DOI: 10.1111/nous.12439

ARTICLE

NOÛS

# The Boltzmann Brains Puzzle

**Ron Avni**

The University of Texas at Austin

**Correspondence**
Ron Avni, The University of Texas at Austin
Email: ron.avni@gmail.com

**Abstract**

Leading cosmological theories engender a controversial puzzle which has prompted philosophers to propose competing epistemological solutions and physicists to propose methodological changes to cosmology. The puzzle arises from the prediction that every brain on Earth will eventually be vastly outnumbered by physical duplicates formed by random collisions of particles in outer space. Supposing that this prediction is correct, shouldn't you believe that your brain is probably one of these vastly more typical extraterrestrial brains, since you cannot infer your brain's origin from your experiential state? But supposing that your brain is one of these extraterrestrial brains, why be confident in the cosmological theories, since you never actually received testimony supporting these theories? Proposals in the literature either deny the rationality of believing theories that make the prediction or deny the typicality of your brain among its duplicates. This paper argues that these proposals are not entirely satisfactory. Instead, one should be confident in theories making the prediction on the supposition that your brain is one of the extraterrestrial brains. The upshots include that it may be rational to believe that your brain is probably an extraterrestrial brain and that cosmologists should not alter their methodology in response to the puzzle.

## 1 INTRODUCTION

Leading cosmological theories engender a controversial puzzle. To solve the puzzle, philosophers have proposed competing epistemological solutions, while physicists have proposed methodological changes to how cosmology is conducted. There is no consensus on how to solve the puzzle.

The puzzle arises as follows. Some leading cosmological theories, which are well-supported by empirical and theoretical considerations, happen to make a strange prediction. The strange prediction is as follows. When the universe becomes sufficiently old, it will be an almost uniform distribution of lone elementary particles spread out across mostly empty space. In this quiescent state, almost nothing will happen. But every once in a while, some elementary particles will by random chance combine to form a larger aggregate of elementary particles. When such a larger aggregate does happen to form, it will thereafter disintegrate.

The story takes a strange turn once we focus on the fact that this quiescent state will last for an extremely long time. Just as a monkey's random typing will eventually produce *Hamlet*, so too will rare collisions between the mostly quiescent elementary particles eventually produce larger aggregates of particles in just about every possible arrangement. Most of these aggregates will be a simple cluster of particles, such as a pocket of hydrogen gas. But just as *Hamlet* will eventually appear in the monkey's text, so too will (for example) a toaster eventually appear in an otherwise mostly empty patch of outer space. The toaster will have been formed by random chance through collisions of particles and will then disintegrate. In fact, the quiescent period will last so very long that vastly many patches of space over time will each contain a toaster. And there will be so very many toasters that for *each* toaster that appears, *that* toaster will have a vast number of physical duplicates scattered across various patches of space at various times. Of course, toasters are but one example of objects that will be formed by random collisions of particles. *Every* physical object on Earth, including socks and toothbrushes, will eventually have vast numbers of physical duplicates floating in outer space.

Importantly, such physical objects include human brains. For every human brain here on Earth, there will eventually be a vast number of physical duplicates floating in outer space. These are the so-called "Boltzmann Brains": brain-like physical objects that were formed as a result of random collisions of elementary particles. Boltzmann Brains (hereafter, "BBs") are named after Ludwig Boltzmann, a physicist who contributed significantly to the development of some of the relevant theories of physics, namely, statistical mechanics and thermodynamics.

We can now sketch the puzzle generated by the strange prediction. To begin, note that you seem to have strong evidence that cosmological theories that make the strange prediction are correct. After all, such theories are among physicists' best theories, and physics has arguably been extremely successful at producing knowledge. But if these theories are correct, then should you be confident that you are a human being on Earth in the twenty-first century rather than a BB floating in outer space at some much later date? It seems that the subjective experience of being in the former situation would be indistinguishable from the subjective experience of being in the latter situation, at least for a short time. Therefore, you seem to have no reason to privilege the hypothesis that you are in the former situation. To the contrary, since these cosmological theories predict that for any brain on Earth, there will be a vast number of duplicate BBs floating in outer space, it seems vastly more likely that anyone having the experience that you yourself are having is a BB rather than a human on Earth. Thus, it seems that you should believe that you are vastly more likely to be a BB than an "ordinary observer" (hereafter, "OO") on Earth.

But there seems to be a problem with believing that you are probably a BB. Supposing that you are a BB, should you remain confident that a cosmological theory which predicts a preponderance of BBs is correct? Your confidence in the theory is presumably based on apparent memories as of receiving testimony from credible physicists, philosophers, or other experts. But if you are a BB, then you never really received any such testimony. Instead, if you are a BB, your brain was randomly formed in outer space with such apparent memories already in place. (As is familiar, your having apparent memories as of events that occurred more than one minute ago is arguably consistent with the claim that the entire universe was formed less than one minute ago.) Thus, it appears that accepting the cosmological theory, which in the first instance seems reasonable, leads to the conclusion that you have insufficient grounds for accepting the theory.

This completes the sketch of how the "Boltzmann Brains Puzzle" (hereafter, "BBP") arises. The rest of this essay proceeds as follows. §2 sharpens BBP by formulating it as a set of three seemingly plausible but jointly inconsistent premises: SCIENCE, TYPICALITY, and UNRELIABLE. Proposals in the literature address BBP by in effect criticizing either SCIENCE or TYPICALITY. §3 provides a defense of SCIENCE and considers one criticism from the literature. §4 provides a defense of TYPICALITY and considers two criticisms from the literature. It is argued in §§3-4 that none of the criticisms of SCIENCE or TYPICALITY are entirely satisfactory.

Then, §5 provides a defense of UNRELIABLE, and it is argued that this defense, while perhaps seemingly plausible, is actually fallacious. The rejection of UNRELIABLE provides a new solution to the Boltzmann Brains Puzzle. §6 concludes by noting some upshots of this new solution.

## 2 | SHARPENING THE BOLTZMANN BRAINS PUZZLE

This section sharpens BBP by formulating it as a set of three seemingly plausible but jointly inconsistent premises. Before proceeding, let us mention two attempts at solutions to BBP only to set them aside. First, some might claim that unlike OOs, BBs are not conscious, and since you are conscious, you are not a BB.[1] However, one might prefer to avoid a commitment to the debatable assumption that BBs are not conscious. Second, an epistemic externalist might deny that you should believe that you are vastly more likely to be a BB than an OO on the mere supposition that BBs vastly outnumber OOs. Instead, what you should believe is determined in part by whether you are a BB or an OO, even if you cannot discern whether you are a BB or an OO. For example, some writers have claimed in effect that you can strongly confirm that you are an OO simply by waiting a few seconds and observing that your brain has not begun disintegrating, which would cause a disordered phenomenology.[2] (This claim is most charitably interpreted as an externalist claim, since, plausibly, you cannot discern whether you are an OO who waited a few seconds or a BB with the apparent memory of having waited a few seconds.) However, one might prefer to avoid a commitment to externalism and its attendant challenges, which are well-known and will not be rehearsed here.

To simplify the premises of BBP, let us make three simplifying assumptions. First, let us assume that there are exactly two possibilities about the relative frequency of BBs and OOs throughout spacetime:

> BB+: BBs vastly outnumber OOs.
> OO+: OOs vastly outnumber BBs.

[1] Cf. Davidson's (1987) "Swampman".

[2] See, for example, Srednicki and Hartle (2013).

Second, let us assume that the agents considered in the premises of BBP have apparent memories as of receiving strong scientific evidence that BBs vastly outnumber OOs. (Whether or not they really did receive such evidence is another matter.) We make this assumption because BBP only arises for agents in such or similar experiential states. Let us also assume that the agents are certain that they have such apparent memories. In so doing, we set aside challenges to the assumption that the experiential states of having such apparent memories can be luminous.[3]

Third, let us assume that every agent is either a BB or an OO. Thus, for any given agent, one of the following is true:

$I_{BB}$: I am a BB.
$I_{OO}$: I am an OO.

To recap: (i) either BB+ or OO+ is true, (ii) every agent has apparent memories as of receiving strong scientific evidence that BB+ is true, and (iii) for any given agent, either $I_{BB}$ or $I_{OO}$ is true.

BBP is fundamentally a puzzle about what it is rational for an agent with a human brain to believe given their evidence. In particular, it is important to consider the *strength* of an agent's beliefs in relation to each other and in relation to the relative frequency of BBs and OOs. This paper therefore considers degrees of belief in a proposition, also known as credences, ranging from zero (perfect certainty in the negation of a proposition) to one (perfect certainty in a proposition). Let us use the shorthand "cr(P)" to refer to the credence that an agent has in a proposition P, and "cr(P|Q)" to refer to the conditional credence that an agent has in a proposition P on the supposition that proposition Q is true. In addition, let us use the symbols "$>>$" to mean "is significantly greater than", the symbol "$<<$" to mean "is significantly less than", the symbol "$\approx$" to mean "is approximately equal to", and the symbol "$\Box$" to mean "it is rationally required that".

With these terms in place, let us turn to briefly introducing the three premises of BBP, with more detailed investigation of each premise to follow in subsequent sections. First, given that an agent has apparent memories as of receiving strong *scientific evidence* that BB+ is true, then if the agent supposes that they are an OO, the agent should be confident that BB+ is indeed true. Thus:

SCIENCE: $\Box cr(BB+|I_{OO})\approx 1$

Second, if an agent supposes that BB+ is true, then the agent should consider themselves *typical* of agents who are in their experiential state, and thus extremely likely to be a BB. Likewise, if an agent supposes that OO+ is true, then the agent should consider themselves typical of agents who are in their experiential state, and thus extremely likely to be an OO. Thus:

TYPICALITY: $\Box cr(I_{BB}|BB+)\approx cr(I_{OO}|OO+)\approx 1$

Third, if an agent supposes that they are a BB, then their apparent memories as of receiving strong scientific evidence that BB+ is true were produced by an *unreliable* process. Therefore, if an agent supposes that they are a BB, then they need not be confident that BB+ is true. Thus:

UNRELIABLE: $\sim\Box cr(BB+|I_{BB})\approx 1$

[3] Perhaps a proponent of Williamson's (2000) views would raise such a challenge.

Each premise seems plausible. But the three premises are jointly inconsistent, as follows. In general, $cr(P|Q)\approx 1$ iff $cr(P\&Q)>>cr(\sim P\&Q)$. Therefore, Science entails that $\Box cr(BB+\&I_{OO})>>cr(OO+\&I_{OO})$, and Typicality entails that $\Box cr(I_{BB}\&BB+)>>cr(I_{OO}\&BB+)$ and also that $\Box cr(I_{OO}\&OO+)>>cr(I_{BB}\&OO+)$. By transitivity of >>, we therefore have that $\Box cr(BB+\&I_{BB})>>cr(OO+\&I_{BB})$, and thus $\Box cr(BB+|I_{BB})\approx 1$. But according to Unreliable, $\sim\Box cr(BB+|I_{BB})\approx 1$, and so the three premises jointly entail a contradiction.

It follows that at least one of the three premises of BBP must be rejected. But which ones, and why?

## 3 | PREMISE SCIENCE

The first premise of BBP is:

Science: $\Box cr(BB+|I_{OO})\approx 1$

A simple but plausible argument for Science is as follows. Consider any agent with apparent memories as of receiving strong scientific evidence that BB+ is true. On the supposition that the agent is an OO, the agent should believe that they really did receive strong scientific evidence that BB+ is true. One should confidently believe any proposition for which one has strong scientific evidence. Thus, a rational agent with apparent memories as of receiving strong scientific evidence that BB+ is true should confidently believe that BB+ is true on the supposition that they are an OO.

### Carroll's Objection: "Cognitively Unstable" Theories Deserve Zero or Low Credence

Carroll (2017) accepts that a puzzle arises from the potential existence of BBs. Carroll does not present arguments that could be used against Typicality or Unreliable. Instead, Carroll's solution to the puzzle is to put the blame squarely on the cosmological theory itself. According to Carroll, crediting the term to David Z. Albert, some theories are "cognitively unstable":

> "[The theory that you live in a] randomly-fluctuating universe… is therefore self-undermining, or as Albert has characterized similar situations in statistical mechanics, *cognitively unstable*… If you reason yourself into believing that you live in such a universe, you have to conclude that you have no justification for accepting your own reasoning. You cannot simultaneously conclude that you live in a randomly-fluctuating universe and believe that you have good reason for concluding that." (p. 22)

Carroll's view seems to be in effect that to accept a theory that entails BB+ is to accept a "cognitively unstable" theory that results in an epistemically untenable position. Carroll proposes the following solution:

> "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)**." (Ibid., bold emphasis added)

Carroll's proposal denies Science, as follows. If rationality requires that an agent's prior credence in BB+ be *exactly* zero, then no evidence whatsoever could prompt an agent to update their credence above zero. A fortiori, the agent's conditional credence in BB+ on the supposition that they are an OO must be exactly zero. This contradicts Science, which requires that an agent who has apparent memories as of receiving strong scientific evidence that BB+ is true has a conditional credence in BB+ of approximately one on the supposition that they are an OO.

Inside the parentheses, Carroll allows in effect for a rational agent's prior credence in BB+ to be "very close" to zero but not exactly zero. However, Carroll's proposal still denies Science even with this caveat. Either it is possible for an agent to receive sufficiently strong scientific evidence to justify updating their credence in BB+ to a high number, or not. If there is such evidence, then Carroll's proposal fails to solve BBP because we may simply stipulate that the agents we are considering have apparent memories as of receiving just such sufficiently strong scientific evidence. If there is no such evidence, then Carroll's proposal denies Science.

What follows from Carroll's denial of Science? Presumably, Carroll would accept that the fact that an agent's apparent memories are as of receiving strong scientific evidence that BB+ is true, together with the supposition that the agent is an OO, jointly entail that the agent really did receive strong scientific evidence that BB+ is true. Moreover, presumably, Carroll would accept that in general, having strong scientific evidence justifies high credence.

It therefore seems that Carroll must take BB+ to be exceptional among scientific hypotheses in that no scientific evidence, however strong, can rationally raise one's credence in BB+ to a high level. In so doing, Carroll in effect proposes to radically revise scientific methodology.[4] This is because there are existing scientific standards by which hypotheses such as BB+ are evaluated, and such standards do not rule out the possibility of strongly confirming BB+.

Moreover, such radical revision to scientific methodology faces significant challenges concerning the epistemic standing of BB+ relative to other scientific hypotheses. Many hypotheses adjacent to BB+, such as the hypothesis that the universe will expand indefinitely, are such that (i) the hypotheses themselves are *not* "cognitively unstable," and (ii) confirmation of the hypotheses would, according to existing scientific standards, support BB+. Is it plausible that arbitrarily many such adjacent hypotheses could be strongly confirmed while rationality requires that credence in BB+ remains zero (or very close to zero)?

Perhaps Carroll is prepared to bite this bullet. Although rejecting Science requires radical revision to scientific methodology that faces significant challenges, perhaps Carroll takes this cost to be the lowest priced solution of BBP. Put differently, perhaps Carroll's intended argument is in effect simply to accept Typicality and Unreliable and take BBP as a reductio ad absurdum of Science. However, §5 proposes a way to reject a premise of BBP without incurring such a hefty cost.

[4] Carroll is not alone in proposing in effect to solve BBP in this way: other physicists, such as Page (2008) and De Simone et. al. (2010), have in effect proposed that cosmologists avoid developing cosmological theories that entail BB+ just because the theories entail BB+.

## 4 | PREMISE TYPICALITY

The second premise of BBP is:

Typicality: $\Box\mathrm{cr}(\mathrm{I}_{\mathrm{BB}}|\mathrm{BB}+)\approx\mathrm{cr}(\mathrm{I}_{\mathrm{OO}}|\mathrm{OO}+)\approx 1$

A plausible argument by analogy for Typicality is as follows. Consider some rational agent $R_1$. Suppose that an unusual deck of cards consisting only of queens and kings is shuffled, and one card is dealt face-down to $R_1$. $R_1$ cannot infer whether the dealt card is a queen based on their experiential state (of seeing a face-down card). Now let Q+ be the proposition that the queens in the deck vastly outnumber the kings in the deck. Prima facie, since $R_1$ is rational, it follows that $R_1$'s conditional credence supposing that Q+ is true that the card dealt to $R_1$ is a queen is approximately one.

Typicality is supported by analogy with the above case. Consider some rational agent $R_2$. Suppose that $R_2$ is either a BB or an OO. Just as $R_1$ cannot infer whether the dealt card is a queen based on $R_1$'s experiential state (of seeing a face-down card), so too $R_2$ cannot infer whether $R_2$ is a BB based on $R_2$'s experiential state (regardless of which experiential state it is). BB+ is the proposition that BBs vastly outnumber OOs, just as Q+ is the proposition that queens vastly outnumber kings. $R_1$ and $R_2$ are in analogous epistemic positions regarding whether $R_1$'s card is a queen and whether $R_2$ is a BB, respectively. Thus, just as $R_1$ should be confident that their card is a queen on the supposition that Q+ is true, so too $R_2$ should be confident that $R_2$ is a BB on the supposition that BB+ is true. In other words, $\mathrm{cr}(\mathrm{I}_{\mathrm{BB}}|\mathrm{BB}+)\approx 1$. By parity of reasoning, it also follows that $\mathrm{cr}(\mathrm{I}_{\mathrm{OO}}|\mathrm{OO}+)\approx 1$.

It's worth noting that Typicality is entailed by the plausible principle that given a possible world, one should divide one's credences as to who one is equally among agents in subjectively indistinguishable states. Such a principle is defended by, for example, Elga (2004, p. 387).

### Kotzen's Objection: The Past Hypothesis Suggests That You Are Not a BB

Kotzen (2021) in effect attempts to refute Typicality. To consider Kotzen's argument, let us review some additional background information on the relevant physics, as follows. The microphysical laws appear to be almost entirely time-symmetric: given any physical process that occurs from time $t_1$ to time $t_2$, it is physically possible that a physical process occurs from $t_2$ to $t_1$ which is physically identical to the first process but for being temporally reversed. For example, consider a cue ball that rolls along a pool table and hits a stationary eight ball, after which the cue ball comes to a stop and the eight ball starts rolling. Since this physical sequence of events is possible, it is also possible that the same events could occur in reverse order: there would be no violation of the microphysical laws if an eight ball were to roll along a pool table and hit a stationary cue ball, after which the eight ball comes to a stop and the cue ball starts rolling.

However, it seems that many events that we regularly observe are never observed in what we would consider to be "reverse" order. A hot cup of tea always cools if left on a table, but a tepid cup of tea never spontaneously becomes hot. A glass vase dropped on the floor shatters into pieces, but pieces of glass on the floor never spontaneously gather themselves up into a solid vase. And so on, despite the fact that these "reverse" processes are just as consistent with the known microphysical laws. (A tiny caveat: there are some time asymmetries in the microphysical laws, but these are considered to be irrelevant in the present context.) Why are such "reverse" processes never observed?

This question may be reformulated as: why do we generally observe entropy, which may be glossed as a measure of disorder in a system, as increasing? One answer, originally due to Boltzmann (1897/2003), is that the universe began in an extremely low entropy state. One might imagine the universe in its first moment as a highly structured vase, which has subsequently been shattering in slow motion for billions of years. This is the reason that entropy increases as time passes. (Or, perhaps what we take to be the direction of time passing *just is* the direction in which entropy increases.) The hypothesis that the universe began with extremely low entropy was dubbed the "Past Hypothesis" by Albert (2000).

Now let us turn to Kotzen's attempt in effect to refute TYPICALITY. Kotzen's argument is presented within a single paragraph. The first half of the text of the argument is as follows:

> "[O]ne 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." (p. 31)

There is a problem with this assessment of the ramifications of the Past Hypothesis, at least on one reading of the above text. The text seems in effect to envisage two sets of universes: PH universes in which the Past Hypothesis holds, and ∼PH universes in which the Past Hypothesis does not hold. Now imagine a photograph of a tree. Suppose that an object that is a physical duplicate of the imagined photograph can be found in a PH universe and also in a ∼PH universe. All things being equal, the object in the PH universe is *more* likely than the object in the ∼PH universe to have been produced by a human who stood in front of a tree and who pressed a button on a camera. This is because the ∼PH universe is in a quiescent state as described in §1 in which any complex (that is, in this context, low-entropy) structure is likely to have arisen by random chance, whereas it will take the PH universe many billions of years for its low-entropy origin to dissipate, during which time complex structures are more likely to exist. As such, the aforementioned object in the PH universe is *more* likely to have been produced by a human than the corresponding object in the ∼PH universe.

However, it does *not* follow that the aforementioned object in *either* universe is likely to have been produced by a human. Whether or not the aforementioned object in either universe is likely to have been produced by a human depends only on whether randomly formed photographs vastly outnumber human produced photographs in that universe. And whether randomly formed photographs vastly outnumber human produced photographs is a question that is independent of whether the Past Hypothesis holds. If the "strange prediction" described in §1 holds for both universes, then it is overwhelmingly likely that the objects in both universes arose from random fluctuations, even if the probability of an origin from random fluctuation is *lower* in the PH universe than in the ∼PH universe. In other words, the Past Hypothesis does provide a boost in the probability that the object in the PH universe was produced by a human, but the Past Hypothesis does not by itself entail that this boosted probability is significantly greater than zero.

The second half of the text of Kotzen's argument is as follows:

> "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." (Ibid.)

Kotzen's comparing himself to the aforementioned photograph is not problematic. But, based on the above reading of the first half of the argument, the intended conclusion does not follow. The Past Hypothesis renders it *more* likely that any given brain was formed in utero on planet Earth. But supposing that BB+ is true, it is still overwhelmingly more likely that any given brain is that of a BB rather than an OO. As such, Kotzen has not provided a reason to reject TYPICALITY.

## Dogramaci's Objection: Do Not Make Bad Statistical Inferences

Dogramaci (2019) in effect rejects TYPICALITY by claiming that if one is rational, then one should not assign a high credence to the proposition that one is a BB, even if one supposes that BB+ is true. Why not?

Let us suppose that BB+ is true. In this case, most brains are those of BBs. It seems natural to conclude that since we have brains, we are probably BBs. According to Dogramaci, this reasoning is an instance of "statistical inference" (p. 3718): since most Fs are Gs, then, given that an individual is F, it is rational to have a high credence that this individual is G. But, according to Dogramaci, while many instances of statistical inference are sound, the above instance of statistical inference relating to BBs is unsound. For when one has the *additional* evidence that a given individual is an H and that most FHs are *not* Gs, then the statistical inference is undermined. Dogramaci claims that we are in the position of having just such additional evidence to undermine the statistical inference that we are BBs. This is because our *total* evidence includes evidence that we are OOs, and OOs are not BBs. Dogramaci argues for this claim as follows:

> "I say that we are in just such an "FH" situation... How do we learn that most minds like mine are BBs? We learn this only via acquiring a huge batch of scientific evidence. That batch includes in it lots that says we're in ordinary human bodies, on ordinary earth, which has existed and circled the sun for billions of years, and so on and so on. We are FHs, who see (through scientific reasoning) that there are (going to be) lots of Fs in the universe that are Gs. But that is not evidence that *we* are a G. We are not a G, i.e. not a BB. That is, on our *total* evidence, it's not rational to make the statistical inference that we're BBs." (p. 3719)

Thus, according to Dogramaci, since our total evidence includes evidence that we are OOs, it would be irrational to make the statistical inference that we are BBs. Dogramaci points out that his view entails that his BB duplicates should believe falsely that they are OOs, even on the supposition that these BBs vastly outnumber him. He writes: "All of the zillions of BBs should think they are not BBs. I will be right, they will be wrong, and we'll all be rational." (p. 3722) Dogramaci goes on to concede: "What I've just said can still sound unsettling: if I'm recommending all these

BBs think something false, aren't I suggesting I'm in a scenario where my total evidence is almost always misleading? Isn't that bad?!" (Ibid.) But he then tries to defend biting this bullet: "[T]he way to respect the evidence we have (and that we share with the BBs), is to respect it as evidence that, among other things, we are on ordinary earth, in ordinary bodies, in a universe we share with zillions of Boltzmann Brains." (Ibid.)

Dogramaci's objection is intriguing. One might challenge Dogramaci's reasoning by offering a theory of statistical inference which denies Dogramaci's judgments about which statistical inferences are sound. Offering such a theory is beyond the scope of this paper. Nonetheless, a worry about the objection can be raised indirectly, as follows. Let us consider a sequence of four statistical inferences. The first two inferences seem to be sound, even by Dogramaci's own lights; each inference in the sequence seems to be sound if its predecessor is sound; and the fourth inference justifies TYPICALITY. The upshot is that the burden is shifted to a proponent of Dogramaci's view to provide a principled reason for why, against appearances, TYPICALITY cannot be justified by statistical inference.

To set the stage, let us suppose that scientists can "envat" any human by plugging their brain into a virtual reality simulation such that the human has the same experiences that they would have had were they not envatted. Let us call any envatted human a "brain in a vat" (hereafter, "BIV"). Also, for convenience, let $A$ be your current age.

Now let us consider the following proposition:

> BIV$_1$: There is an age $a>A$ such that scientists envat 90% of humans aged $a$.

In other words, sometime in the future, 90% of humans born at the same time as you will become BIVs. What should your credence be, supposing that BIV$_1$ is true, that you will be a BIV? Since 90% of those in your birth cohort will be BIVs, statistical inference seems to suggest that the answer is 0.9. Do you have any evidence that you will be embodied, which would undermine this statistical inference? It seems that the answer is no. It also seems that Dogramaci would agree with this assessment. Dogramaci accepts that it may be "reasonable" to use statistical inference even "in the course of a skeptical argument" (p. 3719). Dogramaci gives the following example. Suppose that there will be a blinded drug trial in which 90% of the subjects are given a drug that induces a hallucinatory experience as of seeing an apple on a table. Dogramaci claims that it is rational to conclude, supposing that one is knowingly participating in this drug trial and having an experience as of seeing an apple on a table, that one is probably hallucinating. Using the terminology in the present paper, Dogramaci would agree that $\Box$cr(I am hallucinating|I am participating in the drug trial and having an experience as of seeing an apple on a table) = 0.9. Now, it seems that Dogramaci's drug trial does not differ in any relevant respects from the scenario envisaged in BIV$_1$. The psychopharmacological details of the drug trial differ from the neurophysiological details of the brain envatting, but these differences are not relevant to the question of whether the statistical inferences are sound. Therefore, it seems Dogramaci would agree that a sound statistical inference justifies $\Box$cr(I will be a BIV|BIV$_1$) = 0.9.

Now let us consider another proposition:

> BIV$_2$: There is an age $a<A$ such that scientists envat 90% of humans aged $a$.

The only difference between BIV$_1$ and BIV$_2$ is one of timing: in BIV$_1$, you are currently younger than the age at which 90% of your birth cohort will become envatted, and in BIV$_2$, you are currently older than the age at which 90% of your birth cohort became envatted. What should your credence be, supposing that BIV$_2$ is true, that you are a BIV? Since 90% of those in your birth cohort are BIVs, statistical inference seems to suggest that the answer is 0.9. Do you have any evidence that you are embodied, which would undermine this statistical inference? It seems that the answer is no, as follows. As noted, in the case of BIV$_1$, you have no evidence that you are embodied. The mere difference in timing between BIV$_1$ and BIV$_2$ does not provide you with evidence that you are embodied. Thus, in the case of BIV$_2$, you still have no evidence that you are embodied. It also seems that Dogramaci would agree with this assessment. Dogramaci considers a modification of the drug trial case in which one has an experience as of seeing an apple and *then* learns that one had been secretly enrolled in the drug trial. In this case, according to Dogramaci, the statistical inference that you are probably hallucinating is not undermined:

> "I could have started out believing, even knowing, that there's an apple I see here, and then I later learned that I've been participating in a drug experiment as described above. Then, of course, I would *lose* possession of the evidence that there's an apple I see here. I can't resist the statistical inference that I'm probably hallucinating." (pp. 3719–3720)

Therefore, it seems Dogramaci would agree that a sound statistical inference justifies $\Box$cr(I am a BIV|BIV$_2$) = 0.9.

Now let us consider another proposition:

BIV$_3$: Scientists envat 90% of humans aged zero.

In other words, 90% of humans have always been BIVs. What should your credence be, supposing that BIV$_3$ is true, that you have always been a BIV? Since 90% of humans have always been BIVs, statistical inference seems to suggest that the answer is 0.9. Do you have any evidence that you are embodied, which would undermine this statistical inference? It seems that the answer is no, as follows. As noted, in the case of BIV$_2$, you have no evidence that you are embodied. But BIV$_3$ is merely a special case of BIV$_2$, namely, the case with $a = 0$. (It does not seem that lowering the stipulated age of envatment provides you with evidence that you are embodied.) Thus, in the case of BIV$_3$, you still have no evidence that you are embodied. Therefore, a sound statistical inference justifies $\Box$cr(I have always been a BIV|BIV$_3$) = 0.9.

The fourth and final proposition to consider is BB+, repeated here for ease of reference:

BB+: BBs vastly outnumber OOs.

There are two notable differences between BIV$_3$ and BB+. The first difference is that in BIV$_3$, the disembodied brains have always been BIVs, and in BB+, the disembodied brains have always been BBs. The second difference is that in BIV$_3$, disembodied brains outnumber embodied brains nine to one, and in BB+, disembodied brains vastly outnumber embodied brains. What should your credence be, supposing that BB+ is true, that you have always been a BB? Since the vast majority of brains have always been BBs, statistical inference seems to suggest that the answer is approximately one. Do you have any evidence that you are embodied, which would undermine this statistical inference? It seems that the answer is no, as follows. As noted, in the case of BIV$_3$, you have no evidence that you are embodied. Merely replacing the possibility that you have always been disembodied as a BIV with the possibility that you have always been disembodied as a BB does not provide you with evidence that you are embodied. Moreover, merely increasing the magnitude by which disembodied brains outnumber embodied brains does not provide you with evidence that you are embodied. Thus, in the case of BB+, you still have no evidence that you are embodied. Therefore, a sound statistical inference justifies $\Box \mathrm{cr}(I_{BB}|BB+)\approx 1$. (And by parity of reasoning, a sound statistical inference also justifies $\Box \mathrm{cr}(I_{OO}|OO+)\approx 1$.)

In summary, we considered a sequence of four statistical inferences. The first two inferences seem to be sound; each inference in the sequence seems to be sound if its predecessor is sound; and the fourth inference justifies Typicality. Therefore, it seems that Typicality can be justified on the basis of a sound statistical inference.

Does the foregoing refute Dogramaci's objection to Typicality? Perhaps not. There may well be a principled reason to reject one of the four statistical inferences. However, in the absence of such a principled reason, there is an unresolved worry as to the success of this objection to Typicality.

## 5 | PREMISE UNRELIABLE

The third premise of BBP is:

Unreliable: $\sim\Box \mathrm{cr}(BB+|I_{BB})\approx 1$

Let us consider an argument for Unreliable that may seem plausible. As argued below, this argument for Unreliable is fallacious, but insofar as the argument has any initial appeal, such appeal would help explain why Unreliable may seem plausible. The text of the fallacious argument is enclosed in a box to set it apart from the rest of the paper. The argument is as follows:

When is a high credence in BB+ rationally required? The answer is that the agent must *both* have apparent memories as of receiving strong scientific evidence that BB+ is true, and *also* suppose that they are an OO, as follows.

A high credence in BB+ is not required in the absence of any evidence. This is because a high a priori credence in BB+ would be required only if the theoretical virtues of theories that predict BB+ are significantly greater than the theoretical virtues of theories that predict OO+, but this is not the case. Thus, having apparent memories as of receiving strong scientific evidence that BB+ is true (or having apparent memories similar to such apparent memories) is necessary for requiring a high credence in BB+.

But such apparent memories are not sufficient. Only if the agent supposes that they are an OO would having such apparent memories entail that the agent really did receive strong scientific evidence that BB+ is true, thereby suggesting that BB+ is true. If instead the agent supposes that they are a BB, then their apparent memories as of receiving strong scientific evidence that BB+ is true were produced by an unreliable process: their brain was randomly formed with apparent memories as of receiving the evidence already in place.

Therefore, a high credence in BB+ is required only if an agent both has apparent memories as of receiving strong scientific evidence that BB+ is true, and also supposes that they are an OO. But, if the agent supposes that they are a BB, then ipso facto they suppose that they are not an OO. Therefore, if an agent supposes that they are a BB, then a high credence in BB+ is not rationally required, which is just what Unreliable says.

## Objection: Merely Supposing That You Are a BB May Rationally Imply That BBs Are Prevalent

The above argument is fallacious simply because it is invalid. The argument has two premises. The first premise is that a high a priori credence in BB+ is not required since the theoretical virtues of theories that predict BB+ are not significantly greater than the theoretical virtues of theories that predict OO+. The argument does not provide any justification for this premise. However, let us accept the premise for the sake of argument.

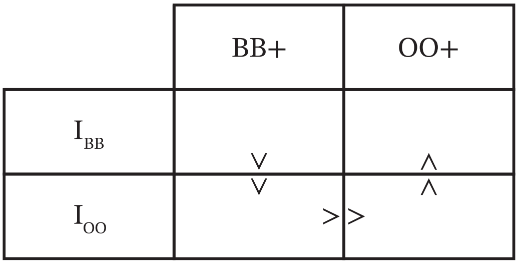
The second premise is that if an agent supposes that they are a BB, then their apparent memories as of receiving strong scientific evidence that BB+ is true were produced by an unreliable process. This claim seems entirely correct.

The problem with these two premises is that they do *not* jointly entail that for a high credence in BB+ to be required, it is *necessary* both to have apparent memories as of receiving strong scientific evidence that BB+ is true and also to suppose that one is an OO. These two conditions are arguably *sufficient* for requiring a high credence in BB+; this much is entailed by SCIENCE, which the argument for UNRELIABLE implicitly endorses. But additional premises would be needed to conclude that these two conditions are necessary for requiring a high credence in BB+. The argument for UNRELIABLE does not offer or defend any such additional premises.

We therefore conclude that the argument presented for UNRELIABLE should be rejected. This leaves two open questions: is there a better argument for UNRELIABLE? And is there a positive argument that UNRELIABLE is false?

The answer to the first question is: not to the knowledge of the present author. The fallacious but perhaps initially appealing argument is presented above in order to explain why UNRELIABLE may seem plausible. But it does not seem obvious that a better argument is available. It is left open for further research whether there is a persuasive argument in favor of UNRELIABLE.

In the meantime, there is a simple positive argument to reject UNRELIABLE, which is to accept SCIENCE and TYPICALITY on the merits of the arguments in their favor presented in §§3-4. As demonstrated by the reasoning at the end of §2, these two premises entail that UNRELIABLE is false. This reasoning can be visualized with a table of credences in order to further clarify the positive view proposed here:



In the table above, the horizontal ">>" symbol indicates that, as per SCIENCE, on the supposition that $I_{OO}$ obtains, credence in BB+ is significantly greater than credence in OO+; the vertical ">>" symbols indicate that, as per TYPICALITY, on the supposition that BB+ obtains, credence in $I_{BB}$ is significantly greater than credence in $I_{OO}$, and on the supposition that OO+ obtains, credence in $I_{OO}$ is significantly greater than credence in $I_{BB}$. The table thus reveals graphically that, on the supposition that $I_{BB}$ obtains, credence in BB+ is significantly greater than credence in OO+. That is, if SCIENCE and TYPICALITY are true, then *merely supposing that you are a BB* supports the hypothesis that BBs vastly outnumber OOs over the hypothesis that OOs vastly outnumber BBs. Contra the fallacious argument for Unreliable, the fact that supposing that you are a BB precludes supposing that you are an OO and thereby precludes using Science to conclude that BB+ is probably true, simply does not entail that supposing that you are a BB does *not* support the hypothesis that BB+ is true.

It's worth noting that there are alternate routes to denying Unreliable.[5] For example, Unreliable would presumably be denied by a proponent of the "Self-Indicating Assumption" (hereafter, "SIA"), a principle elucidated and criticized by Bostrom (2002). SIA is the principle that "…you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist." (p. 66) If you suppose that you are a BB, then more observers in your subjective state exist on the hypothesis that BB+ obtains than on the hypothesis that OO+ obtains. Thus, on this supposition, a proponent of SIA would presumably claim, contra Unreliable, that you should be confident that BB+ obtains. Note, however, that the argument against Unreliable proposed in this paper does not rely on SIA or a similar principle.

## 6 | CONCLUSION

As argued in §§3-4, there are plausible arguments in favor of Science and Typicality, and none of the criticisms of these two premises are entirely satisfactory. However, as argued in §5, we should not accept Unreliable. The BBP is thereby solved.

Should you believe that you are a BB? The solution of BBP proposed herein does not answer this question unconditionally. If you have apparent memories as of receiving strong scientific evidence that BB+ is true, then Science and Typicality jointly entail that you should indeed be confident that you are a BB. But if you do not have such apparent memories, then you need not be confident that you are a BB. At present, you have apparent memories as of receiving *some* scientific evidence that BB+ is true; whether the evidence you have apparent memories as of receiving is strong enough to justify a high credence in BB+ is beyond the scope of this paper.

Should physicists implement methodological changes to the science of cosmology in response to BBP? No. As argued in §3, rejecting Science requires radical revision to scientific methodology that faces significant challenges. Such revision should perhaps be accepted as a necessary cost if there is no other way to resolve BBP. But this cost can be avoided by rejecting Unreliable. Therefore, no methodological change to the science of cosmology is warranted.

## ACKNOWLEDGMENTS

For helpful comments and discussion, the author thanks Josh Dever, Sinan Dogramaci, Cory Juhl, Robert C. Koons, Miriam Schoenfield, and an anonymous referee.

## ORCID

*Ron Avni* https://orcid.org/0000-0002-1236-8430

## REFERENCES

Albert, D. Z. (2000). *Time and Chance.* Cambridge and London: Harvard University Press.

Boltzmann, L. (2003). On Zermelo's Paper "On the Mechanical Explanation of Irreversible Processes" ([S. G. Brush], Trans.). In S. G. Brush & N. S. Hall (Eds.), *The Kinetic Theory of Gases: An Anthology of Classic Papers with Historical Commentary* (pp. 412–419). London: Imperial College Press. (Original work published 1897)

[5] Thanks to an anonymous referee for drawing attention to this point in general and SIA in particular.

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

Carroll, S. M. (2017). Why Boltzmann Brains Are Bad. https://arxiv.org/abs/1702.00850

De Simone, A., Guth, A. H., Linde, A., Noorbala, M., Salem, M. P., & Vilenkin, A. (2010). Boltzmann brains and the scale-factor cutoff measure of the multiverse. *Physical Review D*, 82. https://doi.org/10.1103/PhysRevD.82.063520

Davidson, D. (1987). Knowing One's Own Mind. *Proceedings and Addresses of the American Philosophical Association*, *60*, 441–458. https://doi.org/10.2307/3131782

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

Elga, A. (2004). Defeating Dr. Evil with Self-Locating Belief. *Philosophy and Phenomenological Research*, *69*, 383–396. https://doi.org/10.1111/j.1933-1592.2004.tb00400.x

Kotzen, M. (2021). What Follows from the Possibility of Boltzmann Brains? In S. Dasgupta, R. Dotan, & B. Weslake (Eds.), *Current Controversies in Philosophy of Science* (pp. 21–34). New York and London: Routledge.

Page, D. N. (2008). Is our Universe likely to decay within 20 billion years? *Physical Review D*, 78. https://doi.org/10.1103/PhysRevD.78.063535

Srednicki, M., & Hartle, J. (2013). The Xerographic Distribution: Scientific Reasoning in a Large Universe. *Journal of Physics: Conference Series*, *462*, 12050–12054. https://doi.org/10.1088/1742-6596/462/1/012050

Williamson, T. (2000). *Knowledge and its Limits*. New York: Oxford University Press.

**How to cite this article:** Avni, R. (2022). The Boltzmann Brains Puzzle. *Noûs*, 1–15. https://doi.org/10.1111/nous.12439