Erkenn
DOI 10.1007/s10670-012-9400-9

ORIGINAL ARTICLE

On the 'Simulation Argument' and Selective Scepticism

Jonathan Birch

Received: 5 October 2010 / Accepted: 5 September 2012
© Springer Science+Business Media B.V. 2012

Abstract Nick Bostrom's 'Simulation Argument' purports to show that, unless we are confident that advanced 'posthuman' civilizations are either extremely rare or extremely rarely interested in running simulations of their own ancestors, we should assign significant credence to the hypothesis that we are simulated. I argue that Bostrom does not succeed in grounding this constraint on credence. I first show that the Simulation Argument requires a curious form of selective scepticism, for it presupposes that we possess good evidence for claims about the physical limits of computation and yet lack good evidence for claims about our own physical constitution. I then show that two ways of modifying the argument so as to remove the need for this presupposition fail to preserve the original conclusion. Finally, I argue that, while there are unusual circumstances in which Bostrom's selective scepticism might be reasonable, we do not currently find ourselves in such circumstances. There is no good reason to uphold the selective scepticism the Simulation Argument presupposes. There is thus no good reason to believe its conclusion.

1 The 'Simulation Argument' and its Presuppositions

Nick Bostrom's 'Simulation Argument' purports to show that, unless we are confident that advanced 'posthuman' civilizations are either extremely rare or extremely rarely interested in running simulations of their own ancestors, we should assign significant credence to the hypothesis that we are simulated.[1] This remarkable argument has, perhaps more than any other in recent philosophy,

[1] See Bostrom (2003a, 2005, 2008, 2009), Bostrom and Kulczycki (2011).

J. Birch (✉)
Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, UK
e-mail: jgb37@cam.ac.uk

Published online: 13 September 2012

caught the public imagination, spawning numerous popularizations[2] and speculative discussions.[3] But a high degree of publicity merits a matching degree of scrutiny. In this discussion, I argue that Bostrom cannot plausibly ground the constraint on credence the Simulation Argument purports to yield.

Bostrom begins with an empirical premise: if our civilization reaches a 'posthuman' stage—in which our descendants' capacities "so radically exceed those of present humans as to be no longer unambiguously human by our current standards" (Bostrom 2003b)—we have reason to believe that the computing power available to our descendants will enable them to run hugely many 'ancestor-simulations', each containing vast numbers of simulated beings with conscious experiences of a type indistinguishable from our own. This claim requires that such human-type experiences are substrate independent, at least to the extent that they can be realized either in vivo or in silico; following Bostrom, I will take this as given.

Next, Bostrom argues that, in light of this, *at least one* of the following propositions must be true:

(H_1) $f_p \approx 0$. The fraction (f_p) of all human-level technological civilizations that survive to reach a posthuman stage is close to zero

(H_2) $f_I \approx 0$. The fraction (f_I) of posthuman civilizations that are interested in running simulations of their own evolution or variants thereof is close to zero

(H_3) $f_{sim} \approx 1$. The fraction (f_{sim}) of all observers with human-type experiences that live in simulations is close to unity

We do not know *which* of H_1, H_2 and H_3 are true. But we do know (Bostrom argues) that if H_1 and H_2 are false, then H_3 is true. Why? Because we know that, if the fraction of human-level civilizations reaching a posthuman stage is significantly greater than zero, and if the fraction of such civilizations interested in running ancestor-simulations is significantly greater than zero, then the computer technology available to these civilizations will allow them to run such a mind-boggling number of simulations that the number of simulated observers will vastly outstrip the number of flesh-and-blood observers.

Bostrom now takes a bold 'further step'. He argues that, conditional on H_3, "one's credence in the hypothesis that one is in a simulation should be close to unity" (Bostrom 2003a, 249). Let SIM denote the hypothesis that I am simulated. Bostrom claims:

$$\textbf{CLAIM} : Cr(\mathrm{SIM}|\mathrm{H}_3) \approx 1$$

Why believe CLAIM? Because, Bostrom argues, it derives from a more general "bland indifference principle" (BIP):

$$\textbf{BIP} : Cr(\mathrm{SIM}|f_{sim} = x) = x$$

The argument for BIP is independent of the argument for $H_1 \vee H_2 \vee H_3$. Bostrom defends BIP by means of the following consideration (henceforth: the *DNA*

[2] See, e.g., Tierny (2007), Dupré (2007) and Bostrom (2010).

[3] See e.g., Hanson (2001), Jenkins (2006), Barrow (2007), Steinhardt (2010).

analogy). Suppose I know that 60 % of human beings have a particular DNA sequence labelled S ($f_S = 0.6$). It is 'junk DNA' and correlates with no observable characteristic in those who possess it. Though I know the fraction of people that have S, I have *no other information* as to whether or not I have S. Intuitively, my credence in the hypothesis (H_S) that I have S should be 0.6. In general, an indifference principle (DNA) seems reasonable for assigning credence to H_S in light of learning that x % of the population have S:

$$\textbf{DNA} : Cr(H_S|f_S = x) = x$$

To see why BIP is justified, Bostrom argues, one need only see that we are in an analogous evidential situation with respect to SIM:

> The same reasoning holds if S is not the property of having a certain genetic sequence but instead the property of being in a simulation, assuming only that we have no information that enables us to predict any differences between the experiences of simulated minds and those of the original biological minds (Bostrom 2003a, 250).

CLAIM has some unsettling implications. Though Bostrom discusses these verbally, it will prove helpful to explicate his reasoning in terms of credence functions. Recall that, if neither H_1 nor H_2 is true, then H_3 is true:

$$\text{(i)} \quad \sim(H_1 \vee H_2) \supset H_3$$

It follows that:

$$\text{(ii)} \quad Cr(H_3) \geq 1 - Cr(H_1 \vee H_2)$$

Consequently, if I assign credence significantly less than unity to the claim that either H_1 or H_2 is true, I ought to assign credence significantly greater than zero to H_3. For instance, if I assign credence 0.7 to $H_1 \vee H_2$, I ought to assign credence greater than or equal to 0.3 to H_3. But note that:

$$\text{(iii)} \quad Cr(\text{SIM}) = Cr(\text{SIM}|H_3) \times Cr(H_3) + Cr(\text{SIM}|\sim H_3) \times Cr(\sim H_3)$$

This a priori constraint implies that, if I do assign significant credence to H_3, and if I also accept BIP, I ought to assign significant credence to SIM too. Combining (ii) and (iii) yields the following result:

$$\text{(iv)} \quad Cr(\text{SIM}) \geq Cr(\text{SIM}|H_3) \times (1 - Cr(H_1 \vee H_2))$$

The conjunction of (iv) and CLAIM then entails that, *unless I am virtually certain that either H_1 or H_2 is the case, I ought to assign credence significantly greater than zero to the hypothesis that I am living in a computer simulation.* For instance, if I assign credence 0.7 to $H_1 \vee H_2$, I ought to assign credence greater than or approximately equal to 0.3 to SIM. We are not compelled to assign *high* credence to SIM, since we may instead choose to assign most of our credence to $H_1 \vee H_2$. But unless we are *extremely* confident of the truth of $H_1 \vee H_2$, we must acknowledge SIM to be a serious epistemic possibility.

That is the core of the argument. The conclusion, if true, is profoundly unnerving. But Bostrom's route to that conclusion seems problematic. The argument for the

tripartite disjunction $H_1 \vee H_2 \vee H_3$ explicitly relies on an empirical premise concerning the physical limits of computation, namely the premise that "a posthuman civilization would have enough computing power to run hugely many ancestor-simulations even while using only a tiny fraction of their resources for that purpose" (Bostrom 2003a, 248). We have reason to accept the tripartite disjunction only if we have reason to accept this premise. With this in mind, Bostrom adduces a wealth of scientific detail in support of this claim. In the most general terms, Bostrom's defence of $H_1 \vee H_2 \vee H_3$ requires it to be the case that:

> (*Good Evidence*) Scientific evidence supports claims about the fundamental physical limits of computation.

This claim is *prima facie* plausible, and is made even more so by the detail Bostrom supplies. But the *next* stage of the Simulation Argument—the defence of BIP—tacitly requires a very different claim:

> (*Impoverished Evidence*) My current evidence does not support any empirical claims non-neutral with respect to SIM, such as the claim that I possess two real human hands.

Why does Bostrom need Impoverished Evidence? Recall the DNA analogy by means of which he motivates BIP. In assigning credence to the hypothesis that I have nucleotide sequence *S*, I have *nothing to go on* other than the information that 60 % of individuals possess *S*. This is an indispensable part of the story, for, if *S* were found to correlate with an observable characteristic, I would not continue to let an a priori indifference principle dictate my credence in H_S. Analogously, if my current evidence discriminates between SIM and $\sim$SIM, I should not let BIP dictate my credence in SIM; Bayesian conditionalization should take over. Without Impoverished Evidence, BIP is plainly unjustified.[4]

There is, however, a fairly obvious tension between Good Evidence and Impoverished Evidence. If my evidence is unable to support the mundane claim that I possess two real human hands, how can I nevertheless have good evidence for exotic claims regarding the fundamental physical limits of computation? We can turn this apparent tension into an outright contradiction by introducing a third assumption:

> (*Parity of Evidence*) My epistemic access to the facts about my own constitution is at least as good as my epistemic access to the facts about the physical limits of computation.

Good Evidence, Impoverished Evidence and Parity of Evidence are jointly incompatible. Bostrom must reject one. Each option, however, seems problematic. If he rejects Good Evidence, his argument for the tripartite disjunction falls apart. If he rejects Impoverished Evidence, his bland indifference principle is unjustified. Yet Parity of Evidence has strong intuitive plausibility. Let us consider each option in more detail.

[4] Weatherson (2003) makes a similar point. Bostrom agrees with Weatherson that BIP is only reasonable as a constraint on credences *prior* to Bayesian conditionalization; see Bostrom (2005, 92).

2 Rejecting Impoverished Evidence

Bostrom's least promising option is to repudiate Impoverished Evidence in order to preserve his original argument for the tripartite disjunction. Amending the Simulation Argument accordingly, we now begin with the concession that, while we have a great deal of evidence regarding the fundamental limits of computation, we also have evidence that is non-neutral with respect to SIM—including evidence that we are real, physically embodied human beings. This evidence warrants our assigning very low credence to SIM, and renders BIP indefensible. With this concession in place, the modified Simulation Argument still succeeds in drawing our attention to the reasonably interesting fact that, given our knowledge of the fundamental limits of computation, at least one of H_1, H_2 and H_3 is true. But no constraint on our credence in SIM follows from this result, since there is no longer any reason to allow a bland indifference principle to dictate our credence in SIM. The drawback to this response is plain: if Bostrom hopes to ground a constraint on our credence in SIM, repudiating Impoverished Evidence is tantamount to admitting defeat.

3 Rejecting Good Evidence

Suppose instead that Bostrom abandons Good Evidence, in the hope of retaining BIP and thereby preserving a defensible constraint on our credence in SIM. He is no longer able to defend the tripartite disjunction, for his defence of this disjunction required Good Evidence; but he may yet be able to defend a *quadripartite* disjunction, $\text{SIM} \vee H_1 \vee H_2 \vee H_3$. For he could reason that *either* I am simulated, in which case SIM is true (though, crucially, H_3 need not be, for the true physical limits of computation may be more restrictive than the apparent limits, and may preclude my simulators from running *hugely many* simulations—they might conceivably require a computer the size of a galaxy just to simulate *me*),[5] *or* I am not simulated and the apparent physical limits of computation are the *real* limits, in which case the empirical premise of the Simulation Argument is true, and $H_1 \vee H_2 \vee H_3$ follows.[6] Although this quadripartite disjunction is logically weaker than the tripartite disjunction Bostrom originally sought to defend, it still amounts to an interesting and provocative result.

There are three drawbacks to this response. One is somewhat pedantic: to justify the quadripartite disjunction, Bostrom would need to show how we could rule out the myriad bizarre alternatives to SIM (such as the hypothesis that I am a brain in a vat, or the hypothesis that I am being deceived by an evil demon) without ruling out SIM in the process, and it is hard to see how this could be achieved. A second is dialectical: the central novelty of the Simulation Argument is that the sceptical threat it raises is *empirically motivated*. The argument, in a nutshell, is that recent

[5] See Sect. 4 for further discussion of this point.

[6] See Bostrom (2008) for a version of the Simulation Argument along these lines. I am also grateful to Sacha Golob and an anonymous referee for independently pressing this line of response.

empirical research regarding the possibilities of posthuman computing suggests that we should take SIM a great deal more seriously than we otherwise would. If we reject Good Evidence, however, this novelty disappears: instead of drawing on empirical evidence in support of its conclusion, the modified argument assumes a pervasive scepticism from the outset. It would therefore be unlikely to convince the antecedently non-sceptical scientific realist at whom the original argument was directed.

Let us leave these drawbacks aside for now. For even if we grant Bostrom the quadripartite disjunction $\mathrm{SIM} \vee \mathrm{H}_1 \vee \mathrm{H}_2 \vee \mathrm{H}_3$, he is still unable to derive the constraint on credence his original argument purported to yield.

To see why, recall the use to which Bostrom puts the original tripartite disjunction. $\mathrm{H}_1 \vee \mathrm{H}_2 \vee \mathrm{H}_3$ entails that, if H_1 and H_2 are both false, then H_3 is true: $\sim(\mathrm{H}_1 \vee \mathrm{H}_2) \supset \mathrm{H}_3$. From this we obtained (ii):

$$\text{(ii)} \quad Cr(\mathrm{H}_3) \geq 1 - Cr(\mathrm{H}_1 \vee \mathrm{H}_2)$$

We derived a constraint on my credence in SIM, (iv):

$$\text{(iv)} \quad Cr(\mathrm{SIM}) \geq Cr(\mathrm{SIM}|\mathrm{H}_3) \times (1 - Cr(\mathrm{H}_1 \vee \mathrm{H}_2))$$

When we combine (iv) with CLAIM, it follows that, unless I have credence close to unity that $\mathrm{H}_1 \vee \mathrm{H}_2$ is true, I ought to assign credence significantly greater than zero to SIM.

By contrast, the quadripartite disjunction $\mathrm{SIM} \vee \mathrm{H}_1 \vee \mathrm{H}_2 \vee \mathrm{H}_3$ entails only that, if H_1 and H_2 are both false, then, *if SIM is also false*, H_3 is true:

$$\text{(v)} \quad \sim(\mathrm{H}_1 \vee \mathrm{H}_2) \supset (\sim \mathrm{SIM} \supset \mathrm{H}_3)$$

And from this it follows only that:

$$\text{(vi)} \quad Cr(\sim \mathrm{SIM} \supset \mathrm{H}_3) \geq 1 - Cr(\mathrm{H}_1 \vee \mathrm{H}_2)$$

This inequality will ground the constraint on credence the original argument purported to yield only if a principle analogous to CLAIM applies in this case:

$$\mathbf{CLAIM}^* : Cr(\mathrm{SIM}|\sim \mathrm{SIM} \supset \mathrm{H}_3) \approx 1$$

Is CLAIM* justified? Recall that CLAIM can be defended as a special case of a more general principle, BIP. One might hope that CLAIM* could also be defended as a special case of a more general principle, BIP*:

$$\mathbf{BIP}^* : Cr(\mathrm{SIM}|\sim \mathrm{SIM} \supset (f_{\mathrm{sim}} = x)) \geq x$$

BIP* states that, given the information that $\sim \mathrm{SIM} \supset (f_{\mathrm{sim}} = x)$, one's credence in SIM should be no lower than x. This principle, however, seems questionable. To see why, we can return to the DNA analogy Bostrom uses to motivate BIP. Suppose, as before, that I am told that $f_{\mathrm{S}} = 0.6$. Now suppose, however, that I am also told that only *non-carriers* of the relevant nucleotide sequence are reliably told the true fraction of carriers. My evidence is now *conditional*: *if* I am free of S, *then* 60 % of people indeed have S. If, by contrast, I have S myself, then the true fraction of carriers could be much lower, or much higher, than my information suggests.

In this modified scenario, my evidence is $\sim \mathrm{H}_S \supset (f_{\mathrm{S}} = 0.6)$, and my predicament is closely analogous to that of an observer who learns that $\sim \mathrm{SIM} \supset (f_{\mathrm{sim}} = x)$.

If BIP* is justified, then an analogous principle should hold in this fictional scenario:

$$\mathbf{DNA}^{*}: Cr(\mathrm{H}_S | \sim \mathrm{H}_S \supset (f_\mathrm{S} = x)) \geq x$$

DNA* implies that my credence in H_S—conditional on the information that, if I am free of *S*, then 60 % of people have *S*—should be at least 0.6.

This claim, however, seems far from intuitively compelling. The problem with DNA* can be posed in the form of a dilemma. On the one hand, if I have no antecedent reason to believe that I am not a carrier of *S*, then I have no reason to take the information at $f_\mathrm{S} = 0.6$ at face value. I consequently have no reason to conditionalize on $f_\mathrm{S} = 0.6$; and it is therefore hard to see why my credence in H_S should be constrained as if I had conditionalized on $f_\mathrm{S} = 0.6$. In this position of endemic uncertainty, indifference-based reasoning never gets off the ground: it seems far from clear what my credence in H_S should be, but there is no reason to think that it is constrained to be greater than or equal to the *apparent* fraction of carriers. On the other hand, if I *do* have an antecedent reason to believe I am not a carrier of *S*, then I do have a reason to conditionalize on $f_\mathrm{S} = 0.6$. But since I possess other evidence relevant to H_S, I have no reason to set my credence in H_S by means of an a priori indifference principle. The upshot is that DNA* seems dubious regardless of whether or not I have an antecedent reason to consider myself free of *S*.

The DNA analogy therefore fails to support BIP*. But in the absence of any other motivation for BIP*, the rejection of Good Evidence and consequent retreat to a quadripartite disjunction fails to preserve Bostrom's original constraint on credence. Only one option remains: for his argument to succeed, Bostrom must reject Parity of Evidence.

4 Rejecting Parity of Evidence

If the argument of the preceding sections is correct, the Simulation Argument indispensably presupposes a curious form of selective scepticism. In the early stages of the argument, Bostrom draws on empirical evidence to defend speculative claims about the potential power of posthuman computing. In the latter stages, he assumes that my evidential situation with respect to the physical reality of my own hands is no better than my evidential situation with respect to the hypothesis that my cells contain some random stretch of junk DNA. To save his argument, Bostrom needs to explain how this remarkable conjunction of scientific realism and limb scepticism can be sustained. How could it be the case that I possess good evidence for claims regarding the physical limits of computation and yet lack good evidence for claims regarding my own physical constitution? How could Parity of Evidence be false? On the face of it, it is difficult to see how any epistemic predicament could imperil beliefs of the latter sort and yet fail to imperil beliefs of the former sort. In this section, I want to examine three ways in which this selective scepticism might nevertheless be defended.

4.1 The Substrate-Independence of Computing Power

One possible defence is to argue that my access to the facts about the physical limits of computation is indeed better than my access to the facts about my own material constitution, because the facts about the limits of computation are independent of the material substrate in which computational processes are realized. Spelt out more precisely, this line of thought goes as follows: none of my evidence can tell me whether why the world around me is physically real or simulated. Either way, however, it evidently contains computers of one sort or another. I do not know whether these are real computers running on a physical (silicon) substrate, or whether they are 'virtual machines' running within a larger simulation (cf. Bostrom 2003a, 253). But I don't need to know this to determine the true limits of computation per unit substrate, because those limits are the same regardless of the substrate in question—and regardless of whether the substrate is real or virtual.[7]

The problem with this response is that, if current appearances are any guide, the empirical assumption on which it relies seems highly doubtful: there is no good reason to think that the limits of computation are independent of the material substrate in which the relevant computational processes are realized. Indeed, as Bostrom notes, estimates based on the assumption of conventional material substrate might prove extremely conservative, since alternative substrates such as nuclear matter or plasma might enable faster computing (Bostrom 2003a, 246). By the same token, if suitable materials were in short supply, the power of posthuman computing might be severely curtailed. The upshot is that observations of *simulated* computers will not reliably indicate the true physical limits of the *real* computers on which the simulation is running, because those computers may well be instantiated in a material substrate with very different properties. Moreover, the physical limits of computation depend not only on the specific properties of available substrates, but also on more fundamental physical laws (Bostrom cites the Bremermann-Beckenstein bound and the black hole limit; Bostrom 2003a, 245). Since the laws of physics in the world outside may be very different from the simulated laws, this is another reason to suspect that, if we are simulated, the true physical limits in the world outside the simulation may bear little resemblance to the apparent limits.

4.2 Establishing a Lower Bound

It is possible to foresee an objection at this point. One might concede that observations of a simulated computer cannot tell us the *exact* physical limits of posthuman computation in the world outside the simulation (since the material substrates in the real world may have very different properties to those in the simulated world), but one might still maintain that such observations tell us *something* relevant about those limits. For one might suppose that, even if we are simulated, we can ascertain a *lower bound* on the true limits of computation.

In more detail, the objection goes like this: if I am not simulated, then my evidence is a reliable guide to the true physical limits of computation. But even if I

[7] I thank an anonymous referee for suggesting a response along these lines.

am simulated, I still have some evidence to go on, because the world around me contains simulated computers. Admittedly, these virtual machines may well differ in many respects from the real computers on which they are running. Nevertheless, I can safely infer that the processing capacity of these real computers is *at least* as great as that of the virtual machines they are able to generate. Hence, by investigating the properties of the simulated computers in my surroundings, I can establish a lower bound on the true physical limits of computation. In particular, since I have evidence that a virtual computer the size of a planet could simulate hugely many conscious beings, I also have evidence that a *real* computer of similar size would be capable of similar feats.

This line of thought is seductive but flawed. The difficulty is the tacit assumption that, in order to create the *appearance* that my environment contains computers with sufficient processing power to run hugely many ancestor-simulations, my simulators would need to generate a virtual machine that genuinely possessed the requisite processing capacity—and that my simulators would therefore need access to a real machine of greater or equal capacity. There is no good reason to think that this is the case. The mere *appearance* of hugely powerful machines could be simulated far more straightforwardly by simulating the experiences of a single observer: no virtual machines would need to be constructed at all.

Because of this, the mere appearance of powerful computers in the environment of a simulated observer tells that observer very little about the real processing power of computers in the world outside. The true limits of computation (i.e., the true limits on processing power per unit substrate) could be greater than or equal to the apparent limits, but they might also be much *lower* than they appear to be. It is conceivable that, for one reason or another, posthuman civilizations delight in creating simulated observers for which the apparent laws of physics are far more permissive than the true laws. The true laws, meanwhile, could be such that a typical posthuman civilization is able to simulate very few such observers, not hugely many.

The implication is that an observer with no evidence to locate herself among the physically real inhabitants of the world would face a grim epistemic predicament with respect to the true limits of computation. Simulated experiences may be radically non-veridical, and making inferences about the physics of the outside world on the basis of such experiences is fraught with difficulty. In fact, a simulated observer could infer with confidence only *one* claim about the limits of computation: namely, that these limits are such that *at least one* posthuman civilization could run *at least one* ancestor-simulation. That claim, however, is not enough for Bostrom's argument to go through. I submit, therefore, that appealing to the ability of a simulated observer to investigate its simulated environment is not enough to ground the selective scepticism Bostrom's argument requires.

4.3 Elga on Scepticism and Self-Location

On the face of it, the prospects for Bostrom's brand of selective scepticism look bleak. Yet there are unusual circumstances in which the selective scepticism the Simulation Argument requires might be defensible. Such circumstances arise due to

an important difference between the kind of beliefs I possess with respect to the physical limits of computation and the kind of beliefs I possess with respect to my own physical constitution. Scientific beliefs are typically *de dicto* (sensu Lewis 1979). *De dicto* beliefs locate the actual world to within a set of possible worlds: for instance, my beliefs about the laws of physics locate the actual world to within the set of worlds at which these laws hold. Importantly, however, no amount of *de dicto* belief will suffice for me to locate *myself* within a world: it can specify which world I am in, but it cannot specify which *individual* within that world is *me*, nor can it specify what *time* in that world is *now*. In more precise terms, it does not specify which *centred world* I am in, where a centred world is conceived as an ordered triple of a world, an individual and a time (see Lewis 1979). The belief that I possess two physically real human hands is different: it *does* have a self-locating (or *de se*) component. For, in addition to delimiting the set of worlds I might be in, it helps me locate myself within those worlds, by placing me among the *physically real, human,* and *two-handed* inhabitants. This belief rules out a range of centred worlds within the actual world, including those corresponding to simulated observers, non-human observers, and one-handed observers.

Elga (2004) has argued that, in certain bizarre circumstances, we may possess a great many justified *de dicto* beliefs while lacking the justified *de se* beliefs we would need to locate our experiences as those of a particular individual at a particular time. These are circumstances in which we have good reason to think our current subjective experiences are veridical, yet lack good evidence regarding the centred world at which those experiences are instantiated. Suppose, for instance, that there is another individual in the actual world whose experiences are *subjectively indistinguishable in every respect* from my own (let's suppose it is a brain in a vat). What credence should I assign to the hypothesis (BIV) that my current experiences are those of the envatted brain rather than the human? It is tempting to say that I must assign low credence to BIV on pain of a debilitating global scepticism, but this need not be the case. As the human and the brain are inhabitants of the same world, the same propositions are true regardless of whether I am the former or the latter. On the assumptions that (1) the human has a great many justified beliefs about the external world, obtained through veridical experiences, and (2) the human and the brain have the same beliefs and the same evidence,[8] then the envatted brain will also have a great many justified beliefs about the external world. All that is imperilled in this outré scenario is my justified *de se* belief, since, given (2), my evidence does not discriminate between BIV and $\sim$BIV. Elga calls this *self-locating scepticism*. Self-locating scepticism, he argues, does *not* entail scepticism with respect to one's *de dicto* beliefs.

Much of this is controversial.[9] Nevertheless, I think Elga succeeds in illustrating one set of circumstances in which the scepticism required by Bostrom's argument (that is, scepticism with regard to *any* information that would locate one among the set of *physically real* observers) might be compatible with the retention of justified

[8] This strong brand of evidential internalism is of course contentious; see Williamson (2000) for criticism. I grant it here so as to give Bostrom's argument the best chance of succeeding.

[9] For trenchant criticism of Elga's argument, see Weatherson (2005).

beliefs about the true physical limits of computation. These are circumstances in which I can locate myself to one of two or more *subjectively indistinguishable* centred worlds within the same world, such that (i) one observer is physically real while the other is simulated, (ii) the physically real observer holds justified beliefs about the physical limits of computation, obtained through veridical experiences, and (iii) both observers have the same beliefs and the same evidence. In short, if I had reason to think that my experiences were realized *twice* in the actual world—once in vivo, once in silico—I would have grounds for selective scepticism. I would have no justification for ascribing any properties to myself that were not shared by my simulated duplicate, yet my justification for my *de dicto* beliefs would not be imperilled.

Note, however, that, on Elga's account, I must have reason to believe that the two centred worlds are *subjectively indistinguishable*. It is worth pausing to consider the rationale for this requirement. If the two centred worlds are subjectively indistinguishable, then my inability to locate myself as the non-envatted observer does not undermine my justification for any of my *de dicto* beliefs, since both observers have exactly the same evidence. By contrast, if I am aware of *any* difference in evidence between the non-envatted and envatted predicaments, then this will undermine my justification for at least some of my *de dicto* beliefs, because any such difference introduces a respect in which the evidence of the simulated observer is potentially non-veridical with regard to the nature of the external world. Suppose, for instance, I know that the envatted brain has experiences subjectively indistinguishable from my own in every respect, *except that the colour of bananas is different for the envatted brain*. This difference, small as it may be, would still be enough for my self-locating scepticism to undermine the justification for *some* of my *de dicto* beliefs, namely, those pertaining to the colour of bananas. Only by requiring that the two predicaments are subjectively indistinguishable in *every* respect can Elga ensure that scepticism about *de se* belief cannot undermine *any* of my *de dicto* beliefs.

This, however, is a stringent requirement, and one that makes Elga's brand of selective scepticism significantly different from Bostrom's. Even if we accept that posthuman civilizations could run hugely many simulations of *human-type* experiences, it seems highly unlikely that any of these experiences would be *subjectively indistinguishable in every respect* from those of a particular flesh-and-blood observer, so the kind of scenario Elga describes seems unlikely to arise in the case of posthuman simulations. But can we relax the assumption of perfect mental duplication and still retain a form of selective scepticism capable of rescuing the Simulation Argument?

Suppose that, although the experiences of simulated observers may differ significantly in many respects from those of their flesh-and-blood ancestors, I have reason to believe that real and simulated predicaments are, in general, equally veridical in *all respects epistemically relevant to claims about the fundamental physical limits of computation*. In these circumstances, my inability to locate myself among the real observers would undermine my justification for many of my *de dicto* beliefs, but it would *not* undermine my justification for the *specific class* of *de dicto* beliefs that justify the tripartite disjunction $H_1 \vee H_2 \vee H_3$—since my evidence for

these beliefs would be equally veridical regardless of whether I am real or simulated. If this form of selective scepticism is defensible, it gives us a reason to reject Parity of Evidence.

What is needed for the Simulation Argument to succeed, therefore, is some reason to believe that real and simulated experiences are, in general, equally veridical guides to the limits of computation. If I had reason to believe that real and simulated experiences were equally veridical with respect to the laws of physics, and with respect to the properties and availability of suitable material substrates, my justification for beliefs about the limits of computation would not be imperilled by my inability to locate myself among the real observers.

The difficulty is that I have no such reason. For as we noted above, there is no reason to suppose that posthuman civilizations would not radically mislead their simulated creations with regard to the true laws of physics, and the true properties of material substrates. The result is that my inability to locate myself among the real observers would lead to a pervasive *de dicto* scepticism about *all* aspects of physical reality, including those aspects epistemically relevant to the limits of posthuman computation.

5 Conclusion

Bostrom's Simulation Argument requires that we possess good evidence for claims about the physical limits of computation and yet lack good evidence for claims about our own physical constitution (Sect. 1). Although the argument can be modified in one of two ways to obviate the need for this intuitively implausible conjunction, neither grounds the constraint on credence the original argument purported to yield (Sects. 2, 3). Hence, to preserve his original conclusion, Bostrom must embrace a curious form of selective scepticism. One might hope to defend such a position on the grounds that, although my evidence cannot locate me among the physically real observers, it can establish a *lower bound* on the true physical limits of computation. This defence, however, is not successful (Sect. 4). Alternatively, one might defend such a position by appeal to the distinction between *de dicto* and *de se* belief. Strange as it seems, there *are* unusual circumstances (suggested by Elga) in which selective scepticism might be defensible on these grounds. To the best of our knowledge, however, we do not currently find ourselves in such circumstances (Sect. 4). There is, at present, no good reason to endorse the curious combination of scientific realism and self-locating scepticism that Bostrom's argument requires. There is thus no good reason to believe its conclusion.

Acknowledgments I thank Sorin Bangu, Alex Broadbent, Jeremy Butterfield, Sacha Golob, Nick Jardine, Tim Lewens, Brian Weatherson, three anonymous referees, and an audience at the University of Cambridge for helpful comments. This work was supported by the Arts and Humanities Research Council.

References

Barrow, J. D. (2007). Living in a simulated universe. In B. Carr (Ed.), *Universe or multiverse* (pp. 481–486). Cambridge: Cambridge University Press.

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

Bostrom, N. (2003b). *The transhumanism FAQ*. Retrieved on 25 March 2012 from http://www.transhumanism.org/resources/FAQv21.pdf.

Bostrom, N. (2005). The simulation argument: Reply to Weatherson. *Philosophical Quarterly, 55*, 90–97.

Bostrom, N. (2008). *The simulation argument FAQ*. Retrieved on 25 March 2012 from http://www.simulation-argument.com/faq.html.

Bostrom, N. (2009). The simulation argument: Some explanations. *Analysis, 69*, 458–461.

Bostrom, N. (2010). Ideas of the century: The simulation argument. *Philosopher's Magazine, 50*, 28–29.

Bostrom, N., & Kulczycki, M. (2011). A patch for the simulation argument. *Analysis, 71*, 64–71.

Dupré, B. (2007). *50 Philosophy ideas you really need to know*. London: Quercus.

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

Hanson, R. (2001). How to live in a simulation [Online]. *Journal of Evolution and Technology, 7*. Retrieved on 25 March 2012 from http://www.transhumanist.com/volume7/simulation.html.

Jenkins, P. S. (2006). Historical simulations—Motivational, ethical and legal issues. *Journal of Future Studies, 11*, 23–42.

Lewis, D. K. (1979). Attitudes *de dicto* and *de se*. *Philosophical Review, 88*, 513–543.

Steinhardt, E. (2010). Theological implications of the simulation argument. *Ars Disputandi, 10*, 23–37.

Tierny, J. (2007). Our lives, controlled from some guy's couch. *New York Times*, 17 August 2007.

Weatherson, B. (2003). Are you a sim? *Philosophical Quarterly, 53*, 425–431.

Weatherson, B. (2005). Should we respond to evil with indifference? *Philosophy and Phenomenological Research, 70*, 613–625.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.