A Bayesian analysis of self-undermining arguments in physics

David Wallace*

April 19, 2022

Abstract

Some theories in physics seem to be 'self-undermining': that is, if they are correct, we are probably mistaken about the evidence that apparently supports them. For instance, certain cosmological theories have the apparent consequence that most observers are so-called 'Boltzmann brains', which exist only momentarily and whose apparent experiences and memories are not veridical. I provide a Bayesian analysis to demonstrate why theories of this kind are not after all supported by the apparent evidence in their favor, taking advantage of the split between 'primary evidence', which directly supports a theory, and 'proximal evidence', which is our evidence (largely records and testimony) for the primary evidence.

Certain contexts in physics generate what David Albert (2000) has called *cognitive instability*, or what we might call a *self-undermining argument*. The abstract form of such an argument is: the evidence supports theory T, but conditional on theory T, it's extremely likely that I'm mistaken about the evidence.

The classic example is:

The Boltzmann Brain case: My evidence strongly supports (a certain version of) Statistical Mechanics. But according to (that version of) Statistical Mechanics, it is vastly more likely that my apparent evidence fluctuated into existence by pure change than that I actually made the observations I "remember" making. So Statistical Mechanics undermines its own evidence base.

See Carroll (2021), and references therein, for details of the salient physics. It is generally accepted that this sort of self-undermining, where the apparent 'evidence' for a theory is probably just chance, is a good reason to reject a theory. The point of this article is to show how to understand that from a Bayesian perspective.

There is a *prima facie* problem in doing so. In Bayesian epistemology, one starts with certain priors, and then updates those priors on the basis of evidence

*Department of Philosophy/Department of History and Philosophy of Science, University of Pittsburgh; david.wallace@pitt.edu

learned. If T is some theory, and E is some evidence, then upon learning E I update my credence in T from $\Pr(T)$ to $\Pr(T|E)$.

In the case of a self-undermining argument, it seems, the problem is that $Pr(E|T)$ is low: that is, conditional on the theory, the evidence is unlikely. But this doesn't do us much good, because *ex hypothesi* we have already learned E, so right now we give probability 1 to E. And that to which we give probability 1, we will always give probability 1.

For this reason (among others), Carroll, in his Bayesian analysis of this sort of cognitive instability, concludes that the sensible thing is just to choose priors that rule out self-undermining hypotheses:

> 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). ... we should discard such theories from consideration even before weve looked. (Carroll 2021)

This is perfectly defensible in itself, but we can do better.

Here's how. The "evidence" that supports contemporary statistical mechanics is a conjunction of (many thousands of) claims like:

- The Clausius and Kelvin statements of the Second Law of Thermodynamics have been tested lots of times and have passed the tests;
- The Smoluchowski model of Brownian motion has been tested lots of times and has passed the tests;
- The Boltzmann equation for gas diffusion has been tested lots of times and has passed the tests;
- Last time John Smith ordered a cocktail, the ice in his glass melted over the course of his drinking it;
- ...

More generally, the evidence for any scientific claim is going to be a bunch of experiments, performed over a long period and by many different people, and known to any given scientist only through testimony. Call this evidence the *primary evidence* for the claim.

I don't have direct, introspective, immediate access to the primary evidence. My own immediate data is vastly more impoverished: in the statistical-mechanics case, something like

- I (recall that I) read in a textbook that the Clausius and Kelvin statements of the Second Law, and the Smoluchowski model, and the Boltzmann equation, have all been very thoroughly tested and have passed the test;
- I recall that last night, the ice in my drink melted over the course of the evening;

- ...

Call this more impoverished evidence the *proximal* evidence for a scientific claim. The proximal evidence consists largely of records, testimony and memories that in normal circumstances are taken as being good reason to believe the primary evidence.

(There's a temptation to identify the proximal evidence with some *immediate* set of data: my current neural state, say, or my current sense data. But if we've learned anything from 20th century philosophy of science, it's that there is no really reliable division between immediate observational data and more theory-laden claims. So let's just recognise, more modestly, that my reason for accepting the extensive list of historical claims that comprise the primary evidence is mediated through my acceptance of the proximal evidence.)

For a given theory T, let E be the proposition that the proximal evidence for T obtains, and let H be the proposition that the proximal evidence for T obtains and that the primary evidence for T is approximately what the proximal evidence says that it is: in other words, H is the proposition that the proximal evidence is more or less reliable as to what the primary evidence is. ('More or less reliable' because of course the odd thing may be misrepresented or misreported.) Note that H entails E.

Now: suppose that T might be self-undermining. In this case, we need to recognize the risk that H may be false, i.e., that the proximal evidence may be unreliable as to the primary evidence. But conversely, the proximal evidence is only evidence for T insofar as it's evidence for the primary evidence — my having read a textbook account of a crucial experiment that supports T, for instance, is evidence for T only insofar as it's evidence that the crucial experiment actually occurred.

We can now formalize the claim that T is self-undermining as follows: it is the claim that $\Pr(H|T\&E)$ is some extremely low value ϵ (Caroll suggests $e^{-10^{122}}$). That is: given the theory, and given the proximal evidence, it's highly improbable that the primary evidence for the self-undermining theory actually happened. (Note that I am not here arguing that cosmologies with Boltzmann brains, or any other specific theories in physics, actually are self-undermining; I am taking the self-undermining nature of some T as a premise.)

Now: Bayes' Theorem tells us that

$$\Pr(\neg H|T\&E)\Pr(T|E) = \Pr(T|E\&\neg H)\Pr(\neg H|E). \tag{1}$$

But $\Pr(H|T\&E) = \epsilon$, so $\Pr(\neg H|T\&E) = (1-\epsilon) \simeq 1$, so

$$\Pr(T|E) = \frac{\Pr(T|E\&\neg H)\Pr(\neg H|E)}{1-\epsilon}. \tag{2}$$

$E\&\neg H$ is a sceptical scenario: to condition on it is to condition on the assumption that the proximal evidence, while apparently comprising records and testimony about the primary evidence, is actually wildly inaccurate about that evidence. Assuming that we give low priors to sceptical scenarios, $\Pr(\neg H|E) \ll 1$:

for concreteness, take $\Pr(\neg H|E) = \delta$ for some very small δ. Then

$$\Pr(T|E) < \frac{\delta}{1-\epsilon} \ll 1. \tag{3}$$

That is: if T is self-undermining, then there is no good reason to accept T given the proximal evidence E.

A corollary: since we normally assume that we can neglect sceptical scenarios, generally we feel free to update on the *primary* evidence (and not just the proximal evidence). If we do that, by Bayes' theorem again we have,

$$\Pr(T|H) = \Pr(T|H\&E) = \frac{\Pr(H|T\&E)\Pr(T|E)}{\Pr(H|E)} = \frac{\epsilon}{1-\delta}\Pr(T|E) \tag{4}$$

i.e.

$$\Pr(T|H) \ll \Pr(T|E) \ll 1. \tag{5}$$

So if T is self-undermining then conditionalizing on the *primary* evidence also gives a very low probability of T.

The argument here does not rely on any first-personal aspects of the proximal evidence, but only on the existence of a split between (a) the evidence that directly supports a theory, and (b) the evidence that in turn supports that evidence. For instance, consider the 'Boltzmann bubble' scenario where the Solar system fluctuated into existence long after the Big Bang and has been in existence for the last million years (so that our evidence about terrestrial and Solar-System matters over the last million years is largely correct but our apparent evidence about matters beyond the Solar system or more than one million years ago is just a random fluctuation. In that case we can take the proximal evidence to be the last million years of events on Earth (including the various apparent observations of stars, galaxies, the night sky, etc) that we have made throughout human history) and the primary evidence to be the actual astrophysical and cosmological claims (that the Universe is expanding, that it contains 10^{11} visible galaxies, that it is spatially flat or nearly so, etc) that is taken to directly support a cosmological theory. If that theory is self-undermining then the primary evidence is highly unlikely given the theory, even conditional on the proximal evidence, and again we get the result that conditionalizing either on the proximal, or on the primary, evidence suffices to make the self-undermining theory very unlikely.

References

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

Carroll, S. M. (2021). Why Boltzmann brains are bad. In S. Dasgupta, R. Dotan, and B. Weslake (Eds.), *Current Controversies in Philosophy of Science*, pp. 7–20. Routledge.