Neal, Radford M. (2006). Puzzles of anthropic reasoning resolved using full non-indexical conditioning. *ArXiv Preprint Math/0608592*. [ArXiv preprint at arXiv:math/0608592 [math.ST]]

Nomura, Yasunori. (2015). A note on Boltzmann Brains. *Physics Letters B, 749*, 514–518. [arXiv:1502.05401 [hep-th]]

Olum, Ken D. (2002). The doomsday argument and the number of possible observers. *The Philosophical Quarterly, 52*(207), 164–184. [arXiv:gr-qc/0009081]

Page, Don N. (2006). The lifetime of the universe. *Journal of the Korean Physical Society, 49*, 711–714. [arXiv: hep-th/0510003]

Percival, Ian C. (1961). Almost periodicity and the quantal H theorem. *Journal of Mathematical Physics, 2*(2), 235–239.

Perlmutter, S., Aldering, G., Goldhaber, G., et al. (1999). Measurements of omega and lambda from 42 high-redshift supernovae. *The Astronomical Journal, 517*(2), 565–586. [arXiv: astro-ph/9812133]

Poincaré, Henri. (1890). Sur le problème des trois corps et les équations de la dynamique. *Acta Mathematica, 13*, 1–270.

Riess, Adam G., Filippenko, Alexei V., Challis, Peter, et al. (1998). Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal, 116*, 1009. [arXiv:astro-ph/9805201]

Schulman, Lawrence .. S. (1978). Note on the quantum recurrence theorem. *Physical Review A, 18*(5), 2379–2380.

Srednicki, Mark, & Hartle, James. (2010). Science in a very large universe. *Physical Review D, 81*(12), 123524. [arXiv:0906.0042 [hep-th]]

Vilenkin, Alexander. (1995). Predictions from quantum cosmology. *Physical Review Letters, 74*(6), 846. [arXiv:gr-qc/9406010]

Zermelo, Ernst. (1966a). Über einen Satz der Dynamik und die mechanische Wärmetheorie. In Stephen. G. Brush (Trans.), *Kinetic Theory* (p. 382). Oxford, UK: Oxford University Press. [Originally published in 1896, in *Annalen der Physik, 293*(3), 485–494]

Zermelo, Ernst. (1966b). Über mechanische Erklärungen irreversibler Vorgänge. Eine Antwort auf Hrn. Boltzmann's „Entgegnung." In Stephen G. Brush (Trans.), *Kinetic Theory* (p. 403). Oxford, UK: Oxford University Press. [Originally published in 1896, in *Annalen der Physik, 295*(12), 793–801]

2 What Follows from the Possibility of Boltzmann Brains?

Matthew Kotzen

2.1 The Boltzmann Brain Problems

A Boltzmann Brain is a hypothesized observer that comes into existence by way of an extremely low-probability quantum or thermodynamic[1] "fluctuation" and that is capable of conscious experience (including sensory experience and apparent memories) and at least some degree of reflection about itself and its environment. Boltzmann Brains do not have histories that are anything like the ones that we seriously consider as candidates for own history; they did not come into existence on a large, stable planet, and their existence is not the result of any sort of evolutionary process or intelligent design. Rather, they are staggeringly improbable cosmic "accidents" that are (at least typically) massively deluded about their own predicament and history. It is uncontroversial that Boltzmann Brains are both metaphysically and physically possible, and yet that they are staggeringly unlikely to fluctuate into existence at any particular moment.[2] Throughout the following, I will use the term "ordinary observer" to refer to an observer who is not a Boltzmann Brain. We naturally take ourselves to be ordinary observers, and I will not be arguing that we are in any way wrong to do so.

There are several deep and fascinating philosophical and cosmological questions that are raised by the possibility of Boltzmann Brains. Here are just a few of them: Do I have compelling reasons to believe that I am not a Boltzmann Brain? Is it possible, under any circumstances, for me to coherently believe that I am a Boltzmann Brain, or to suspect that I might be? If I am persuaded that most of the observers in the universe are indeed Boltzmann Brains, does that rationally compel me to believe that I am most likely a Boltzmann Brain? If I am persuaded that most of the observers in the universe *who are in my subjective state* are Boltzmann Brains, does that rationally compel me to believe that I am most likely a Boltzmann Brain? If a given cosmology entails that I am most likely a Boltzmann Brain, does this provide a strong reason to reject that cosmology? Does a cosmology according to which the universe has an infinite past or future (or both), or infinite space, entail that I am most likely a Boltzmann Brain – and, if so, should such a cosmology for this reason be rejected? Does a cosmology according to which the entire universe, or the portion of the universe in which we live, is the result of a quantum or thermodynamic fluctuation entail that I am most likely a Boltzmann Brain – and, if so, should such a cosmology for this reason be rejected? Are there

available and plausible and evidentially supported cosmologies according to which I am most likely *not* a Boltzmann Brain – and, if so, does this provide an additional reason to accept such cosmologies? Does the possibility of Boltzmann Brains present a genuine philosophical paradox, where several independently compelling hypotheses turn out to be logically inconsistent with each other, or does it simply present problems for particular cosmological assumptions and not others?

I cannot hope to resolve – or even to substantially address – all of these large and difficult questions here. However, in this chapter, I will try to do three things. First, in Section 2.2, I will argue that though David Albert and Sean Carroll's notion of "cognitive instability" is an interesting and important one, considerations of cognitive instability alone are not sufficient to rule out the hypothesis that I am a Boltzmann Brain. In Section 2.3, I will argue against James Hartle and Mark Srednicki's conclusion that we can coherently believe both that most observers in the universe are Boltzmann Brains and yet that we (likely) aren't. And in Section 2.4, I will briefly survey the nature of the problems that Boltzmann Brains pose for several categories of cosmological hypotheses.

2.2 Cognitive Instability

Carroll, building on some suggestions by Albert, has argued that I should reject the hypothesis that I am a Boltzmann Brain because there is an important sense in which belief in such a hypothesis is "cognitively unstable" and hence self-undermining. Carroll writes:

> Is it possible that you and your surrounding environment, including all of your purported knowledge of the past and the outside world, randomly fluctuated into existence out of a chaotic soup of particles? Sure, it's possible. But you should never attach very high credence to the possibility. Such a scenario is *cognitively unstable*, in the words of David Albert. You use your hard-won scientific knowledge to put together a picture of the world, and you realize that in that picture it is overwhelmingly likely that you have just randomly fluctuated into existence. But in that case, your hard-won scientific knowledge just randomly fluctuated into existence as well; you have no reason to actually think that it represents an accurate view of reality. It is impossible for a scenario like this to be true and at the same time for us to have good reasons to believe in it. The best response is to assign it a very low credence and move on with our lives.
>
> (2016, p. 92)

There is an ambiguity here: is Carroll's conclusion merely that we should *never* attach a *very high* credence to the possibility that we are Boltzmann Brains, or is it that we *should* attach a *very low* credence to the possibility? After all, there is significant credal distance between "very high" and "very low." Some version of the former conclusion may be plausible, but I am not persuaded of the latter conclusion. After all, there are plenty of other widely discussed "skeptical" hypotheses

that also, arguably, exhibit the phenomenon of cognitive instability, but I do not think that this alone provides adequate reason for us to reject (or to assign very low credence to) those hypotheses.

For instance, Descartes' famous "Dreaming Hypothesis" – the hypothesis that I'm currently experiencing a lifelike dream as opposed to interacting directly with the external world – may exhibit some degree of cognitive instability. If I were to believe that I am currently dreaming on the basis of my available evidence, there is a threat that the significance of that evidence is significantly undermined by the Dreaming Hypothesis itself; if I am dreaming right now, then (at least much of) my available evidence is dreamt evidence, and dreamt evidence (much like randomly fluctuated evidence) also fails to represent an accurate view of reality. Similar considerations apply to more contemporary skeptical hypothesis such as the hypothesis that all of my experiences are being caused by electrical stimulation via a supercomputer such as The Matrix; experiences caused by electrical stimulation via The Matrix similarly fail to represent an accurate view of reality. But surely this observation alone is not a general solution to Cartesian skepticism; even if I cannot reasonably assign very high credence to the Dreaming Hypothesis on the basis of evidence which would be unreliable if the Dreaming Hypothesis is true, it doesn't follow that I should assign a very low credence to the Dreaming Hypothesis. Rather, perhaps I just have no rational basis for determining whether I am dreaming or not, in which case one very natural response seems to be to assign some sort of middling credence – neither very high nor very low – to the Dreaming Hypothesis.[3]

Another example of cognitively unstable hypotheses is a class of "conspiracy theories" according to which a powerful agent or group has endeavored to mislead the public about their activities. If I come to believe that such a powerful agent is manipulating all of my evidence, including my evidence about the evidence-manipulating activities of that very agent, then it is hard to see how I could coherently hold an evidence-based belief in the truth of the conspiracy theory in question. But again, it is not obvious that the best response in each case is to simply assign a very low credence to the conspiracy theory in question; if the theory is independently plausible and does a good job of explaining the available evidence, then it may well deserve to be taken seriously and to be assigned a credence that is higher than "very low."

Though introduced for largely distinct dialectical purposes, Adam Elga's example of hypoxia (a condition in which the body and brain are deprived of adequate oxygen) also raises similar issues:

> Hypoxia impairs reasoning and judgment, which is bad enough. But what makes the condition really insidious is that it can be undetectable. In other words, when hypoxia impairs someone's reasoning, it can do so in such a way that the impaired reasoning seems (to the hypoxic individual) to be perfectly fine. It is a sad truth that airline pilots have crashed planes by happily and confidently making horrible judgment calls as a result of suffering from hypoxia.
>
> (Elga, 2008, p. 3)

Is it possible for the hypothesis that I'm currently hypoxic to be true and at the same time for me to have good reasons to believe it? The answer to this question may well depend on some of the details of the effects of hypoxia. Does hypoxia cause people to reason poorly in general, including when they are performing basic deductive and inductive inferences, or are the effects restricted to "higher-level" inferential skills such as those deployed in complicated mathematical calculations, strategic reasoning, and spatial reasoning? If hypoxia's effects are limited to higher-level forms of reasoning, then a hypoxic person might well double-check her mathematical calculations with a calculator, or measure the oxygen level in her own blood, and thereby come to the reasonable conclusion that she is hypoxic. But if hypoxia impacts basic reasoning, then it seems that almost *any* conclusion that a hypoxic person draws about her hypoxia – including the conclusion that she is hypoxic – is bound to be an unreasonable one. The conclusion in question might well seem like a reasonable one to the hypoxic person given her evidence, of course, but that's just a symptom of the hypoxia. If this is how hypoxia works, then it is hard to see how it could be possible for the hypothesis that a particular person is hypoxic to be true and at the same time for her to have a justified belief that it is true. However, the lesson here cannot be that the appropriate response is for pilots to simply disregard or discount the hypothesis that they are hypoxic. The "sad truth" that Elga refers to is that pilots suffering from hypoxia often assign *far too low* a credence to the hypothesis that they are hypoxic, and – notwithstanding the cognitive instability of the hypothesis that an individual is hypoxic – we certainly wouldn't want to instruct new pilots to adopt a policy of assigning a very low credence to the hypothesis that they are hypoxic and (hope that they) move on with their lives.

Some philosophers might try to distinguish examples like hypoxia from the case of Boltzmann Brains by arguing that whereas hypoxia makes it impossible for me to *take appropriate account of* the evidence I have that I'm hypoxic, I still *have* the evidence that I am hypoxic (say, in the form of my poor decision-making), right there in front of me; by contrast, since (as Carroll argues) fluctuated evidence doesn't actually represent an accurate view of reality, it is impossible for me to even *have* good reasons to believe that I am Boltzmann Brain.[4] In the philosophical jargon, another way to put this point might be that the truth of the hypoxia hypothesis merely makes it impossible for me to have a *doxastically justified belief* in that hypothesis, whereas the truth of the Boltzmann Brain hypothesis additionally makes it impossible for me to even have *propositional justification* to believe that hypothesis.[5] But, though I am convinced that this distinction between propositional and doxastic justification is an extremely important one in other contexts, I don't see any good reason to think that it is particularly useful here. First, it isn't completely clear what evidence I really *have* even in the hypoxia case (or in the hallucinatory drug case); when I'm hypoxic (or under the influence of the imagined hallucinatory drug), things seem perfectly normal to me, just as they do to Boltzmann Brains who are having experiences indistinguishable from those of ordinary observers.[6] Second, if I'm experiencing a particularly life-like dream, or if some conspiracy theory is true, I might not even *have* any evidence for those hypotheses, and yet I think that there are circumstances under which some such

hypotheses shouldn't be simply disregarded (or assigned extremely low credences). And third, it seems clear to me that it is possible *in principle* for me to have some evidence in favor of the hypothesis that I am a Boltzmann Brain; for example I could have the sorts of chaotic and disordered experiences that would presumably be quite typical of Boltzmann Brains and that are quite untypical of ordinary observers. I'll return to this point in Section 2.4 below.

None of this is to deny that the phenomenon of cognitive instability is an interesting or important one, or that it may often be epistemically relevant. But it seems to me that the cognitive instability of a particular hypothesis is not always a good reason to assign it a low credence. Rather, the cognitive instability of a hypothesis seems to be one way in which *a hypothesis, when true, is able to "hide itself" from rational discovery*.[7] If the hypotheses that I am dreaming or in The Matrix were true, those hypotheses would be very good at hiding themselves from rational discovery, as it would be difficult or impossible for me to acquire evidence that they are true. Similarly with the hypothesis that I am hypoxic: the perniciousness of hypoxia in the cases Elga refers to is precisely that, when a pilot is hypoxic, that fact often eludes rational discovery. And so too with conspiracy theories: by design, they are such that, when they are true, they are often difficult or impossible to get evidence for.

Different kinds and degrees of cognitive instability seem to correspond to varying senses in which, and extents to which, a hypothesis (when true) is able to elude discovery by making it (in varying ways and to varying extents) difficult or impossible for agents to form a reasonable belief that it is true. In some cases, this is because, when the hypothesis is true, the world is very much like (or very often like) the way the world is when the hypothesis is false; in such cases, it is very improbable for an event to occur to which even *could* count as a reason to believe or disbelieve the hypothesis. In other cases, the *entire* world might differ in important and widespread ways depending on whether the relevant hypothesis is true or false, and yet the world will tend to look the same *to observers* in either case. In still other cases, even though the truth of the hypothesis would impact the world in ways that are *in principle* detectable by observers, the truth of the hypothesis would also cause those very observers to fail to *notice* or *take appropriate account of* the ways in which the world is so impacted. These are all interesting – and interestingly different – cases of cognitive instability, but it strikes me as far too hasty to simply rule all of the relevant hypotheses out from serious consideration in one fell swoop.

Of course, many cognitively unstable hypotheses do indeed deserve low credences; indeed, many of them deserve *extraordinarily* low credences, and should be almost entirely disregarded in nearly every epistemic and practical context. Some conspiracy theories, for example, are so preposterously specific and ad hoc and reliant on coincidence that they cannot be taken seriously; even if they can be "cooked up" so that the likelihood that they assign to all of the available evidence is as high as one would like, their extraordinarily small prior probabilities guarantee that they should never be taken seriously in almost any practical or epistemic context other than the most extremely theoretical ones. In such cases, the cognitive instability of the relevant hypothesis is not *unrelated* to its low

prior probability; it is entirely possible for specificity or ad hocness or reliance on coincidence, for example, to explain *both* the cognitive instability of a hypothesis *and* its low prior probability. But cognitive instability and low prior probability are logically distinct – even if oftentimes correlated – properties of a hypothesis. Cognitive instability, all by itself, is not a sufficient reason to reject a hypothesis. Fortunately, there are plenty of *other* good reasons to reject the hypothesis that I am a Boltzmann Brain; I will return to this issue in Section 2.4.

Finally, it is not even completely obvious that considerations of cognitive instability always rule out assigning high credence to cognitively unstable hypotheses. Take the hypothesis that all of my current experiences are being caused by electrical stimulations via the Matrix – the thought here was that this hypothesis is cognitively unstable because, if it's true, then any evidence that I have in its favor is unreliable. But suppose that I were to have the experience of a digital scroll moving across my visual field which reads "You are in The Matrix, and we're trying to get you out!!" If the Matrix Hypothesis is true, then that visual-scroll-evidence is Matrix-generated and hence in some sense an unreliable representation of what the world is like (since there is not in fact any text floating in front of me, as my visual experience represents there to be). But, it would be quite unreasonable to dismiss the Matrix Hypothesis, or to refuse to assign it a high credence, on this basis; even if the visual-scroll evidence is an unreliable representation of the world around me, it is still an excellent indication that I am in The Matrix, since it is hard to imagine what other hypothesis could explain the scroll. Similarly, if I were to have the sort of disordered or chaotic experiences that are (presumably) quite typical of Boltzmann Brains and quite atypical of ordinary observers, those experiences could (at least in principle) constitute strong evidence that I am a Boltzmann Brain; as I'll suggest in Section 2.4, the fact that I do *not* have these disordered and chaotic experiences should be seen as some evidence that I am *not* a Boltzmann Brain. So a great deal depends on the exact nature and strength of the putative evidence in favor of the hypothesis that we are Boltzmann Brains; unfortunately, I do not see any *general* reason to insist that nothing even in principle could count as evidence in favor of that hypothesis.

2.3 Hartle and Srednicki

Another reaction to the problems posed by Boltzmann Brains, defended by Hartle and Srednicki, is that "it is perfectly possible (and not necessarily unlikely) for us to live in a universe in which we are not typical" (2007, p. 123523-1). One important dialectical consequence of this claim is that, if it is true, then it would allow us to rationally hold both that the vast majority of the observers in the universe are Boltzmann Brains, and yet that *we* are (most likely) not; we would of course be quite atypical observers in this scenario, but on Hartle and Srednicki's view that is not (necessarily) a mark against the rationality of believing in such a scenario. According to Hartle and Srednicki, "[a] theory is not incorrect merely because it predicts that we are atypical" (ibid.). Thus, even if some cosmology were to entail that most of the observers in the universe are Boltzmann Brains, it would not follow that I am probably a Boltzmann Brain, and hence my rejection of the

hypothesis that I am a Boltzmann Brain would not (necessarily) be a reason to reject that cosmology.

Hartle and Srednicki's argument relies to a significant extent on two analogies. First:

> Consider two theories of the development of planet-based intelligent life based on the appropriate physics, chemistry, biology, and ecology. Theory A predicts that there are likely to be intelligent beings living in the atmosphere of Jupiter; theory B predicts that there are no such beings. Because Jupiter is much larger than the Earth, theory A predicts that there are today many more jovians than humans.
>
> Would we reject theory A solely because humans would not then be typical of intelligent beings in our solar system? Would we use this theory to predict that there are no jovians, because that is the only way we could be typical? Such a conclusion seems absurd.
>
> (2007, p. 123523-2)

Hartle and Sredinicki are correct that a presumption in favor of our own typicality would (all by itself) be no reason to reject the hypothesis that there are (likely to be) intelligent beings living in the atmosphere of Jupiter. But, I believe, this analogy profoundly misunderstands the nature and dialectical role of assumptions regarding our own typicality. The point is not that, *regardless of what evidence we have*, we should always take ourselves to be typical of any specified class of observers. If I have excellent evidence that I have just won the lottery, I certainly shouldn't reject the hypothesis that I've won simply because being a lottery winner would make me a quite atypical human being. And even if I'm certain that I just won the lottery, that doesn't give me any reason to believe that most of the other lottery players have probably won too (as that would make me a more typical lottery player), or that most other people are lottery players (as that would make me a more typical human in one sense), or that everyone in the world is now rich (as that would make me a more typical human in another sense), or any other such thing.

Rather, the point of the most plausible versions of epistemic presumptions in favor of typicality is that, if there are two observers at a world *who have the same evidence*, then I can't have any good reason to believe that I am one of them rather than the other. For example, Adam Elga's self-locating "Indifference Principle"[8] entails that if there are two "subjectively indistinguishable" agents at a world – i.e., individuals who "have the same apparent memories and are undergoing experiences that feel just the same" (2004, p. 387) – then a rational agent should be equally confident in the hypothesis that she is identical to the first of these agents as she is that she is identical to the second of these agents. As a consequence (at least in the finite case), if there are multiple subjectively indistinguishable observers at a world, then a rational agent should be more confident that she is one of the (more numerous) typical ones, rather than one of the (less numerous) atypical ones. A natural generalization of this principle to the infinite case will entail that if, among a set of subjectively indistinguishable observers, there is a subset of typical

observers that has higher standard measure than the subset of atypical observers, then a rational agent should similarly be more confident that she is one of the typical observers. Indeed, a crucial component of several of the scientific and philosophical worries about Boltzmann Brains is precisely that (on certain cosmological assumptions) it is very likely that there are very many Boltzmann Brains *in precisely my current subjective state* in this world; if that is indeed the case, the thought goes, then my current subjective state would furnish me with no rational basis on which to conclude that I am an (atypical) ordinary observer, rather than a (typical) Boltzmann Brain.

However, Elga's Indifference Principle certainly does *not* entail that I should be more confident in hypotheses according to which there are fewer observers in the universe, or in hypotheses according to which there are fewer observers who are different from me in the universe.[9] If the humans and the jovians in the example would be in different subjective states, then Elga's Indifference Principle obviously doesn't apply. And even if they would be in the *same* subjective state, Elga's Indifference Principle still doesn't entail that we should have a low credence that there are jovians; all that it entails is that, on the assumption that there are jovians in the same subjective state as the humans, I should be more confident that I am one of the (more numerous) jovians than that I am one of the (less numerous) humans. So it seems to me that Hartle and Srednicki have an *extremely* different sort of typicality assumption in mind here from the ones that have currency in contemporary philosophical discussions.

Hartle and Srednicki's second analogy is as follows:

> Consider a model universe which has N cycles in time, $k = 1,\ldots,N$. In each cycle the universe may have one of two global properties: red (R) or blue (B)...Two competing theories of this model universe are proposed. One, *all red* or *AR*, in which all the cycles are red, and another, *some red* or *SR*, in which some number of particular cycles are red and the rest are blue. We (an idealized observing system) seek to discriminate between these two theories on the basis of our data....Suppose that we (a particular observing system) observe red. Our data D is then (E,R), which in the context of the model could be more fully described as 'there is at least one cycle in which an observing system exists and the universe is red.
>
> (2007, p. 123523–4)

If there are a low number of cycles (and if the fraction or measure of red cycles according to *SR* is significantly lower than 1), then $p(E,R \mid SR)$ can be significantly lower than $p(E,R \mid AR)$, in which case (E,R) significantly favors *AR* over *SR* (at least on Hartle and Srednicki's Bayesian assumptions, which I'm happy to stipulate). But Hartle and Srednicki point out that, as the number of red cycles increases according to each hypothesis, it becomes overwhelmingly probable that there will be *at least one* red cycle regardless of whether *AR* or *SR* is true, even if the fraction or measure of red cycles is fairly small according to *SR*. Thus, as the number of red cycles increases, $p(E,R \mid SR) \approx p(E,R \mid AR)$, and hence (E,R) no longer significantly favors *AR* over *SR*. Hartle and Srednicki conclude that, on these assumptions,

"[e]ven though the typical observing system in the *SR* theory is observing blue, our data provides no evidence that we are typical" (2007, p. 123523–5). Though this isn't made totally explicit, the analogical suggestion here is supposed to be that, as long as the conditions are such that the existence of *at least one* observer with my evidence is extremely likely, it is of no particular epistemological significance for a given cosmological theory that that theory entails that I am very atypical among all the observers who share my evidence. In particular, as long as a given cosmological theory entails that it is very likely that some ordinary observer with my evidence will exist at some point in the history of the universe, it is no mark against that theory if it *also* entails that the vast majority of observers in the universe with my evidence are Boltzmann Brains. If *SR* is true, most observers observe blue; but as long as it's very likely that someone will observe red, my observing red puts no rational pressure on me to reject *SR*. By analogy, if a particular cosmological theory is true, most observers are Boltzmann Brains; but as long as it's very likely that there will be an ordinary observer, my confident belief that I'm an ordinary observer puts no rational pressure on me to reject the cosmological theory.

However, it seems to me that Hartle and Srednicki's argument again contains a critical error, which can be traced back to the assumption that the relevant data is the proposition "there is at least one cycle in which an observing system exists and the universe is red." This is a flagrant violation of the Principle of Total Evidence,[10] according to which we should always take account of the *strongest* statement of our evidence that we have available to us. And the *strongest* statement of our evidence isn't the claim that there is *at least one* cycle in which an observing system exists and the universe is red; it is that *this* observing system – i.e., me – exists and observes red. And *that* is much more likely if *AR* is true than if *SR* is true, and hence (contra Hartle and Srednicki) our evidence confirms *AR* over *SR*.[11]

For comparison, suppose that 100 conscious observers are about to be created, and that either (*AR*) all 100 of them are going to be put in red rooms, or that (*SR*) only one of them of them is going to be put in a red room and the remaining 99 are going to be put into blue rooms. I wake up in a red room and am told all of this. Can there be any serious doubt that my observing red is strong evidence for *AR* over *SR*? After all, if *AR* is true, then I will certainly observe red; and if *SR* is true, then I am rather unlikely to observe red. It is irrelevant that, on either theory, it is certain that *at least one observer* will observe red. That's a weaker statement of my evidence than what I know, and I would violate the Principle of Total Evidence by conditionalizing on it and nothing stronger.

2.4 The Badness of Boltzmann Brains

Why *should* I reject the hypothesis that I am a Boltzmann Brain, and what implications does this have for cosmology? There are two good (if fairly obvious) reasons to reject the hypothesis that I am a Boltzmann Brain. First, the physical probability is staggeringly low that I would fluctuate into existence. And second (to return to a point from Section 2.2), even if I did fluctuate into existence, it is overwhelmingly improbable that I would be having anything like the ordered

and coherent stream of thoughts and experiences that I am in fact having; the vast majority of conscious Boltzmann Brains would have wildly disordered and incoherent thoughts and experiences. The relevance of these points to different cosmologies should be analyzed separately.

Sometimes, cosmological consequences related to Boltzmann Brains are appealed to in order to argue that the universe must not be infinite in time or space or both.[12] If the universe is infinite, it looks to follow that there will be infinitely many Boltzmann Brains. Suppose, in addition, that we are considering a cosmology according to which the set of all Boltzmann Brains vastly outnumbers (or: has a vastly higher asymptotic density[13] than) the set of ordinary observers. And suppose even further we are considering a cosmology according to which the set of all Boltzmann Brains *with ordered and coherent streams of thoughts and experiences* (or perhaps even stronger: with *my* total evidence) vastly outnumbers (or has a vastly higher asymptotic density than) the set of all ordinary observers with ordered and coherent streams of thoughts and experiences.

If I were certain (or nearly certain) that such a cosmology were true, then it seems unavoidable (for reasons related to the discussion of typicality in Section 2.3) that I should be overwhelmingly confident that I am one of the Boltzmann Brains with ordered and coherent streams of thoughts and experiences, rather than an ordinary observer. But, notwithstanding all of that, the fact remains that, if such a cosmology were true, it would be overwhelmingly probable (again for reasons related to the typicality considerations from Section 2.3) that I would have had the *disordered* and *incoherent* streams of thoughts and experiences that the vast majority of Boltzmann Brains (and hence the vast majority of observers in the universe) have. And since my *actual* thoughts and experiences are so highly ordered and coherent, I take myself to have strong evidence against a cosmology like that – evidence that doesn't depend on the *premise* that I'm not a Boltzmann Brain. The point here is not that there couldn't be a Boltzmann Brain that has ordered and coherent thoughts and experiences; indeed, if the universe were infinite and constantly fluctuating, there would (with probability 1) be *infinitely many* such Boltzmann Brains. And the point is not that it is more improbable for a particular Boltzmann Brain to have ordered experiences than it is for that Boltzmann Brain to come into existence to begin with. The point, rather, is that on the assumption that *I* am a Boltzmann Brain, it is incredibly unlikely that *I* would have such ordered and coherent thoughts and experiences, regardless of how likely it is that I would come into existence as a Boltzmann Brain to begin with. And that gives us reason to prefer a cosmology on which it's more probable that a randomly selected observer would have ordered thoughts and experiences.

Considerations having to do with Boltzmann Brains are *also* sometimes taken to cause problems for cosmologies according to which the state of our entire universe, or the portion of the universe in which we live, is the result of a quantum or thermodynamic fluctuation.[14] As I understand this concern, it is independent of whether the entire universe is infinite or not; whereas the prior concern was specifically about cosmologies according to which the universe is infinite, this set of worries applies even to cosmologies according to which the universe is finite.

The idea here is that, though the fluctuation of a Boltzmann Brain is indeed wildly improbable, the fluctuation of an *entire universe* or even of the *portion* of the universe we're able to observe into a low-entropy state is even *more* wildly improbable, and by a large degree, since low-entropy states of universes (or large portions thereof) require larger fluctuations in order to arise than Boltzmann Brains do. Thus, it is overwhelmingly more probable for a Boltzmann Brain with my current experiences to fluctuate into existence than for an entire universe (or a large portion thereof) to fluctuate into a low-entropy state, and hence (if I accept the cosmology under consideration) I should be more confident in the former hypothesis than in the latter one.

This reasoning strikes me as overwhelmingly compelling. Moreover, I am completely persuaded by Albert's (2000) and Carroll's (2010, chapters 8–9) arguments that we should accept the "Past Hypothesis" that the observable universe began in a state of very low entropy. Though a full discussion of the Past Hypothesis is not possible here, one of the central virtues of the Past Hypothesis is that it explains why, e.g., a photograph (or memory) is very likely to have been caused by the actual events that it represents; even though the most likely way *in the space of all possible evolutions of the universe* for the photograph to have come into existence is for it to have randomly fluctuated from a higher-entropy past, it is also the case that the most likely way *in the space of all evolutions of the universe from a low-entropy beginning* for this photograph to have come into existence is for it to have been caused by the event it represents. Similarly for me: even though the most likely way in the space of all possible evolutions of the universe for me to have come into existence is to have randomly fluctuated from a higher-entropy past, it is also the case that the most likely way in the space of all evolutions of the universe from a low-entropy beginning for me to have come into existence is through a process characteristic of ordinary observers. So, reasons to accept a cosmology that includes the Past Hypothesis (of which I think there are powerful ones) are also reasons to reject the hypothesis that I am a Boltzmann Brain.

The question remains, of course, of why the Past Hypothesis is true. And I think that one of the lessons here is that it will not do to say that the Past Hypothesis was itself true as a result of random fluctuation; if *that* were the only way that the Past Hypothesis could have been true, then I should prefer the hypothesis that I am a Boltzmann Brain on the grounds that this latter hypothesis is so much more probable. But it seems to me that the Inflation Theory[15] – according to which there was a period of exponential expansion of the universe during its first few moments – offers some realistic hope of explaining the truth of the Past Hypothesis without invoking the sorts of minuscule probabilities that would be associated with the Past Hypothesis being true as a result of random fluctuation. In brief, the idea here is that, prior to Inflation, the portion of the universe that was to become the observable universe was microscopic, and that quantum (and perhaps thermal) fluctuations on this microscopic scale expanded during Inflation to regions of low entropy that would make the Past Hypothesis true. The Inflation Theory has had many successes,[16] and I think that there are grounds for a great deal of optimism about both its truth and its capacity to explain our manifest experience without appealing to any wildy improbable fluctuations.

Notes

1 Some cosmologists have argued that Boltzmann Brains that arise as quantum fluctuations in the vacuum pose more serious problems than those that arise as thermal fluctuations – see, e.g., Davenport and Olum (2010). In this paper, I will ignore any differences that exist between these different sorts of Boltzmann Brains.

2 For a non-standard view of quantum fluctuations in de Sitter space, see Boddy, Carroll, and Pollack (2017); they argue that quantum fluctuations in isolated quantum systems are an epistemic phenomenon rather than a genuine physical one, and hence that Boltzmann Brains won't appear in the true de Sitter vacuum. However, as far as I know, there is absolutely no reason to doubt the physical possibility of Boltzmann Brains that arise by way of thermal fluctuation out of a thermal equilibrium.

3 Or, if such a thing is possible, to assign *no particular credence at all* – the credal equivalent of "withholding judgment" – to the hypothesis.

4 Thanks to Ram Neta for helpful discussion of this proposal.

5 See Ichikawa and Steup (2014, Section 1.3.2).

6 See Neta (2008) for a discussion what it means for an agent to "have" some particular piece of evidence.

7 I do not claim that cognitive instability *just is* the ability of a hypothesis to hide itself from rational discovery, nor do I claim that the *only* way for a hypothesis to hide itself from rational discovery is to be cognitively unstable. A hypothesis might be able to hide itself from rational discovery, for instance, simply by having a low prior probability and by making all of the same empirical predictions as a hypothesis with a high prior probability; no cognitive instability would be required. I claim only that cognitive instability is one way for a true hypothesis to evade rational discovery. This ability may well make it unreasonable for observers to assign the hypothesis a very high credence, but I am arguing here that it does *not* automatically make it reasonable for observers to assign the hypothesis a very low credence.

8 See Elga (2000, 2004).

9 It is controversial whether some version of the "Self-Indication Assumption" formulated in Bostrom (2002) is true: "Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist." (p. 66) I think that the most plausible version of the SIA adds a clause about sharing your evidence: "Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers who have your evidence exist over hypotheses on which few observers exist who have your evidence." If this latter principle is true, and if the humans and jovians in Hartle and Srednicki's analogy share the same evidence, then there is a case to be made that our existence provides some reason to believe that there *are* jovians.

10 See Carnap (1947).

11 For arguments along similar lines (though applied in somewhat different contexts), see Kotzen (2013) and White (2000).

12 See, e.g., Page (2007, 2008a, 2008b, 2008c). For useful discussion, see also Dyson, Kleban, and Susskind (2002); Bousso and Freigovel (2007); Linde (2007); Vilenkin (2007); and Banks (2007).

13 On the assumption that the set of observers in the universe is countably infinite, we cannot appeal to the standard Lebesgue measure here, since the Lebesgue measure can be definted only in spaces that can be represented as Euclidean n-dimensional *real-valued* spaces, and if the set of observers in the universe is countable, then the space of possible numbers of observers is natural-number-valued. Thus, we must appeal to asymptotic densities, which yield the "proportion" of natural numbers up to n that have some property, in the limit as n approaches ∞. See Nathanson (2000) and Tenenbaum (1995) for discussions of asymptotic densities. See Buck (1946) for a discussion of the analogy between measures and asymptotic densities.

14 See, e.g., Albrecht and Sorbo (2004), and Carroll (2016, Chapter 11).

15 The Inflation Theory was developed by Alan Guth, Andrei Linde, Paul Steinhardt, and Andreas Albrecht.

16 For instance, the Inflation Theory is often thought to explain the nearly-flat geometry of the universe, the uniformity of the cosmic background radiation, and the absence of stable magnetic monopoles. For an accessible overview, see NASA (n.d.).

References

Albert, David Z. (2003). *Time and chance*. Cambridge, MA: Harvard University Press.

Albrecht, Andreas, & Sorbo, Lorenzo. (2004). Can the universe afford inflation? *Physical Review D, 70*(6). [arXiv:hep-th/0405270]

Banks, Tom. (2007). Entropy and initial conditions in cosmology. [ArXiv preprint at arXiv:hep-th/0701146]

Boddy, Kimberly K., Carroll, Sean M., & Pollack, Jason. (2017). Why Boltzmann brains don't fluctuate into existence from the De Sitter vacuum. In Khalil Chamcham, Joseph Silk, John D. Barrow, et al. (Eds.), *The philosophy of cosmology* (pp. 228–240). Cambridge, UK: Cambridge University Press. [arXiv:1505.02780 [hep-th]

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

Bostrom, Nick. (2003). Are we living in a computer simulation? *The Philosophical Quarterly, 53*(211), 243–255.

Bousso, Raphael, & Freivogel, Ben. (2007). A paradox in the global description of the multiverse. *Journal of High Energy Physics, 06*, 018. [arXiv:hep-th/0610132]

Buck, R. Creighton. (1946). The measure theoretic approach to density. *American Journal of Mathematics, 68*(4), 560–580.

Carnap, Rudolf. (1947). On the application of inductive logic. *Philosophy and Phenomenological Research, 8*(1), 133–148.

Carroll, Sean M. (2010). *From eternity to here: the quest for the ultimate theory of time*. New York, NY: Penguin.

Carroll, Sean M. (2016). *The big picture: on the origins of life, meaning, and the universe itself*. Boston, MA: Dutton.

Davenport, Matthew, & Olum, Ken D. (2010). Are there Boltzmann brains in the vacuum? [ArXiv preprint at arXiv:1008.0808 [hep-th]]

Descartes, Rene. (n.d.). *Meditationes de prima philosophia, in qua Dei existentia et animae immortalitas demonstrantur [Meditations on first philosophy]*. Paris: Michel Soly.

Dyson, Lisa, Kleban, Matthew, & Susskind, Leonard. (2002). Disturbing implications of a cosmological constant. *Journal of High Energy Physics, 10*, 011. [arXiv:hep-th/0208013]

Elga, Adam. (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis, 60*(266), 143–147.

Elga, Adam. (2008). *Lucky to be rational*. Unpublished manuscript. Available at www.princeton.edu/~adame/papers/ bellingham-lucky.pdf

Elga, Adam. (2004). Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research, 69*(2), 383–396.

Hartle, James B., & Srednicki, Mark. (2007). Are we typical? *Physical Review D, 75*(12), 123523. [arXiv:0704.2630 [hep-th]]

Ichikawa, Jonathan J., & Steup, Matthias. (2014). The analysis of knowledge. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2014 Edition). Available at* http://plato.stanford.edu/archives/ spr2014/entries/knowledge-analysis/.

Kotzen, Matthew. (2013). Multiple studies and evidential defeat. *Noûs, 47*(1), 154–180.

Kotzen, Matthew. (2012). Selection biases in likelihood arguments. *The British Journal for the Philosophy of Science, 63*(4), 825–839.

Linde, Andrei D. (2007). Sinks in the landscape and the invasion of Boltzmann Brains. *Journal of Cosmology and Astroparticle Physics, 01*, 022. [arXiv:hep-th/0611043]

Nathanson, Melvyn B. (2000). *Elementary methods in number theory*. New York: Springer-Verlag.

National Aeronautics and Space Administration [NASA]. (n.d.). *Our universe: what is the inflation theory?* Available at http://wmap.gsfc.nasa.gov/universe/bb_cosmo_infl.html

Neta, Ram. (2008). What evidence do you have? *The British Journal for the Philosophy of Science, 59*(1), 89–119.

Page, Don N. (2007). Susskind's challenge to the Hartle–Hawking no-boundary proposal and possible resolutions. *Journal of Cosmology and Astroparticle Physics, 01*, 004. [arXiv:hep-th/0610199]

Page, Don N. (2008a). Is our universe decaying at an astronomical rate? *Physics Letters B, 669*(3–4), 197–200. [arXiv:hep-th/0612137]

Page, Don N. (2008b). Is our universe likely to decay within 20 billion years? *Physical Review D, 78*(6), 063535. [arXiv:hep-th/0610079]

Page, Don N. (2008c). Return of the Boltzmann brains. *Physical Review D, 78*(6), 063536. [arXiv:hep-th/0611158]

Tenenbaum, Gérald. (1995). *Introduction to analytic and probabilistic number theory. Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.

Vilenkin, Alexander. (2007). Freak observers and the measure of the multiverse. *Journal of High Energy Physics, 2007*(01), 092. [arXiv:hep-th/0611271]

White, Roger. (2000). Fine-tuning and multiple universes. *Noûs, 34*(2), 260–276.

Study Questions for Part I

1. What is a Boltzmann brain? In what are they similar and in what are they different from ordinary observers?
2. What is the standard argument against cosmologies dominated by Boltzmann brains? Where does this argument go wrong according to Carroll?
3. What is "cognitive instability"? Why does it pose a problem for cosmologies dominated by Boltzmann brains according to Carroll? Why, according to Kotzen, is it not a sufficient reason to reject a hypothesis?
4. Hartle and Srednicki argue that "[a] theory is not incorrect merely because it predicts that we are atypical." Explain one of the examples they use to support this claim and how it is supposed to bear on the Boltzmann brains problem. Does Kotzen think the analogy works? Why?
5. How can issues related to Boltzmann brains be used to argue against cosmologies in which time, space, or both are infinite?