Aporia vol. 27 no. 2–2017

Safety vs. Simulation: A Skeptic's Argument

Bryan Ragan

I will show that Safety-based theories of knowledge which rely on nearby possible worlds fail to resolve the Brain-in-a-Vat (hereafter BiV) version of the Skeptic's Argument. These theories rely on the ideas that being a BiV is not a belief which is credible enough to be sincerely held, and that such a belief has a much lower probability of being true than false. BiV is a more modern evolution of Descartes' "Deceiver God," and advances in computer technology have made further evolution of this basic epistemic problem possible. By using the Simulation Hypothesis (hereafter Simulation) as put forward by Nick Bostrom, a new version of the Skeptic's Argument can be constructed. Simulation not only replaces the BiV framing of the Skeptic's Argument but goes further in arguing that this is actually more likely to be true than the alternative: that we live in a natural (not simulated) reality. The challenges this presents to epistemological theories are nearly identical to the BiV argument or Descartes' Deceiver God, but the nature of Simulation is especially problematic for any theory of knowledge based on Safety, such as those presented by Ernest Sosa and Duncan Pritchard.

The Skeptic's Argument

The Skeptic's Argument has persisted in the same general form since its conception. It began properly with Descartes in the 1600s when he wrote,

Bryan Ragan is a United States Marine Corps veteran who is attending North Carolina State University on the Post-9/11 GI Bill. He is in his senior year working towards dual majors in biology and philosophy. His philosophical interests include philosophy of mind and cognitive science.

> Because I have no reason for thinking that there is a God who is a deceiver (and of course I do not yet sufficiently know whether there even is a God), the basis for doubting depending as it does merely on the above hypothesis, is very tenuous and, so to speak, metaphysical. But in order to remove even this basis for doubt, I should at the first opportunity inquire whether there is a God, and, if there is, whether or not he can be a deceiver. For if I am ignorant of this, it appears I am never capable of being completely certain about anything else. (Descartes 48)

More formally,

> (P1) I do not know that God is not a deceiver.
>
> (P2) If I do not know that God is not a deceiver, then I do not know X.
>
> Therefore, (C) I do not know X.

Descartes attempts to resolve his predicament by establishing both that God exists and that it is not in God's nature to deceive. In short, Descartes attacks (P1) in order to evade (C). Unfortunately for Descartes, his skeptical argument in modern guise presents a chronic challenge, where his arguments to resolve it have not proven nearly as persistent.

A modern version as presented by Duncan Pritchard (and others), the Brain-in-a-Vat Hypothesis (BiV), can be presented as follows,

> (P1) I do not know that I am not a BiV.
>
> (P2) If I do not know that I am not a BiV, then I do not know X.
>
> (C) I do not know X. (Pritchard 441)

This version removes the theological aspects of Descartes' version while preserving its underlying core. Whether we are inclined to accept Descartes' proof that God is not a deceiver is no longer relevant because the framing has changed. Under this science fiction derived hypothesis, we need only accept the possibility that some human or alien mad scientist is capable of being a deceiver, and that, of course, is eminently reasonable to assume.

Sensitivity and Safety

Safety-based theories (hereafter Safety) are a refinement of Sensitivity-based theories (hereafter Sensitivity), and for that reason a brief

overview of both Sensitivity and the reasons for its refinement are in order. Sensitivity-based theories begin by stating that for anything to constitute knowledge, it must be sensitive. Per Pritchard, Sensitivity is defined as: "An agent S has a *sensitive* belief in a true contingent proposition *p* iff in the nearest possible worlds in which *p* is not true, *S* no longer believes *p*" (438). Pritchard further defines "nearest possible worlds" as those worlds "most similar to the actual world" (438). Sosa defines Sensitivity in slightly different terms, stating that "a belief by *S* that *p* is 'sensitive' iff were it not so that *p*, *S* would not believe that *p*" (141). The key component to Sensitivity is that beliefs must have some way of being sensitive to their own truth conditions before they can truly be a candidate for knowledge. For example, my belief that I am sitting down is sensitive because in nearby (similar) worlds where I am standing up I do not believe that I am sitting down.

Sensitivity can deal with the Skeptic's Argument, but it does so at great cost. Like Descartes, Sensitivity focuses on (P1) in an attempt to defeat the Skeptic's Argument. But, rather than demonstrate that (P1) is false, Sensitivity demonstrates that the belief that I am not a Brain-in-a-Vat is not sensitive. If I believe I am not a BiV, in possible worlds where I am a BiV, I would still be able to believe I am not a BiV (Pritchard 442). Therefore, the belief that I am not a BiV is not sensitive, so it is not a candidate for knowledge. Sensitivity does not so much defeat the Skeptic's Argument as make it epistemologically irrelevant. Sensitivity theorists therefore argue that even though (P1) is true, it just doesn't matter.

Where Sensitivity starts to run into issues is with (P2). The problem is that we have previously established that all sorts of everyday knowledge are sensitive. When we combine our ability to know common facts with our inability to know that we are not BiVs, some very awkward things follow. For instance, "Suppose that I know that I am seated, and I also know that if I am seated, then I am not a BiV. Then it follows via closure that I know that I am not a BiV" (Pritchard 442). While my belief that I am seated is sensitive and a candidate to be knowledge, believing that I am not a BiV is not sensitive and thus cannot be known. So, it seems that I know something—knowledge that entails something else—but I cannot know what it entails. Or as DeRose put it, "the abominable conjunction that while you don't know that you're not a bodiless (and handless!) BiV, still, you know you have hands" (21).

This is such a large problem because in order to defeat the Skeptic's Argument, we are forced to abandon the closure principle which its second premise relies on. Closure can be defined as follows: "For all *S*, P1, and P2, if *S* knows P1, and *S* knows that P1 entails P2, then *S* also knows P2" (Pritchard 441). As illustrated above, under Sensitivity I can

potentially know things which entail things that are impossible for me to know. Closure is so strongly intuitive that for most its loss is simply too steep a price to pay—even if doing so allows one to defeat the Skeptic's Argument (Pritchard 442). If resolving the Skeptic's Argument requires that we abandon something as useful and intuitive as the Closure principle, it certainly seems like we are on the wrong track.

Safety-based theories maintain Sensitivity's ability to resolve the Skeptic's Argument while simultaneously disposing of the "abominable conjunction" and retaining Closure. Per Pritchard, Safety can be defined thusly: "An agent S has a *safe* belief in a true contingent proposition *p* iff in most nearby possible worlds in which S believes *p*, *p* is true" (Pritchard 446). Sosa defines "safe" as follows:

> Call a belief by S that *p* "safe" if: S would believe that *p* only if it were so *p*. (Alternatively, a belief by S that *p* is "safe" iff: S would not believe that *p* without it being the case that *p*; or, better, iff: as a matter of fact, though perhaps not as a matter of strict necessity, not easily would S believe that *p* without it being the case that *p*). (Sosa 142)

Safety can be applied to the BiV problem through the following: I believe I am not a BiV, and in most nearby possible worlds where I believe I am not a BiV, it is true that I am not a BiV; therefore, my belief that I am not a BiV is "safe." With respect to everyday knowledge under Safety, if I believe I am sitting down, and in most nearby possible worlds where I believe that I am sitting down it is true that I am sitting down, my belief that I am sitting down is "safe." With respect to Closure, my belief that I am sitting is "safe" and thus a candidate for knowledge, and my belief that I am not a BiV is also "safe" and thus also a candidate for knowledge. The first entails the second, and both can be known; hence, Closure is preserved.

While we have resolved the abominable conjunction, how Safety resolves the BiV argument still requires some exploration. As Sosa puts it, "In the actual world, and for quite a distance away from the actual world, up to quite remote possible worlds, our belief that we are not radically deceived matches the fact as to whether we are or are not radically deceived" (147). The problem with this resides in the notion of distance as it relates to possible worlds. If I believe that I am not a BiV, and I am not a BiV, then the nearest possible worlds will also be worlds in which I am not a BiV. So, my belief that I am not a BiV will be "safe." If I believe that I am a BiV, and I am a BiV, then the nearest possible worlds will also be worlds in which I am a BiV. So, my belief that I am a BiV will be "safe." Both cases are perfectly parallel. Similarly, if I believe that I am a BiV, and I am not a

BiV, then the nearest possible worlds will be worlds where I am not a BiV and my belief will not be "safe." Likewise, if I believe I am not a BiV, and I am a BiV, it is not "safe." Whether or not I can know I'm a BiV is entirely dependent on whether I actually am a BiV. Safety only resolves the BiV argument because we start with the assumption that we are not BiVs. Safety itself does nothing to resolve the problem; the assumption that I am not a BiV does all the work. This assumption is entirely the result of the framing of the BiV argument being so unbelievable or seemingly improbable.

While the scientist overseeing the direct stimulation of his vast collection of brains would certainly be able to deceive them, why would anyone (or anything) do this? Furthermore, BiVs would by their very nature still be occupying space in the original reality from which they came, so it seems reasonable to assume that the majority of the brains that exist and have ever existed do not reside in vats, but in bodies. While the apparent improbability of a BiV might, on the surface, seem irrelevant for a thought experiment, it is this very incredulity which Safety relies on to resolve the BiV problem. Absent this incredulity, both cases (BiV and non-BiV worlds) are perfectly parallel with respect to the assessment of whether they are "safe." How much stronger then would the Skeptic's Argument be if it were possible to remove said incredulity entirely, or at least reduce it massively?

Simulation: A New Skeptical Hypothesis

The Simulation Hypothesis as put forward by Bostrom makes a compelling basis for a very believable version of the Skeptic's Argument. The hypothesis relies on several assumptions which are informed by science as much as, or more than, science fiction. (A1) First, the prediction from computer scientists that, per Bostrom, "enormous amounts of computing power will be available in the future," is true (243). (A2) Second, the assumption of substrate-independence, as put forth in philosophy of mind and cognitive science, is true. Substrate-independence being "the idea that mental states can supervene on any of the broad class of physical substrates. Provided a system implements the right sort of computational structures and processes, it can be associated with conscious experiences" (Bostrom 244). (A3) Finally, it is possible for a civilization like ours to achieve a posthuman level of technological development, and such a civilization would have an interest in devoting some of its vast computational power to running "ancestor simulations" either for scientific or even entertainment purposes (Bostrom 245).

(A1) is speculative to a certain degree, but on its surface appears well within the realm of possibility. Obviously, our current level of technological

advancement is inadequate for the production of ancestor simulations, but, for Bostrom, "Persuasive arguments have been given to the effect that *if* technological progress continues unabated *then* these shortcomings will eventually be overcome" (245). It is important to note that timescale is irrelevant. It does not matter if the necessary computational ability is obtained in fifty years, five thousand years, or even hundreds of thousands or millions of years from the present; the only thing that matters is that it is possible. Bostrom goes into great detail with regards to resolving this issue and ultimately concludes that "posthuman civilizations would have enough computing power to run hugely many ancestor-simulations even while using only a tiny fraction of their resources for that purpose" (248).

(A2) is not entirely uncontroversial, but only the weak version of substrate-independence is required for simulation to be possible (Bostrom 244). We only need to assume that, "a computer running a suitable program would be conscious," and that the computational processes of the human brain could be replicated by such a program to a suitably fine-grained degree if not in totality (Bostrom 244). Both assumptions are widely accepted within the relevant fields of study (Bostrom 244).

(A3), by its nature, is the most speculative assumption because it relies on predicting future states of civilization. It is to deal with (A3) that Bostrom puts forth his disjunctive argument for Simulation Hypothesis as follows:

> At least one of the following propositions is true:
>
> (1) The human species is very likely to go extinct before reaching a "posthuman" stage.
>
> (2) Any posthuman civilization is extremely unlikely to run a significant number of simulations of their evolutionary history (or variations thereof).
>
> (3) We are almost certainly living in a computer simulation.

Furthermore, it follows that the belief that there is a significant chance that we will one day become posthumans who run ancestor-simulations is false—unless we are currently living in a simulation (Bostrom 255).

For our purposes, the justification for (3) being true in the event (1) and (2) are false is of importance and necessitates reiteration. If a posthuman stage is achieved, and most posthumans have interest in running ancestor-simulations, and the computational abilities of such a posthuman civilization would be great enough to allow for many such simulations, it

follows that the majority of minds like ours in existence would not belong to the original race but to simulations (Bostrom 255).

With Simulation understood, the Skeptic's Argument can be reformulated with a new framing as follows:

> (P1) I do not know that I am not living in a simulation.
>
> (P2) If I do not know that I am not living in a simulation, then I do not know X.
>
> Therefore, (C) I do not know X.

Safety vs. Simulation

In different historical approaches to dealing with the Skeptic's Argument, a trend becomes apparent: to defeat the Skeptic's Argument by attacking some aspect of its framing. Descartes defeated his original form theologically because he had framed the argument in theological terms, namely a Deceiver God. However, there is a sense in which the nature of the deceiver is somewhat external to the argument itself—and therefore arbitrary. In resolving the BiV version of the Skeptic's Argument, Safety-based theories commit a similar error to Descartes; they defeat the argument by attacking the Deceiver God, or rather what stands in for it. BiV does away with any theological weaknesses which could be exploited by someone like Descartes, but it encounters other weaknesses in doing so.

The concept of a BiV is, put simply, ridiculous at face value. It does not seem true, and a person assumes a kind of default position that it is false. The BiV is a kind of construct designed specifically for the Skeptic's Argument—it doesn't flow naturally. Furthermore, baked into the nature of its construction is the almost certain proposition that even if BiVs exist, it would be a minority of brains across all possible worlds which would be BiVs. BiV has two weaknesses which are exploited by Safety: (1) An intuitive incredulity towards the concept, and (2) a low probability of any individual brain being a BiV. Even among those who find the BiV version of the Skeptic's Argument persuasive, you would be hard-pressed to find anyone who sincerely believed they were themselves a BiV.

Simulation, unlike BiV, developed independently from the Skeptic's Argument itself. For this reason, the use of Simulation as a new type of framing device for the Skeptic's Argument does not have the same intrinsic failings found in BiV versions. The Simulation argument is first and foremost an argument for Simulation being entailed by reasonable premises. Furthermore, a non-trivial number of people exposed to

the Simulation argument find it compelling. Even some very public and well-regarded individual like Elon Musk openly profess their honest belief in it (Koebler).

Serendipitously for our purposes, Simulation relies on similar reasoning to that which is used in Safety for its own justification; both Simulation and Safety make appeals to nearby possible worlds and rely on probability with respect to the nature of those nearby possible worlds. Simulation (1) is intuitively credible when a person is exposed to it and (2) maintains that the probability of any given brain being simulated is higher than the probability that it is natural. Under Safety, if I believe in this world that I am a simulation and it is true that I am a simulation, then in nearby possible worlds where I believe I am a simulation it will also be true that I am a simulation; thus, my belief that I am a simulation is "safe." As discussed above, the same is true for the parallel position. So, we are left with an unresolved problem: Is the reality we experience simulated or not? Am I being radically deceived? For Safety to provide us with any help we must know which possible worlds are nearby, and to know that we must already know the precise thing we are trying to determine—the very thing which Simulation calls into question in the first place. The circularity here is an obvious issue.

Conclusion

While Safety-based theories can defeat skeptical arguments (e.g., Brain-in-a-Vat) by attacking weaknesses built into the external framing of the BiV hypothesis itself, they do not address the core of the Skeptic's Argument. Given a version of the Skeptic's Argument built on the Simulation Hypothesis, as I have detailed, the belief that I am living in a simulation meets the requirements for being "safe" laid out by Duncan Pritchard and Ernest Sosa for Safety-based theories of knowledge. Admittedly, it could be argued that this belief is not Sensitive. But, if forced to fall back to Sensitivity, we are forced to abandon Closure—the very problem which Safety was developed to solve. Nevertheless, Safety cannot be maintained when presented with the Simulation-based version of the Skeptic's Argument. It may be possible to formulate an epistemological theory which exploits some weakness in Simulation itself, but this only serves to defeat a given description of the nature of the deceiver used by the argument. An epistemological theory which truly wishes to defeat the Skeptic's Argument must find a way to engage with its core. Any other approach, based on the historical trend outlined, will only lead to the

eventual development of new and different versions of Descartes' Deceiver God.

Works Cited

Bostrom, Nick. "Are You Living in a Computer Simulation?" *Philosophical Quarterly*, vol. 53, no. 211, Apr. 2003, pp. 243–55.

DeRose, Keith. *The Appearance of Ignorance: Knowledge, Skepticism, and Context*. Oxford UP, 2018, p 21.

Descartes, Rene. "Meditations." *Modern Philosophy: An Anthology of Primary Sources*, by Roger Ariew and Eric Watkins, Hackett Publishing, 2009, pp. 1–106.

Koebler, Jason. "Elon Musk Says There's a 'One in Billions' Chance Reality is Not a Simulation." *Vice*, Motherboard, 2 June 2016.

Pritchard, Duncan. "Sensitivity, Safety, and Anti-Luck Epistemology." *The Oxford Handbook of Skepticism*, by John Greco, Oxford UP, 2008, pp. 436–55.

Sosa, Ernest. "How to Defeat Opposition to Moore." *Philosophical Perspectives*, vol. 13, 1999, pp. 14–53.