, how confident should you end up that you took Labelscramble?

That is not settled by the description given so far. It will be useful to fill in the details in several ways and analyze the resulting versions of the case. We'll see that in each version, no undermining or instability is present.

Start with a version of the case in which you are certain that unless you take Labelscramble, the bottle contains what it seems:

Version 1 Before you look at the label you are certain that Labelscramble is the only hallucinogen around, that you have consumed nothing but a pill from the bottle, that the bottle is accurately labeled, and that your visual perception is perfect unless you take a hallucinogen. As a result, you rule out scenarios in which you seem to see "Labelscramble…" without having taken Labelscramble.

In Version 1 you get decisive evidence that you took Labelscramble: $P(L|E) = 1$, because $P(E\&\bar{L}) = 0$. So it is reasonable for you to conclude that you took Labelscramble.

How can that be? Doesn't confirmation that you have taken Labelscramble amount to confirmation that your label-reading abilities are no good? And didn't we claim above that believing that your label-reading abilities are no good would leave you with no reason to doubt them? The answer is that you *do* have reason to doubt your label-reading abilities: namely, that your visual impression of "Labelscramble…" was antecedently much more likely to arise if you have taken Labelscramble than if not.[8] This reason is grounded in whatever evidence you had that ruled out scenarios in which you seem to see "Labelscramble…" without having taken Labelscramble. And this reason in no way depends on the thought that you can reliably read labels.

[7] This model is in the spirit of analyses given in Egan and Elga (2005, p. 81), Talbott (2020, p. 2295), and White (work in preparation, §2).

[8] In this sort of context, when I speak of one probability or conditional probability being "much more likely" than another I mean that the ratio of the one probability to the other is large (or undefined because only the latter probability equals zero). That is compatible with both probabilities being rather close to zero.

Moral: the apparent instability of becoming confident that you have taken Labelscramble is illusory. The illusion arises because it is counterintuitive that the deliverances of a faculty can simultaneously be evidence that the faculty is working improperly and also that it has happened to be correct in the present instance (Egan & Elga, 2005, p. 82).

The above version of the case involves extreme assumptions about your confidence. In it, that you take Labelscramble is the *only* "fishy" hypothesis you initially take seriously—the only hypothesis according to which your impression of the bottle's contents is not to be trusted. Suppose instead that you initially give some small credence to a large number of fishy hypotheses: that the bottle contains Labelscramble, that the bottle contain a particular other hallucinogen, that the bottle was randomly labeled, that your eyes are failing despite not having taken a hallucinogen, and so on. Given this background story, what should you think after reading the label?

Simplify matters by supposing that the fishy hypotheses $F_1, F_2, \ldots, F_k$ are mutually incompatible, and let F be that some fishy hypothesis or other is true. Plausibly, your evidence strongly confirms F. That is because E was much more to be expected given that something fishy was going on than not. For even though a fishy scenario was fairly unlikely to produce the particular visual impression "Labelscramble…", a non-fishy scenario—one in which you took no hallucinogen, the pill bottle is accurately labeled, your eyes weren't failing, and so on—was much less likely to do so.[9]

So seeing "Labelscramble…" should make you confident that something fishy is going on. But after seeing the label, how should you allocate your credence among the various fishy hypotheses? Suppose that each fishy hypothesis gets approximately the same initial credence. Then the fishy hypotheses that end up with the most credence are the ones according to which it was most to be expected that you would see "Labelscramble…". In particular, the final odds (ratio of probabilities) between any two fishy hypotheses F_a and F_b is given by the ratio $P(E|F_a)/P(E|F_b)$. And those ratios depend on further details about the case.

Let **Version 2** of the case be one in which seeing "Labelscramble…" was approximately as likely given one fishy hypothesis as it was given another. This would be plausible if each potential hallucinogen and each potential eye failure was equally likely to cause a hallucination of "Labelscramble…", and if this likelihood matched the probability of the bottle being mislabeled "Labelscramble…" given that it is mislabeled at all.

In Version 2 you should end up counting each of the k fishy hypotheses as approximately equally likely, and so should assign each of them a probability of approximately $1/k$. Since k was assumed to be large, this probability is fairly low, though not as low as it was before you looked at the label. So in this case, your evidence slightly increases your probability that you took Labelscramble, but not appreciably more than it increases your other probability in various competing fishy hypotheses.

How can that be? What about the instability argument from the previous section, according to which "if you believe that you have not taken Labelscramble, you have no reason to doubt your visual perception—so you should trust what you read and conclude that you have taken Labelscramble"? The answer is that you *do* have reason to doubt that the pill bottle contains what it appears to contain. The reason is that your visual impression of "Labelscramble…" was antecedently much more likely to arise if something fishy is going on than if not. It's just that even though each fishy hypothesis gets a boost in probability, there are so many of them that no one of them gets much of a boost.

[9] Here again "much less likely" is to be understood in terms of ratios of probabilities rather than differences: what is being claimed is that $P(E|F)/P(E|\bar{F})$ is very large. As a result, getting E multiplies your odds for F by a very large factor, and hence plausibly ends up being large as well.

Moral: the apparent instability of remaining doubtful that you have taken Labelscramble is illusory. The illusion arises because it is tempting to think that if your evidence doesn't significantly increase the probability of any particular fishy hypothesis, it also doesn't significantly increase the probability that something or other fishy is going on.

So far we have seen a case (Version 1) in which you get decisive confirmation that you have taken Labelscramble, and hence decisive confirmation of a particular fishy hypothesis. We have seen a case (Version 2) in which you get slight confirmation that you have taken Labelscramble, but also strong confirmation that something fishy is going on. Now let us turn to a case in which you get strong *disconfirmation* that you have taken Labelscramble while still getting strong confirmation that something fishy is going on:

Version 3 Like Version 2 except that you realize the following. When pill bottles are randomly labeled (let this be hypothesis F_1), the label always matches the label of a real drug, determined uniformly at random. But when Labelscramble causes one to hallucinate a label (hypothesis L), the hallucinated text is a sequence of 100 characters determined uniformly at random (and hence is almost always gibberish).

In Version 3, E was vastly more to be expected on the hypothesis that the pill bottles are randomly labeled than on the hypothesis that you take Labelscramble. In other words, the ratio $P(E|L)/P(E|F_1)$ is miniscule. That is because there are many orders of magnitude more 100-character strings than there are drugs. Indeed, since $P(E|L)$ is so very small, the ratio $P(E|L)/P(E|\bar{L})$ is also miniscule. So in this version of the case, even though seeing "Labelscramble…" counts as strong evidence that something fishy is going on, it also counts as almost decisive evidence that you did *not* take Labelscramble.[10] And here, too, there is no threat of instability.

These three versions of the case are of course not exhaustive. But they illustrate why both pieces of the instability argument are mistaken, even though each is tempting.

5 | NO INSTABILITY IN THE BOLTZMANN BRAIN CASE

Now return to the apparent threat of undermining in the Boltzmann brain case. The threat, recall, was this: It is unstable to end up confident that you are a Boltzmann brain because then you should think that your apparent cosmological evidence was randomly generated and hence that your confidence was unwarranted. And it is unstable to end up confident that you are not a Boltzmann brain because then you should follow your cosmological evidence to the conclusion that Boltzmann brains dominate, and hence that you are probably a Boltzmann brain.

It is the aim of this section to carry over lessons from the discussion of the Labelscramble case to show that neither part of the above argument is correct. However, the Boltzmann brain case differs from the Labelscramble example in several ways. First, the Boltzmann brain case essentially involves self-locating uncertainty. And it is a vexed question what prior probabilities are appropriate and how to update one's probabilities in such cases. Second, the argument for instability in the Boltzmann brain case depends on a contested statistical rule or principle of indifference.

[10] This comparison is meant to set the stage for a point due to Page (2024, p. 62) and Dogramaci and Schoenfield (forthcoming, §2a) about the range of Boltzmann brain experiences, a point that will be discussed in §5.

A general treatment of the Boltzmann brain case would have to grapple with both of these complexities, and more. But we needn't do so in order to get evidence that the above argument (that instability is inevitable) is no good. For we can follow the lead of Dogramaci and Schoenfield (forthcoming), Page (2024), and Wallace (2023) and use a simple Bayesian model to analyze the Boltzmann brain case—just as we used such a model to analyze the Labelscramble case in the previous section.[11]

Let P be your prior probability function and let E be your evidence (including the apparent cosmological evidence that the universe is very large). Assume that you are rational and that your probability function is the result of conditionalizing your prior on your evidence. Grant for the sake of the argument that in a large universe, Boltzmann brains dominate. (If not, there is not even an apparent threat of cognitive instability.)

Introduce some terminology:

Small The universe is small enough that no Boltzmann brains exist.
Large The universe is so large that many Boltzmann brains exist.
Human You are an ordinary human.
BB You are a Boltzmann brain.

As noted above, the proper analysis of the case depends on whether an appropriate principle of indifference is true. Rather than weighing in on whether such a principle is true, let us consider what happens either way.

First assume that given your evidence and that the universe is large, it is reasonable to be highly confident that you are an ordinary human: $P(\text{Human}|E\&\text{Large}) \approx 1$. To assume this is to deny (in an extreme way) the relevant principle of indifference. Under that assumption there is no obstacle to ending up with significant probability that you are an ordinary human inhabiting a large universe. For example, you end up with significant probability in this if you follow ordinary scientific standards and count E as supporting significant credence that the universe is large.[12] No threat of undermining or instability is in sight.[13]

Next assume instead that the relevant principle of indifference is true. It follows that given your evidence and that the universe is large, it is reasonable to be highly confident that you are a Boltzmann brain: $P(\text{BB}|E\&\text{Large}) \approx 1$. It further follows that given E, you should end up highly doubtful that you are an ordinary human inhabiting a large universe:[14]

[11] Proceeding in this way is not neutral on the contested question of how to update self-locating probabilities, since the discussion below implicitly assumes an update rule in the spirit of the "Self-Sampling Assumption" (Bostrom, 2002, p. 57). Since it is unclear which update rule is correct, a fuller discussion would evaluate the argument according to various competing rules. The present approach still makes progress since many of the considerations raised below arise for many candidate updating rules.

[12] Suppose that you violate the principle of indifference in this extreme way: $P(\text{Human}|E\&\text{Large}) \approx 1$. And suppose that you follow ordinary scientific standards and count E as supporting significant credence that the universe is large: $P(\text{Large}|E)$ is substantial. It follows that your credence that you are an ordinary human inhabiting a large universe, $P(\text{Large}\&\text{Human}|E)$, is also substantial. Proof: $P(\text{Large}\&\text{Human}|E) = P(\text{Large}|E) \cdot P(\text{Human}|E\&\text{Large}) \approx P(\text{Large}|E) \cdot 1 = P(\text{Large}|E)$. So if $P(\text{Large}|E)$ is substantial, then so is $P(\text{Large}\&\text{Human}|E)$.

[13] Responses to the Boltzmann brain problem that favor this route include Dogramaci (2020) and Dogramaci and Schoenfield (forthcoming).

[14] Proof: Since $P(\text{BB}|E\&\text{Large}) \approx 1$, we have that $P(\text{Human}|E\&\text{Large}) \approx 0$. Therefore $P(\text{Large}\&\text{Human}|E) = P(\text{Large}|E) \cdot P(\text{Human}|E\&\text{Large}) \approx P(\text{Large}|E) \cdot 0 = 0$.

So: you should end up highly doubtful that you are an ordinary human inhabiting a large universe. But what situation should you think you are in? Your credence should be distributed among competing hypotheses, two of which are:

SmallHuman The universe is small, no Boltzmann brains exist, and you are an ordinary human in unexceptional circumstances.

LargeBB The universe is large, many Boltzmann brains exist, and you are a Boltzmann brain.

Start by looking at the ratio $P(E|\text{SmallHuman})/P(E|\text{LargeBB})$. This ratio measures the relative support that E provides to SmallHuman over LargeBB. $P(E|\text{SmallHuman})$ is fairly small both because E is very specific and because given that you are a human inhabiting a small universe, it is not much to be expected that you receive cosmological evidence indicating (according to normal cosmological standards) that the universe is large.

However, as Page (2024, p. 62) and Dogramaci and Schoenfield (forthcoming, §2a) point out, there is a case to be made that $P(E|\text{LargeBB})$ is much, much smaller than $P(E|\text{SmallHuman})$. For it might be thought that the range of potential Boltzmann brain experiences is much wider than the range of potential human experiences. As a result, any particular experience within the human range is vastly more to be expected given SmallHuman than given LargeBB. The idea is that the situation is analogous to Version 3 of the Labelscramble case, in which seeming to see "Labelscramble…" was vastly more to be expected on the hypothesis that the bottle was randomly assigned the label of a real drug than on the hypothesis that the bottle was assigned a random 100 character string.

If this is correct, then E massively confirms SmallHuman over LargeBB. So unless LargeBB gets a prior probability that is many orders of magnitude greater than the prior probability of SmallHuman, SmallHuman ends up with a much higher posterior probability than LargeBB. Assuming that no other hypotheses are in the running, the upshot is that in the light of the cosmological evidence you should be confident that the universe is small and that there are no Boltzmann brains.[15]

What about the argument that being confident that you are not a Boltzmann brain is unstable because it leads to the conclusion that the universe is large, and hence that you are probably a Boltzmann brain? Answer: that argument is unsound. For under the present assumptions, the antecedent unlikeliness of receiving E given that the universe is small is swamped by the *extreme* unlikeliness of receiving E given that the universe is large. So being confident that you are not a Boltzmann brain does not lead to the conclusion that the universe is large, and hence does not lead to instability.

On the above pattern of assumptions, even though one's evidence includes cosmological considerations that seem to point in the direction of a large universe, one should conclude that the universe is small. But let us adopt assumptions to give the argument for instability the best possible chance of succeeding. In particular, suppose that the cosmological evidence seemingly supporting a large universe is *tremendously* strong. Suppose that it is so strong that $P(E|\text{SmallHuman})$ is much smaller than $P(E|\text{LargeBB})$ (which is already miniscule). This is an extreme modification—perhaps an impossible one. But let us make it for the sake of the argument.

Even under these extreme assumptions the case for instability does not succeed. Note first that under the present assumptions, the ratio of $P(E|\text{LargeBB})$ to $P(E|\text{SmallHuman})$ is large—so E considerably confirms LargeBB over SmallHuman. Suppose that the two hypotheses get similar prior probabilities. It follows that LargeBB ends up with a higher posterior probability than SmallHuman. Does this mean that you should end up confident that you are a Boltzmann

[15] This is the conclusion suggested by Page (2024, p. 62).

brain? That depends on the status of competing hypotheses that we have not yet discussed. These hypotheses, like LargeBB, are "fishy" in the sense that in them your evidence is highly misleading. Such hypotheses include, for example:

Conspiracy You are the object of an elaborate conspiracy: everyone you have ever interacted with is an actor and all scientific testimony you've heard is manufactured and inaccurate.
Insane You are insane, suffering a massive episode that only has the illusion of coherence.
BIV You are a brain in a vat operated by alien scientists.

We have not so far discussed such hypotheses because on previous assumptions they have had prior probabilities and levels of confirmation low enough that they could be reasonably disregarded. But on the present assumptions that needn't be true.

That is because on present assumptions, $P(E|\text{SmallHuman})$ is *tremendously* small. And it may well be that, for example, $P(E|\text{Conspiracy})$ is not as small. If so, E might confirm Conspiracy over LargeBB strongly enough that Conspiracy ends up more likely than LargeBB, despite having had a lower prior probability. And the same might be true for many other fishy hypotheses. If so, it could be that your evidence slightly increases your probability in LargeBB but also increases your probabilities of various competing fishy hypotheses to a similar level. Then as in Version 2 of the Labelscramble case, you should think: "Some fishy hypothesis or other is true, but I am unsure which one". And as in that case, you should not put much credence in any particular fishy hypothesis. So you should end up confident that you are not a Boltzmann brain.

What about the argument that being confident that you are not a Boltzmann brain is unstable because it leads to the conclusion that the universe is large, and hence that you are a Boltzmann brain? Answer: Under the present assumptions, you highly doubt that the universe is large.

Is there a pattern of assumptions according to which your evidence *should* make you confident that you are a Boltzmann brain? Not a plausible pattern, in my view. But as another concession for the sake of the argument, suppose (implausibly) that E is vastly more to be expected given LargeBB than given *any* competing fishy hypothesis: $P(E|\text{LargeBB}) \gg P(E|F_i)$ for each competing fishy hypothesis F_i. Provided that the number of fishy hypotheses is not gargantuan, we finally have a version of the case in which E may support significant confidence in LargeBB. In effect we have reproduced the structure of Version 1 of the Labelscramble case.

But isn't this unstable? Doesn't confirmation that you are a Boltzmann brain amount to confirmation that you have no reason to think that you are a Boltzmann brain? The answer is that (under the given assumptions) you do have reason to think you are a Boltzmann brain: namely, that your evidence was antecedently much more likely to arise if LargeBB is true than if not. And this reason in no way depends on the thought that you are an ordinary human, just as in Version 1 of the Labelscramble case your reason for thinking that you took Labelscramble in no way depends on the thought that you can reliably read labels. So even these tremendously implausible concessions do not produce a case that involves instability.

The above patterns of assumptions about the Boltzmann brains case are of course not exhaustive. And the discussion does not settle how we should react to our cosmological evidence. But it illustrates why both pieces of the instability argument are mistaken, even though each is tempting.[16]

[16] In memory of William Talbott, 1949–2023. An earlier version of this paper was presented at the 2024 Rutgers Epistemology Conference. For helpful conversations and feedback, thanks to Tyler Brooke-Wilson, David Builes, David Elga, Samuel Elga, Jake Nebel, Gideon Rosen, Nico Silins, participants in the 2024 Rutgers Epistemology Conference, and two anonymous *PPR* referees. Thanks to Charlotte Elga for advising at a crucial sticking point: "Just sit down and write your paper, Dad."

ORCID

Adam Elga https://orcid.org/0000-0002-1354-6032

REFERENCES

Albert, D. Z. (2000). *Time and chance*. Harvard University Press.

Avni, R. (2022). The Boltzmann brains puzzle. *Noûs*, *57*(4), 958–972. https://doi.org/10.1111/nous.12439

Bostrom, N. (2002). *Anthropic bias: Observation selection effects in science and philosophy*. Studies in philosophy. Routledge, New York.

Builes, D. (2024). Center indifference and skepticism. *Noûs*, *58*(3), 778–798. https://doi.org/10.1111/nous.12478

Carroll, S. M. (2020). Why Boltzmann brains are bad. In S. Dasgupta, R. Dotan, & B. Weslake (Eds.), *Current controversies in philosophy of science* (pp. 7–20). Routledge.

Chalmers, D. J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. W. W. Norton & Company.

Christensen, D. (2010). Rational reflection. *Philosophical Perspectives*, *24*, 121–140.

Christensen, D. (2024). Epistemic akrasia: No apology required. *Noûs*, *58*, 54–76. https://doi.org/10.1111/nous.12441

Dogramaci, S. (2020). Does my total evidence support that I'm a Boltzmann Brain? *Philosophical Studies*, *177*(12), 3717–3723. https://doi.org/10.1007/s11098-019-01404-y

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*(1), 77–93. https://doi.org/10.1111/j.1520-8583.2005.00054.x

Elga, A. (2004). Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research*, *LXIX*(2), 383–396.

Lasonen-Aarnio, M. (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, *88*(2), 314–345. https://doi.org/10.1111/phpr.12090

Myrvold, W. C. (2016). Probabilities in statistical mechanics. In A. Hájek & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199607617.013.26

Page, D. N. (2024). Bayes keeps Boltzmann brains at bay. *Foundations of Physics*, *54*(5), 62. https://doi.org/10.1007/s10701-024-00791-5

Saad, B. (forthcoming). Lessons from the void: What Boltzmann brains teach. *Analytic Philosophy*. https://doi.org/10.1111/phib.12353

Talbott, W. J. (2020). Is epistemic circularity a fallacy? *Philosophical Studies*, *177*(8), 2277–2298. https://doi.org/10.1007/s11098-019-01310-3

Wallace, D. (2023). A Bayesian analysis of self-undermining arguments in physics. *Analysis*. https://doi.org/10.1093/analys/anac096

White, R. (work in preparation). Who's afraid of Boltzmann brains?.

Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.

Williamson, T. (2014). Very improbable knowing. *Erkenntnis*, *79*(5), 971–999. https://doi.org/10.1007/s10670-013-9590-9

Winsberg, E. (2012). Bumps on the road to here (from Eternity). *Entropy*, *14*(3), 390–406. https://doi.org/10.3390/e14030390

How to cite this article: Elga, A. (2025). Boltzmann brains and cognitive instability. *Philosophy and Phenomenological Research*, 1–10. https://doi.org/10.1111/phpr.70014