

The Philosophical Quarterly

CONTENTS

ARTICLES

On the Composition of the <i>Prototractatus</i>	<i>Jinho Kang</i>	1
Confused Thought and Modes of Presentation	<i>Krista Lawlor</i>	21
The Structure of Sceptical Arguments	<i>Duncan Pritchard</i>	37
Deflationism and the Success Argument	<i>Nic Damjanovic</i>	53
Two Sides of 'Silencing'	<i>Jeffrey Seidman</i>	68

DISCUSSIONS

Logical Syntax in Wittgenstein's <i>Tractatus</i>	<i>Cora Diamond</i>	78
The Simulation Argument: Reply to Weatherson	<i>Nick Bostrom</i>	90
Punishment as Deterrence: Reply to Sprague	<i>Anthony Ellis</i>	98
Privileged Access Revisited	<i>Jordi Fernández</i>	102

CRITICAL STUDIES

Budd and Brady on the Aesthetics of Nature	<i>Allen Carlson</i>	106
Myles Burnyeat's <i>Map of Metaphysics Zeta</i>	<i>Mary Louise Gill</i>	114

BOOK REVIEWS

122

BLACKWELL PUBLISHING
FOR
THE SCOTS PHILOSOPHICAL CLUB
AND THE
UNIVERSITY OF ST ANDREWS

THE SIMULATION ARGUMENT: REPLY TO WEATHERSON

BY NICK BOSTROM

I reply to some recent comments by Brian Weatherston on my 'simulation argument'. I clarify some interpretational matters, and address issues relating to epistemological externalism, the difference from traditional brain-in-a-vat arguments, and a challenge based on 'grue'-like predicates.

1. Introduction

In an earlier volume of this journal, I introduced a line of reasoning which I call 'the simulation argument'.¹ Compressed into a nutshell, the reasoning ran as follows:

- *Metaphysical assumption.* I assumed a weak and quite widely accepted form of supervenience, according to which conscious experiences would in fact result from the running of computer programs emulating human brains at some suitable level of detail, such as that of neurons and synapses. I called this assumption 'substrate-independence'.
- *Empirical assumptions.* I gave (and to some limited extent justified) a rough estimate of the computational requirements of emulating a human mind on a computer, and of simulating such aspects of the environment as human-like minds would be able to detect. I also estimated a lower bound on the computing power available to a technologically mature civilization (defined as one which has developed those technologies that we already have good grounds for believing to be feasible). Even allowing for a large margin of error in these estimates, a technologically mature civilization would have so much computing power that even if it devoted only a tiny fraction of it to running 'ancestor-simulations', there would be vastly many more simulated ancestral people than non-simulated ones.

From this, I argued that we have grounds for accepting a tripartite disjunction, namely, that at least one of the following three propositions is true:

1. $f_p \approx 0$
2. $f_I \approx 0$
3. $f_{sim} \approx 1$.

¹ N. Bostrom, 'Are You Living in a Computer Simulation?', *The Philosophical Quarterly*, 53 (2003), pp. 243–55.

Here f_p is the fraction of civilizations at our current stage that eventually become technologically mature; f_t is the fraction of technologically mature civilizations that apply some non-negligible portion of their computational power to running ancestor-simulations; and f_{sim} is the fraction of all people with human-like experiences who live in simulations.

I further argued that if (3) is assumed to be true, we should think that we are almost certainly living in a computer simulation.

Brian Weatherson, in his recent commentary on my paper, is prepared to accept the correctness of the simulation argument up to but not including this last step.² He also (p. 425) accepts (3). By contrast, I do not accept (3), but only the disjunction $(1) \vee (2) \vee (3)$. My view is that we do not currently have strong evidence for or against any of the particular disjuncts. At any rate, the disjunction is all that the simulation argument purports to show; it does not seek to establish that we are probably living in a computer simulation.

For the purposes of this paper, however, I shall assume (3), so that I can focus on the particular claim with which Weatherson takes issue.

2. *Weatherson's four interpretations of the bland principle of indifference*

The question I am concerned with is therefore this: given (3), do we have reason to assign a high credence to the thesis (SIM) that we live in a simulation?

I noted in the original paper that it is not in general true that if we knew (3) we should then assign a high credence to (SIM). We might have some additional piece of specific relevant information. For example, if we were convinced that all simulated people would be dressed in black trench coats, then the fact that we are not all wearing these coats would be evidence against (SIM). Or suppose that we had somehow deduced that the simulators would make all people in simulations aware of their condition (perhaps by making a 'window' stating 'YOU ARE LIVING IN A COMPUTER SIMULATION' hover in their visual field). Then our failure to observe such a signal would give us grounds for concluding that we are not simulated. More generally, we might believe that people in ancestor-simulations would be either more or less likely than non-simulated people to have some particular kind of experience. Our having that kind of experience might then provide probabilistic evidence for or against (SIM). I suggested, however, that we do not in fact seem to have any such evidence, or at least none that is sufficiently strongly correlated with (SIM) to alter the conclusion that, given (3), we should assign a high probability to (SIM).

I suggested the following principle for assigning probability to (SIM) in the absence of more specific relevant information (I termed it the 'bland principle of indifference'):

BPI. $Cr(SIM \mid f_{sim} = x) = x$.

Weatherson does not attack this principle directly. Instead, he formulates a more general principle which he declares is 'the main reason for believing' (BPI):

BPI⁺. $\forall \Phi : Cr(\Phi \mid f_{\Phi} = x) = x$.

² B. Weatherson, 'Are You a Sim?', *The Philosophical Quarterly*, 53 (2003), pp. 425–31.

Here Φ stands for the statement ‘I have property ϕ ’. In the special case where ϕ is the property of being simulated, (BPI⁺) reduces to (BPI). (BPI⁺), more general and therefore *prima facie* more vulnerable, is the primary target of Weatherston’s attack.

He discusses four different interpretations of it, and argues that none of them lends support to the disputed final step in the simulation argument.

On the first interpretation, the scope of the quantifier in (BPI⁺) is unrestricted. Weatherston says (p. 426) that this results in a clearly false principle, because a rational agent ‘may well know that the proportion of human-like agents who like spaghetti westerns is rather low, while rationally being quite confident that he likes spaghetti westerns’. Yet this example is inconsistent with (BPI⁺) only if the principle is misconstrued as stipulating that our credence in Φ should equal x whenever we know that $f_\phi = x$. What the principle actually asserts, however, is that the conditional credence of Φ given $f_\phi = x$ should be x . If one has other relevant information, such as the fact in Weatherston’s example that one likes spaghetti westerns (where Φ is the statement ‘I have the property of liking spaghetti westerns’), then one needs to conditionalize on this extra evidence too, in order to arrive at the correct posterior credence assignment to Φ . This follows from the principle of total evidence.

For comparison, suppose someone proposes the principle ‘The conditional credence of randomly selecting a black ball from an urn in which the fraction of black balls equals x should be x ’. Does one provide an example to counter this (quite reasonable) urn-principle by pointing out that an agent may well already have looked at the ball he has drawn, and may be rationally quite confident that it is black? Only if the urn-principle is misinterpreted as claiming a lot more than it does.

Weatherston’s second and third interpretations restrict the quantifier in (BPI⁺) in various ways, and he argues that because of ‘grue’-like problems, these restrictions do not succeed in making it immune from the same sort of problem that he thinks affects the principle in its first interpretation. But since the first interpretation did not suffer from the alleged problem (unless the principle is mistakenly taken to imply that no other possible information could affect one’s posterior credence in Φ if one knows that $f_\phi = x$), I do not need to discuss these two interpretations further. (I shall say something about the ‘grue’ problem towards the end of the paper.)

In Weatherston’s fourth interpretation, he removes the red herring present in the other three interpretations by recognizing that (BPI) is really a constraint on prior credence functions. Here he also slides back to discussing (BPI) rather than (BPI⁺). It seems that he accepts (BPI) under its correct interpretation. At any rate, he makes no attempt to counter any of the arguments for accepting (BPI) that were given in the original paper.

Weatherston then reconstructs the final step of the simulation argument as follows (I have slightly adapted his notation to make it conform to the rest of this paper):

- P1. *A priori*, our rational credence in (SIM), given that $f_{sim} \approx x$, is x
- P2. All our evidence is probabilistically independent of the property of being in a simulation
- C. Our rational current credence in (SIM), given that $f_{sim} \approx x$, is x .

‘This interpretation’, he writes (p. 429), ‘may be reasonably faithful to what Bostrom

had in mind'. And indeed, this is getting closer to the argument I presented, although I still have a few quibbles. The first and most serious is that (P2) is incorrectly formulated. The whole point of the simulation argument was to argue that we have evidence that is interestingly relevant to the hypothesis that we are in a simulation! A better formulation of (P2) (which, I presume, is also what Weatherston intended) is

P2*: All our evidence is probabilistically independent of (SIM), after we conditionalize on $f_{sim} \approx x$ (for some particular x).

The second quibble is that I do not wish to claim that even (P2*) is strictly true. It may well be that we have *some* evidence that is not entirely independent of (SIM) conditionalized on $f_{sim} \approx x$. We can speculate that hypothetical simulators may be especially interested in simulating certain kinds of events or people and that they would run a disproportionate number of ancestor-simulations of such kinds. For example, one might think that pivotal moments in history, or unusually interesting lives, would be simulated relatively more frequently than they occur in non-simulated histories. Then, depending on whether we think that our own lives and places in history fit these presumed criteria (or not), we may have additional probabilistic evidence for (or against) (SIM) even after conditionalizing on $f_{sim} \approx x$.

My view is that such speculations are tenuous. We have very little understanding of the psychology or motivations of our hypothetical simulators (who would surely have to be an extremely advanced life-form in order to be technically capable of creating ancestor-simulations). It is possible that as we or our descendants develop towards mastering the requisite technological capabilities, more information might become available about the operative motives for running ancestor-simulations (or about the absence of such motives). This could give further probabilistic evidence about (SIM). Yet for now, it seems fair to say that whatever evidence we have either for or against (SIM), other than the evidence that exerts its influence via $f_{sim} \approx x$, is at best weak and speculative.

For the purposes of the simulation argument, all we need from (P2*) is that our current empirical evidence is too weak to result in substantial changes to the posterior credence of (SIM) after we have already conditionalized on $f_{sim} \approx 1$. The plausibility of this assumption is enhanced by the fact that the approximate equality $f_{sim} \approx 1$, which is the disjunct (3) in the summary of the simulation argument above, is indeed very close to exact equality if the empirical estimates referred to in my introduction are in the right ballpark. The credence of (SIM) obtained by conditionalizing on $f_{sim} \approx x$, where x has a value that is very close to 1, would remain fairly close to 1 unless we had strongly relevant specific evidence to the contrary, of the sort suggested above. We must therefore understand 'probabilistically independent' in (P2*) in a loose sense that accommodates these reservations.

A further minor quibble with Weatherston's reconstruction is that (P1) has a technical defect that is avoided by (BPI). Since f_{sim} is a constrained variable – it takes on values only within the unit interval – it is not true that, for any x , $Cr(SIM | f_{sim} \approx x)$ should be (exactly) x . Conditional on $f_{sim} \approx 1$, our rational credence in (SIM) is not 1 but slightly less than 1, because $f_{sim} \approx 1$ is compatible with $f_{sim} < 1$. Likewise, on the condition of $f_{sim} \approx 0$, our rational credence in (SIM) is a little greater than 0. (One

way to apply the correct principle $Cr(SM \mid f_{sim} = x) = x$, if our evidence is of the form $f_{sim} \approx 1$, is to transform the latter approximate proposition into some suitable set of more specific propositions with an associated measure. These propositions may be of the form ' $f_{sim} = 1 - x$ ', for $\epsilon \geq x \geq 0$, where ϵ is a small value corresponding to the degree of inexactitude of the approximation in $f_{sim} \approx 1$.)

3. *The issue of epistemological externalism*

The point of contention having now been clarified, Weatherson's objections boil down to two, both of which are directed against (P2*). The first objection (p. 429) is that our evidence might be constituted by more than our conscious phenomenal states. (I shall deal with the second objection in §5.)

[A rational person's] evidence might be constituted by more than his conscious phenomenal states.... On the externalist version, [his] perceptual evidence is constituted in part by the objects he is perceiving.... A Sim may not know that he has different evidence from someone seeing a dagger when he sim-sees a sim-dagger; but that does not imply that he does not have different evidence, unless one also assumes, implausibly, that agents know exactly what their evidence is.

Since this is not the place to enter into a full-throttle discussion about the nature of evidence, I shall confine myself to a couple of brief remarks. Epistemological externalism is sometimes motivated by a desire to avoid radical scepticism about the external world. One might be tempted to think that we normally have knowledge of truths such as 'I have hands', even though there is a theoretical possibility that we might be deceived about the matter. But any plausible way of answering the radical sceptic must take care to avoid making postulates so strong that perceptual appearance becomes an absolute epistemic trump. We must acknowledge that there are, at least, special circumstances in which we would have good reason to doubt our senses. For example, if you enter a funny mirror hall, and you know that it often appears to people in this hall as if they have three hands, then the fact that you seem to perceive yourself as having three hands should not make you believe that your body has sprouted a new limb. Instead, you should believe that you are experiencing an illusion. I would claim that given (3), we have grounds for concluding that we are in just such a special circumstance in which illusions are ubiquitous and in which we should distrust our senses in regard to one particular (narrowly circumscribed) set of facts, namely, facts that have to do with how we are physically implemented. For if (3) is true, then almost all people in the world have perceptions which, if interpreted naively, are misleading about such facts.

Simulated people living in ancestor-simulations may in general have beliefs as true as we have if we are not simulated. They would not suffer from a general breakdown of epistemic functioning. Only with regard to beliefs about the nature of their reality, whether it is a simulated virtual reality or basic physical reality, would they be systematically deceived (unless, that is, they had properly understood the simulation argument). This kind of error is not significantly worse, from a foundational point of view, than the error of a very naïve realist who infers from appearances that ordinary objects are completely solid at the subatomic level. It

might make him a bad solid-state physicist but need not make him otherwise incompetent.³

Thus I would maintain that externalist epistemology, of any reasonable stripe, should regard (3) as implying a case such that if we knew on theoretical grounds that this was the actual case, then we should not take our perception of two hands as giving us strong reason to think that they are two non-simulated hands.

For a sequence of hypothetical cases, where the fraction of people known to be living in simulations is 90%, 98%, 99%, 99.99%, and so on, in the limiting case, where we know that 100% of all people live in simulations, we can logically deduce that we are living in a simulation: there is no wiggle room, and we must assign a credence of 1 to the hypothesis that we are living in a simulation. Surely our credence in the other cases should gradually approach the credence in the limiting case, rather than discretely and suddenly jump to 1 when an arbitrarily small change is made from the case where 99.99999...% of people live in a simulation to the case where everybody does.⁴

4. *On the difference between the simulation argument and brain-in-a-vat arguments*

It might be worth making a brief digression here to emphasize that the simulation argument is fundamentally different from traditional brain-in-a-vat arguments. The purpose is different: not to set up a sceptical problem as a challenge to epistemological theories and common sense, but rather to argue that we have interesting empirical reasons to believe that a certain disjunctive claim about the world is true, that is, (1) \vee (2) \vee (3). The simulation argument relies crucially on non-obvious empirical premises about future technological abilities. And the conclusion of the simulation argument is not simply that we cannot be certain that we are not living in a simulation. If we knew that f_{sim} was very small but non-zero, we might not be able to be completely certain that we are not in a simulation, but this would not be a very interesting contention. (If we think that somewhere in our infinite universe there are a few ‘envatted brains’, then maybe we should not assign a strictly zero credence to our being envatted brains either, but so long as we thought that the proportion of brains in vats to brains in crania was small enough, we would have no ground for seriously doubting that we are not brains in vats, at least if we lacked specific evidence to the contrary.) The simulation argument is also different from ordinary brain-in-a-vat arguments in that it does not begin from a starting-point of doubt and ask for some compelling reason for cancelling that doubt. Rather, it begins from the starting-point that things are as we believe they are, and then, while granting us that we might be justified in assigning a high initial credence to these beliefs, nevertheless tries to show that we have specific empirically-grounded reasons for revising these initial beliefs in a certain way – not so as to make us generally agnostic about the existence of an external world, but so as to accept the disjunctive conclusion. Thus the simulation argument is not best thought of as a sceptical argument that would

³ David Chalmers makes a similar point in ‘The Matrix as Metaphysics’, philosophy section of <http://whatisthematrix.com> (2003).

⁴ Additional arguments for (BPI) were given in the original paper. See also my *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (New York: Routledge, 2002).

have us be more agnostic, but rather as an argument that would have us increase our credence in one particular disjunction (and decrease our credence in its negation). It aims to tell us something about the world, rather than to advise us that we know less about the world than we thought we did.

The *structure* of the simulation argument does not depend on the nature of the hypothetical beings that would be created by the technologically mature civilizations. If instead of computer simulations they created enormous numbers of brains in vats connected to a suitable virtual reality simulation, the same effect could in principle be achieved. One could even imagine a Truman-Show scenario, in which enormous numbers of ordinary biological humans are created and systematically misled to believe that they live in (the equivalent of) the twenty-first century by means of clever stage designs, faked historical records, etc. However, I see no empirical reason for taking very seriously the hypothesis that these kinds of deceived agents would be created in sufficient numbers to dominate either the people living in original history or the people living in ancestor-simulations. By contrast, if the computing power of a technologically mature civilization is as huge as it seems it would be, then its inhabitants could, by using only a tiny fraction of this computing power for the purpose of creating ancestor-simulations, create astronomically many. (The contrast with the Truman hypothesis is especially stark, for this would require the typical mature civilization to pour enormous resources into creating such scenarios – and even then, if there were but a few mature civilizations that chose to pursue the ancestor-simulation path instead, they could easily create many more simulated people like us than would be living in the Truman Shows.)

5. *Human vs 'suman'*

For his final objection against (BPI), Weatherson defines a gruesome predicate, 'suman'. Let *C* be some property that you can be fairly sure applies uniquely to you. For example, *C* may be something like the property of being somebody who has seen more than ten thousand daffodils today and has *As You Like It* as well as a book by Aldous Huxley lying on his desk and who is currently perceiving a small coconut plant near the right perimeter of his visual field. Given such a *C*, Weatherson then gives the definition

x is a suman \equiv_{df} x is a [non-simulated] human who is *C* or a Sim who is not *C*.

The difficulty which Weatherson now perceives (p. 430) is this:

[We] might wonder just why we might even think that [a rational person's] evidence is probabilistically independent of the hypothesis that he is [a non-simulated] human. To be sure, his evidence does not entail that he is human. But that cannot be enough to show that it is probabilistically independent. For the evidence also does not entail that he is suman. And if (P2) is true, then the evidence must have quite a bit of bearing on whether he is suman.... More generally, we still need a distinction here between the property of being human and the property of being suman that shows why ordinary evidence should be independent of the first property but not the second.

As explained above, the simulation argument does not presuppose that our evidence

is probabilistically independent of the hypothesis that we are (non-simulated) humans. Rather, it relies on (P2*), loosely interpreted as not implying strict independence. Nevertheless, it is of course correct that for (P2*) to hold, it is not enough that our evidence does not entail (SIM) or its negation. It is also true that our evidence might have quite a bit of bearing on whether one is a suman. One's evidence, we may assume, includes either the fact that one has the property *C*, or the fact that one lacks the property *C*. In the former case, one can infer that one is suman iff one is non-simulated; and in the latter case one can infer that one is suman iff one is simulated. Thus, conditional on *C*, the credence that one is suman should be equal to the credence that one is non-simulated. This conditional credence in one's being suman may be smaller or greater than, or equal to, one's unconditional credence in one's being suman (i.e., one's credence before learning that one is *C*), depending on what credence one assigns to being simulated. There is, *pace* Weather-son, no general assumption that 'ordinary evidence' should not be independent of the hypothesis that one is suman – it all depends on what one's ordinary evidence is. If one assigns a low credence to (SIM), then one's having *C* will be evidence in favour of the hypothesis that one is suman. If one assigns a high credence to (SIM), then one's having *C* will be evidence against the hypothesis that one is suman. And if one assigns a 50% credence to (SIM), then the evidence that one has *C* will be independent of (i.e., uncorrelated with) the hypothesis that one is suman.

Things would be different if we had reason to think that people with *C* were either more or less likely to be simulated than people lacking *C*. If simulators had a known preference for simulating people with *C*, then having *C* would be evidence in favour of (SIM), and the opposite would be evidence against (SIM). In fact, however, we have no grounds (or at most extremely weak grounds) for thinking that simulators would preferentially create simulations of people with *C* rather than with some other set of similar properties. In the absence of such grounds, our evidence that we have *C* has no bearing on (SIM).

There is thus no need, at least as far as the simulation argument is concerned, to search for some deep philosophical difference between predicates like 'suman' and 'non-simulated human'. We can legitimately use either predicate to ask questions or formulate hypotheses. The difference is merely one of practicality: the hypotheses most readily and conveniently expressed with the predicate 'suman' are not ones that relate very transparently and interestingly to our evidence or to the questions we are interested in asking.⁵

Oxford University

⁵ For their comments on an earlier draft, I am grateful to Jeff Medina and Brian Weather-son. I am also grateful to all those who have responded to the original paper. Some of these responses can be found at www.simulation-argument.com.