Bracketing and Boltzmann Brains

Adam Elga

Version of April 15, 2024 *

1

According to some theories that cosmologists take seriously—call them "large cosmologies"—the universe is so spatiotemporally large that just about any finite configuration of matter will repeatedly form, simply due to random fluctuations. A *Boltzmann Brain* (abbreviation: "BB") is such a randomly-formed configuration that is conscious (at least for a little while). If a large cosmology is true, then BBs are so numerous and varied that the vast majority of the entities in your subjective state (in the same phenomenal state and having the same apparent memories as you) are BBs rather than humans. It seems to follow that you should have significant confidence that you are a BB—about as much confidence as you have that a large cosmology is true.

But it also seems crazy to have significant confidence that you are a randomly-formed configuration of matter. That is the problem of Boltzmann Brains.

My plan is to assess whether and how considerations of instability and self-undermining can help us address that problem.

2

How exactly is the argument that you are likely to be a BB supposed to go? Dogramaci (2020) usefully divides the argument into two steps. First, ordinary scientific evidence makes it reasonable to have significant credence that the universe has a great many BBs—so many that the overwhelming

*For the 2024 Rutgers Epistemology Workshop. In memory of William Talbott, 1949–2023.

proportion of entities in your subjective state are BBs. Second, a statistical rule or a principle of indifference entails that conditional on the overwhelming proportion of entities in your subjective state being BBs, one should be confident that one is a BB.

There are many places to resist this argument. At the first step one might deny that randomly-generated entities are conscious.[1] Or one might posit a constraint on self-locating rational credences according to which your evidence disconfirms a cosmological hypothesis when that hypothesis entails that an especially small fraction of observers have evidence that matches yours.[2]

At the second step, externalists about evidence will not be tempted by statistical rules or principles of indifference stated in terms of subjective states (Williamson 2000). Instead it would be natural for them to say that while humans have plenty of strong evidence about their environments, grounded in their interactions and memories of interactions with apples, tables, and so on, BBs—having never interacted with such things—have evidence that is extremely impoverished. And there is little pressure in favor of a principle of indifference that requires dividing one's credence equally among predicaments that one's evidence distinguishes between.[3]

Having flagged these lines of resistance, I would like to set them aside. This reflects no prejudice against them—one of them may well be the right way to solve the problem of Boltzmann Brains. It is rather to focus attention

[1]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 was part of a long-lived but nevertheless still randomly formed Boltzmann "bubble" (randomly formed mini-universe) of an intermediate size (Saad Forthcoming).

[2]Arntzenius and Dorr (2017) proposes a constraint with similar consequences (when combined with suitable background assumptions). See also Bostrom (2002), Kotzen (2020). Depending on one's prior credences, large cosmologies might deserve little credence given a constraint of this kind because according to them, such a small fraction of observers have qualitative evidence that matches yours.

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

on the viability of two strategies that appeal to undermining and instability.

3

Several theorists have pointed out that following the BB argument to the conclusion that you are a BB seems to leave you in an unstable or self-undermining state. Here is how the instability is supposed to arise:

On the one hand, you are confident that you are a BB on the basis of (apparent) cosmological evidence about the size of the universe. On the other hand, you realize that BBs have memories that were randomly generated and so are not to be trusted. Therefore, confidence that you are a BB rationally requires confidence that you have no reason to think that you are a BB. And this combination of attitudes (confidence in a claim combined with confidence that your evidence offers no support for that claim) is unreasonable.[4]

In response it has been proposed (Carroll 2020, Chalmers 2022) that theories that lead to this sort instability get very low prior probability—that it is reasonable to essentially rule them out at the start of inquiry. If this is right, then the first step of the BB argument—that ordinary scientific evidence makes reasonable significant credence that the universe has a great many BBs—fails:

> The best we can do is to decline to entertain the possibility that the universe is described by a cognitively unstable theory, by setting our prior for such a possibility to zero (or at least very close to it). [...] If we discover that a certain otherwise innocuous cosmological model doesn't allow us to have a reasonable degree of confidence in science and the empirical method, it makes sense to reject that model [...]. This includes theories

[4]One might say that such combinations of attitudes can be rational (Christensen 2024, Lasonen-Aarnio 2014, Williamson 2014), but let us grant the relevant "level-bridging", "rational-reflection", or "anti-akrasia" principle for the sake of the argument.

> in which the universe is dominated by Boltzmann Brains and other random fluctuations. (Carroll 2020, 17)

While it is far from decisive, I would like to register a worry about this proposal. The worry is that the spirit of the proposal would naturally extend beyond cases that involve BBs, to other cases in which large numbers of highly-misled folks—such as computer-simulated creatures—come to exist. In those cases, we humans might have control over whether or how many such creatures exist. If so, implementing the "zero-prior" proposal requires making a choice: is it hypotheses according to which undermining scenarios *might* arise that get zero prior probability, or ones in which undermining scenarios *actually* arise? If the former, then the proposal ends up ruling out many scenarios in which there isn't any actual undermining (because no misled creatures end up being created). That seems unwarranted. If the latter, then the proposal would lead to a strange result, as follows.

Assume that scenarios in which many misled creatures are created gets prior probability near zero. We might try to use this fact to manipulate the future. Suppose that we resolve to ensure that [many misled creatures get created] if and only if [humanity ever starts a nuclear war]. Assuming that we have control over the creation of misled creatures, there would seem to be no in-principle obstacle to our doing this. We would set up a default condition in which no misled creatures are created, and also set up sensors that reliably trigger the creation of many misled creatures if there is ever a nuclear war. By the prior-zero assumption, we can rule out that many misled creatures get created. But then we must expect one of the following: (1) all attempts to effectively make the above default-and-sensor arrangements will be mysteriously and persistently thwarted, however carefully planned and executed, or (2) we are able to make the arrangements, and no nuclear wars ever occur.

Expectation (1) would involve an unprecedented and seemingly conspiratorial limit to our engineering powers. Expectation (2) would mean that the above strategy would indeed be a utility-maximizing way to pre-

vent nuclear war (and—using similar arrangements—force or prevent any future condition that we can reliably test for). Either expectation would be quite strange.

4

Dogramaci (2020) also appeals to instability considerations to block the BB conclusion, but employs a different strategy—call it the "instability-blocks-undermining" strategy. Recall that the second step of the BB argument moves from the intermediate conclusion that the overwhelming proportion of entities in your subjective state are BBs to the final conclusion that you are a BB. The move is supposed to be underwritten by a statistical rule. The instability-blocks-undermining strategy is to say that the statistical rule does not apply in this case. Why? Because your (ordinary scientific) evidence "includes in it lots that says [you have an] ordinary human [body], on ordinary earth, which has existed and circled the sun for billions of years" (Dogramaci 2020, 3719). So (the proposal continues) your total evidence strongly supports that you are an ordinary human (and not a BB). And this is true even if your total evidence *also* supports that there are many BBs. Why doesn't evidence that there are many BBs compromise, negate, defeat, or undercut your evidence that you are an ordinary human? Because

> my belief that I'm a BB would have to be entirely based on, and epistemically dependent for its justification on, the ordinary scientific evidence that the universe hosts zillions of BBs. This means that no matter how hard I try to take on the belief that I'm a BB, I cannot rationally do so while kicking away my beliefs in ordinary science, for those beliefs in science were the basis for the conclusion that I'm a BB! [. . .] This, I suggest, explains why, in the BB case, the ordinary scientific evidence must remain a part of my total evidence [. . .]. (Dogramaci 2020, 3720)

The appeal to instability here is indirect, and is best understood by way

of a contrast with another case—call it the **apple drug case**. Suppose that I see an apple on a table and then am informed that I have participated "in an experiment where [I have likely been] given a drug that [induces] a hallucination as of an apple on a table" (Dogramaci 2020, 3720). In this case the instability-blocks-undermining theorist can grant that the information about the experiment significantly undermines your visual evidence that there is an apple on the table. They can grant that you should think "my evidence suggests that I may well be hallucinating, and so I should have less trust in how things visually appear to me". But the theorist will point out what they take to be a crucial difference between the apple drug case and the BB case:

> In the drug case, my knowledge that I'm participating in the drug trial, and my belief that I'm hallucinating, are in no part based on, or dependent for their justification on, my views about whether or not there's an apple on the table. By contrast, my belief that I'm a BB would have to be entirely based on, and epistemically dependent for its justification on, the ordinary scientific evidence that the universe hosts zillions of BBs. (Dogramaci 2020, 3721)

In other words, in the apple drug case:

> Evidence that you've taken the drug *does undermine* your evidence that there is an apple
>
> because
>
> the thought that you are on the drug *does not depend* on the thought that there is an apple.

But in the BB case:

> Scientific evidence that there are many BBs *does not undermine* your ordinary scientific beliefs

because

the thought that there are many BBs *does depend* on your ordinary scientific beliefs.

That is the proposal.

5

To assess this proposal it will help to explore the instability phenomenon on which it depends. Recall that in the apple drug case, your thought that you have taken the hallucinogenic drug does not depend on the thought that there is an apple in front of you. That is why (according to the proposal) the evidence that you have taken the drug does undermine your visual evidence that there is an apple.

Here is a modified case in which your evidence that you have taken a hallucinogenic drug *does* depend on your visual impressions:

The undermining drug case Feeling adventuresome, you swallow a pill from the bottle in front of you without reading the bottle's label. You've taken no other pills.

When you look at the label it seems to read:

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

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 should trust what you read and conclude that you have taken it. But on the other hand, if you believe that you *have* taken it, 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 analysis (in the spirit of Egan and Elga (2005, 81) and Talbott (2020, 2295)) shows that the above reasoning is mistaken. The case need not involve any instability at all: upon looking at the bottle, you may stably and significantly increase your confidence that you have taken Labelscramble. Let me explain.

Let P be your probability function before you read the label and introduce a few named propositions:

L You have taken a Labelscramble pill.

R Your visual impression of the label was randomly generated.

V The label visually appears to you to read "Labelscramble (50mg)…".

The question is: how much (if at all) does your visual impression of the label (V) confirm that you have taken Labelscramble (L)? Assuming that you update by conditionalization and making other reasonable assumptions about the case,[5] the answer[6] is that your odds for L (i.e., $P(L)/P(\overline{L})$) get multiplied by a factor of at least $P(R \mid L)/P(R \mid \overline{L})$. But this ratio is much greater than 1, since you count it much more likely that your visual

[5]Assumptions:

1. You are certain that you have consumed nothing but a pill from the bottle and that the bottle is accurately labeled.
2. You are certain that: your visual impression of the label is either randomly generated or perfectly accurate. (It follows from (1) and (2) that $P(V \mid \overline{R}L) = 1$ and $P(V \mid \overline{R}\overline{L}) = 0$.)
3. You count it much more likely that your visual impression was randomly generated if you have taken Labelscramble than if you have not (i.e., $P(R \mid L)/P(R \mid \overline{L}) \gg 1$).
4. Conditional on your visual impression being randomly generated, you count it equally likely to be "Labelscramble…" whether or not you have taken a Labelscramble pill (i.e., $P(V \mid RL) = P(V \mid R\overline{L})$).

[6]The odds form of Bayes' theorem says that your new odds equals your old odds times

impression was randomly generated if you have taken a Labelscramble pill than if not. So according to this analysis, your visual impression of the label significantly confirms that you have taken a Labelscramble pill.

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 note 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.

Bottom line: upon looking at the label you may stably conclude: "My visual experience is evidence both that I have taken Labelscramble *and also* that I hallucinated seeing a Labelscramble label."

I admit that the above analysis can leave a lingering impression of paradox. The following case is intended to dispel that impression. Consider the **random-or-perfect case**, which is the undermining drug case with an added assumption:

> Before you look at the label you are rationally *certain* of the following: your visual perception will be randomly-generated if you have taken Labelscramble and will be perfectly accurate

---

the likelihood ratio:

$$\frac{P(L \mid V)}{P(\overline{L} \mid V)} = \frac{P(L)}{P(\overline{L})} \cdot \frac{P(V \mid L)}{P(V \mid \overline{L})}.$$

But this likelihood ratio is much greater than 1 because:

$$\begin{aligned}
\frac{P(V \mid L)}{P(V \mid \overline{L})} &= \frac{P(R \mid L)P(V \mid RL) + P(\overline{R} \mid L)P(V \mid \overline{R}L)}{P(R \mid \overline{L})P(V \mid R\overline{L}) + P(\overline{R} \mid \overline{L})P(V \mid \overline{R}\overline{L})} \\
&\geq \frac{P(R \mid L)P(V \mid RL)}{P(R \mid \overline{L})P(V \mid R\overline{L})} && \text{(since } P(V \mid \overline{R}\overline{L}) = 0\text{)} \\
&= \frac{P(R \mid L)}{P(R \mid \overline{L})} \gg 1.
\end{aligned}$$

otherwise.

In the random-or-perfect case, before you look at the label you are rationally certain that you will not seem to read "Labelscramble…" unless you are hallucinating. So together with the proposition that the label *does* appear to read "Labelscramble…", your evidence makes it rational for you to be certain that you have taken Labelscramble.[7]

Nagging concern: "But if you are certain that your visual impressions of labels are hallucinations, how can it be rational to trust your visual impression of a label?" Reply: the just-mentioned route to the conclusion that you have taken Labelscramble does not rely on your trusting your visual perception of labels. Instead, what justifies the conclusion is V (that you *seem* to see "Labelscramble…"), together with whatever prior evidence you had that ruled out scenarios in which you seem to see "Labelscramble…" without having taken Labelscramble. And nothing in the case requires that this prior evidence depended on your being able to reliably read labels.

6

In the undermining drug case, your visual impression prompted some disillusionment. You initially trusted your ability to read labels and were forced to give up that trust. But the disillusionment was focused on an extremely narrow domain: *just* your ability to read pill-bottle labels. The narrowness of the disillusionment resulted from two factors. First, Labelscramble was assumed to have an extremely limited effect. Second, in order to simplify the analysis it was assumed that your having taken Labelscramble was the *only* "fishy" hypothesis that you put any credence in. For example, you put no credence in the hypothesis that the pill bottle was mislabeled, or that Labelscramble didn't have the effects advertised.

Relaxing these assumptions opens the door to cases that prompt significantly more diffuse disillusionment. For example, consider a **diffuse under-**

[7]Symbolically: $P(V\bar{L}) = 0$ and so $P(L \mid V) = 0$. Thanks here to Tyler Brooke-Wilson.

mining case that is like the undermining drug case except that you initially put some credence in a wide range of hypotheses according to which the setup is not exactly as was initially described—such as the hypothesis that the pill bottle was mislabeled, or the hypothesis that Labelscramble doesn't have the effects advertised, or .... Call these hypotheses "fishy". Then (given appropriate assumptions about your priors), seeming to see "Labelscramble..." should prompt a more diffuse sort of disillusionment: instead of concluding that you took Labelscramble, you should conclude that *either* you took Labelscramble, or the bottle was mislabeled, or ... . You might in such a case say "*Something* fishy is going on here, but I'm not sure exactly what." And your credence would be increased a bit in each fishy hypothesis, but you wouldn't count any one of them as particularly likely.

Now turn back to the case of cosmological theories that seems to support that there are many Boltzmann Brains. The instability-blocks-undermining theorist, recall, reasons as follows:

1. It would be unreasonable to become confident that you are a BB on the basis of the deliverances of cosmological theories, since confidence that you are a BB would rationally require you to reject those theories (along with all of ordinary science).

2. Therefore, it is permissible for you to keep trusting ordinary science and have significant confidence that there are many BBs, without being correspondingly confident that you are a BB.

In the light of the analysis of the undermining drug cases, I am persuaded of claim (1): it would indeed be unreasonable for you to be confident that you are a BB. Doing so would be as bad as in the diffuse undermining case being very confident that you had taken Labelscramble. But claim (2) does not follow. For all that has been said, a rational reaction in the BB case is analogous to the rational reaction in the diffuse undermining case: becoming confident that something fishy is going on (and very slightly

increasing your credence that you are a BB), without becoming confident that you are a BB.

Which fishy hypotheses you should end up seriously considering will of course depend on your initial credence in them, and on how likely your evidence was on each of them. In many versions of the diffuse undermining case, rather mundane hypotheses (such as that the bottle was mislabeled) will receive the lion's share of your confidence, and radical ones (such as that you are having a psychotic episode) will receive precious little. As a result such cases will prompt disillusionment that is diffuse but not that deep: the cases will not make it reasonable for you to reconsider strongly held and central beliefs about what your evidence supports.

In the BB case, it remains to be seen how well your evidence is explained by rather mundane fishy hypotheses, and so remains to be seen whether that evidence prompts deep disillusionment. But in order to assess whether it does, we will need to motivate a more general Bayesian theory of undermining evidence than was employed in the analysis of the undermining drug case from §5.

7

To analyze the BB case, we will need to generalize the framework from §5. For that framework presupposed simple conditionalization: that a rational agent starts with a prior probability function $P(\cdot)$, then learns a new proposition E and comes to have a new probability function $P(\cdot \mid E)$. However, the BB case crucially involves potential mistrust in one's own memories—and simple conditionalization cannot adequately represent the development of such mistrust. This is shown (for example) by a variant of the Shangri-La example from Arntzenius (2003, 356):

Shangri-La There are two routes to Shangri-La: A and B. But right after entering the city, everyone who takes route A has their memory of taking that route replaced by a false memory of taking route B. Their

> memories of what song they sing during the journey, however, are unaffected. An acolyte who knows all this (but otherwise trusts her memory completely) tosses a fair coin to determine which route to take to Shangri-La. She takes route B. She tosses another fair coin to determine whether to sing about peace or about joy. She sings about peace. Right before entering Shangri-La, she is certain that she took route B. But after entering Shangri-La, how confident should she be that she took route B?

Answer: after entering Shangri-La, the acolyte should not trust her memory of having taken route B, and so should significantly reduce her confidence that she has taken route B. But we cannot understand the transition from her "right-before-entering-Shangri-La" credences and her "after-entering-Shangri-La" credences as simple conditionalization (Arntzenius 2003, 367). For (we may suppose) right before entering Shangri-La the acolyte had credence 1 that she took route B. And simple conditionalization can never reduce one's credence in a claim from 1 to a lower value.

Still, many think it clear that in the Shangri-La case, the acolyte should end up 50% confident that she has taken route B. The motivation for thinking this is that once she is in Shangri-La, the acolyte shouldn't trust her apparent memories of what route she took. So she should "bracket" them: set them aside. In other words, she should depend on a credence function that is "prior to" her current credence function in that it does not reflect trust in her apparent memory of which route she took.

One might be tempted to identify this prior credence function with the credence function she actually had at some earlier time—say, before tossing the coins and beginning her journey. But that won't work. Doing so would get the right result for her credence that she took route B, since that credence was 50% before she tossed the coins. But it would get the wrong result for her credence that she sings about peace, since that credence was then 50% as well. That result is wrong because now she should be much more than 50% confident that she sang about peace, since she has no reason

to doubt her memory of what she sang.

So the "prior" credence function that should constrain the acolyte's new credences is not the credence function she had at some previous time. Rather, it is the result of starting with her current credence function and bracketing off some aspect of it: trust in her apparent memories of what route she took.[8]

Appeal to this sort of prior credence function is helpful in a variant of the Shangri-La case as well. In the variant, the acolyte realizes that not everyone who enters by route A has their memory altered—just a randomly-chosen 30% of them. Again, the acolyte enters by route B and again we ask: once she enters Shangri-La, how confident she should be that she took route B? Here we cannot say that this credence should match her prior credence that she took route B. For since it could have happened that she took route A and did not undergo memory alteration, her apparent memories of having taken route B constitute some (though not decisive) evidence that she did so. This suggests a natural proposal: her credence that she took route B should equal: her prior credence that she took route B *conditional on* her having apparent memories as of having taken route B.

This recipe can be generalized: Suppose that one has evidence that threatens to compromise one's trust in the deliverances of a faculty. Then one's credences should depend on one's prior credence function—the credence function gotten from your current one by bracketing off or setting aside trust in that faculty. 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. Call this the *bracketing constraint*.[9]

Perhaps you have had the misfortune of speaking with someone who you initially trusted but later came to think was a liar. As you lost your

[8]Compare Elga (2007, 489).

[9]A special case of this constraint, applied to the case of peer disagreement, is the view given in Elga (2007, n. 26), a view which is similar in spirit to the one defended in Christensen (2007).

trust in this person, you started to incorporate what they said in a different way. You stopped adopting the content of their claims directly, but instead retreated to updating on the proposition that they *made* those claims. The bracketing constraint enjoins a similar retreat—except that the entity that you stop trusting (or come to trust less) is an aspect of yourself.

So: we have a constraint on how you take into account undermining evidence. This constraint helps recover all of the conclusions we have made about the apple drug case and the various versions of the undermining drug case and the Shangri-La case. But what does it entail about how we should react to the cosmological evidence in the light of BBs?

8

The short answer is: I have no idea, and maybe nothing.

The good news is that the bracketing approach does help avoid the sort of instability that the instability-blocks-undermining theorist was worried about. For it suggests that in the BB case one might adopt a credence function according to which one is stably confident that something fishy is going on, without requiring an unjustifiably-high level of confidence in any particular fishy hypotheses.

But bad news part 1 is that even though such a reaction wouldn't involve high confidence that you are a BB, it would involve high confidence that something extremely fishy is going on. Depending on the details, that might still involve giving up most ordinary scientific beliefs (just as in the diffuse undermining case one is not confident that the whole setup is as originally described). The fishy hypotheses might include, for example, that one is a BB, that one is a brain in a vat, that one is in a simulation, that contemporary science is entirely on the wrong track, that there is a vast conspiracy to produce misleading science, that one is suffering from severe hallucinations and false memories, and so on. And the conclusion that something *that* fishy is going on seems almost as crazy as the conclusion that one is a BB, even if it is part of a stable belief state.

Bad news part 2 is that the bracketing involved in the BB case is tremendously broad. For what your evidence threatens to compromise is nothing less than your entire ordinary scientific and commonsense worldview, including essentially all of your memories. Now it is one thing to ask what your credences are, setting aside your trust in one particular memory. But it is another thing to ask what they are, setting aside trust in almost all of your core beliefs.[10] *Perhaps* there is a fact of the matter about what this prior credence function is. But even if so, I find myself boggling at the prospect of pinning that function down with enough precision to deliver answers to the question: what are my prior credences, conditional on my current evidence? So it is not clear that the bracketing constraint delivers any particular verdict at all.

In this connection it is worth considering other batches of evidence that threaten to compromise one's core beliefs. For example:

> What would count as a reason to think [that you are a brain in a vat]? Perhaps experiences like these: You are walking down the street, hearing the ordinary soundtrack of life – birds chirping, the wind in the leaves, etc. Then all of a sudden there is a needle-on-vinyl-like glitch in the soundtrack, immediately followed by a strange, alien voice saying, "How is the experiment proceeding with the human brain in a vat?" This is followed by another strange, alien voice responding, "The experiment is proceeding very well… Hey, don't lean on that button. Get your tentacle off that button!" And this is abruptly followed by a return to the soundtrack of ordinary life. (Markosian 2014, 162)

Certainly in the light of that evidence you should become confident that something fishy is going on. Perhaps in the example as described, some not-too-wild hypotheses would get the lion's share of the confirmation:

[10] Compare Elga (2007, 496).

that you had a short seizure, that your friends played an elaborate practical joke on you involving loudspeakers (why were those "aliens" speaking English, anyway?), that you had an unusually vivid hallucination due to a medicinal side-effect, and so on. But one might modify the example to make those sorts of hypotheses harder to maintain (for example, by including experiences of periodic and systematic eavesdropped conversations from tentacled overseers — experiences that persist in the light of full medical workups). In the resulting case there is pressure to think you should become confident that something *extremely* fishy is going on. But even then it is not clear that it is reasonable to be very confident in any one particular fishy scenario. For once one takes any global deception scenario seriously, very many global deception scenarios may provide strong competing explanations for one's experiences (Elga 2003).

9

Where does that leave us on whether undermining considerations produce a reasonable response to the problem of Boltzmann Brains? We can endorse the zero-prior proposal, but that invites strange consequences in cases where we can influence whether large numbers of misled folks are created. The instability-blocks-undermining idea may block the conclusion that we should be confident that we are BBs, but does not on its own block the almost-as-unacceptable conclusion that we should be confident that something extremely fishy is going on.

I tentatively conclude that instability considerations do not provide a satisfying resolution to the problem of Boltzmann Brains.[11]

[11]For helpful conversations, thanks to Tyler Brooke-Wilson and Gideon Rosen. Thanks to Charlotte Elga for advising at a crucial sticking point: "Just sit down and write your paper, Dad."

References

Frank Arntzenius. Some problems for Conditionalization and Reflection. *Journal of Philosophy*, 100(7):356–371, 2003.

Frank Arntzenius and Cian Dorr. Self-Locating Priors and Cosmological Measures. In Khalil Chamcham, John Barrow, Simon Saunders, and Joe Silk, editors, *The Philosophy of Cosmology*, pages 396–428. Cambridge University Press, 2017.

Nick Bostrom. *Anthropic bias: observation selection effects in science and philosophy*. Studies in philosophy. Routledge, New York, 2002. ISBN 978-0-415-93858-7 978-0-415-88394-8.

Sean M. Carroll. Why Boltzmann brains are bad. In Shamik Dasgupta, Ravit Dotan, and Brad Weslake, editors, *Current Controversies in Philosophy of Science*, pages 7–20. Routledge, October 2020. ISBN 978-1-317-49715-8.

David John Chalmers. *Reality+: virtual worlds and the problems of philosophy*. W. W. Norton & Company, New York (N.Y.), 2022. ISBN 978-0-393-63580-5.

David Christensen. Epistemology of Disagreement: The Good News. *Philosophical Review*, 116(2):187–217, April 2007. ISSN 0031-8108, 1558-1470. doi: 10.1215/00318108-2006-035.

David Christensen. Epistemic Akrasia: No Apology Required. *Noûs*, 58: 54–76, 2024. ISSN 1468-0068. doi: 10.1111/nous.12441.

Sinan Dogramaci. Does my total evidence support that I'm a Boltzmann Brain? *Philosophical Studies*, 177(12):3717–3723, December 2020. ISSN 1573-0883. doi: 10.1007/s11098-019-01404-y.

Andy Egan and Adam Elga. I Can't Believe I'm Stupid. *Philosophical Perspectives*, 19(1):77–93, December 2005. ISSN 1520-8583. doi: 10.1111/j.1520-8583.2005.00054.x.

Adam Elga. Handout for MIT IAP Presentation: Why Neo was too confident that he had escaped the Matrix. January 2003.

Adam Elga. Reflection and Disagreement. *Noûs*, 41(3):478–502, 2007. ISSN 1468-0068. doi: 10.1111/j.1468-0068.2007.00656.x.

Matthew Kotzen. What Follows from the Possibility of Boltzmann Brains? In *Current Controversies in Philosophy of Science*. Routledge, 2020. ISBN 978-1-315-71315-1.

Maria Lasonen-Aarnio. Higher-Order Evidence and the Limits of Defeat. *Philosophy and Phenomenological Research*, 88(2):314–345, 2014. doi: 10.1111/phpr.12090.

Ned Markosian. Do You Know That You Are Not a Brain in a Vat?:. *Logos & Episteme*, 5(2):161–181, 2014. ISSN 2069-0533. doi: 10.5840/logos-episteme20145214.

Bradford Saad. Lessons From the Void: What Boltzmann Brains Teach. *Analytic Philosophy*, Forthcoming.

William J. Talbott. Is epistemic circularity a fallacy? *Philosophical Studies*, 177(8):2277–2298, August 2020. ISSN 0031-8116, 1573-0883. doi: 10.1007/s11098-019-01310-3.

Timothy Williamson. *Knowledge and its limits*. Oxford University Press, Oxford, 2000.

Timothy Williamson. Very Improbable Knowing. *Erkenntnis*, 79(5):971–999, October 2014. ISSN 1572-8420. doi: 10.1007/s10670-013-9590-9.

Git revision: ca244bd