The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

TWO NEW DOUBTS ABOUT SIMULATION ARGUMENTS

Micah Summers and Marcus Arvan

ABSTRACT

Various theorists contend that we may live in a computer simulation. David Chalmers in turn argues that the simulation hypothesis is a *metaphysical* hypothesis about the nature of our reality, rather than a sceptical scenario. We use recent work on consciousness to motivate new doubts about both sets of arguments. First, we argue that if either panpsychism or panqualityism is true, then the only way to live in a simulation may be as brains-in-vats, in which case it is unlikely that we live in a simulation. We then argue that if panpsychism or panqualityism is true, then viable simulation hypotheses are substantially sceptical scenarios. We conclude that the nature of consciousness has wide-ranging implications for simulation arguments.

KEYWORDS consciousness; panqualityism; panspychism; simulation; scepticism.

1. Introduction

Nick Bostrom [2003] argues that it is highly probable that we live in a computer simulation. Other theorists argue that versions of the simulation hypothesis may explain features of fundamental physics, such as quantum phenomena [Whitworth 2008; Arvan 2014], relativistic physics [Whitworth 2008, 2010], or cosmological fine-tuning [Mizrahi 2017]. David Chalmers [2005] in turn argues that if we do live in a simulation, then most of our beliefs are true—that is, that the simulation hypothesis is a *metaphysical* hypothesis about the nature of our world, rather than a sceptical hypothesis calling into doubt our knowledge of it. However, others doubt whether it is probable that we live in a simulation (Weatherson [2003]; Brueckner [2008]; Birch [2013]; cf. Huemer [2016]), and others argue that some variants of the simulation hypothesis are genuinely sceptical scenarios [Hanley 2017]. The present article uses recent work in the philosophy of consciousness, including Chalmers'

[2013] work on panpsychism and 'panqualityism', to deepen these critiques and to challenge Chalmers' interpretation of simulation scenarios. Section 2 argues that if either panpsychism or panqualityism is true, then there may be only one possible way to live in a simulation: namely, as brains-in-vats hooked up to an external simulation. Following this, we argue that if this is the only way that our world could be a simulation, then—*contra* Bostrom—the probability that we live in a simulation is immeasurably small; and *contra* other simulation theorists, it is unlikely that some simulation hypothesis is a true explanation of fundamental physics. Section 3 then argues that if either panpsychism or panqualityism is true, then viable simulation hypotheses are (*contra* Chalmers) substantially sceptical scenarios. Finally, Section 4 responds to objections. We conclude that the nature of consciousness has far-reaching implications for the likelihood that we live in a simulation, and for the metaphysical and epistemological implications thereunto.

2. Why There May Only Be One Way to Live in a Simulation

The most influential argument that we may live in a simulation, Bostrom's [2003] Simulation Argument, holds that one of the following propositions must be true:

1. We will likely not reach 'posthumanity', a future stage where humanity has achieved most technological possibilities; or
2. Very few posthuman civilizations (PHC) will want to create simulations; or
3. Most people with experiences like ours live in a simulation [ibid.: 255].

Bostrom begins by assuming *substrate independence*, the assumption that 'mental states can supervene on any of a broad class of physical substrates' [ibid.: 244]. According to this assumption, phenomenally conscious mental states—such as what pain is like—can supervene not only on the states of fleshy, carbon-based brains like ours, but also on digital analogues of fleshy brains, such as the 'brains' of *simulated beings*. This assumption plays a critical role in Bostrom's argument, which goes as follows. First, Bostrom defines a

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

posthuman civilization (PHC) as a society that has 'acquired most of the technological capabilities that one can currently show to be consistent with physical laws and with material and energy constraints' [ibid.: 245]. Second, Bostrom argues that a PHC could make *ancestor-simulations*, or simulations of people who previously existed [ibid.: 248]. Third, he argues that a PHC would have the computing resources 'to run *hugely many* ancestor-simulations even while using only a tiny fraction of their resources for that purpose' ([ibid.]; italics added). Fourth, he argues that it may be possible for simulated civilizations to become PHCs themselves, running ancestor-simulations on *virtual computers* within their own ancestor-simulation [ibid.: 252–3]. Fifth, from these premises, Bostrom derives the following equation describing the *fraction* of human-like beings with experiences like ours who live in ancestor simulations [ibid.: 248]:

$$f_{sim} = \frac{f_p f_1 \bar{N}_1}{(f_p f_1 \bar{N}_1) + 1}$$

Here, f_{sim} refers to the fraction of human-like beings with experiences like ours in ancestor-simulations; *fp* to the fraction of human civilizations that reach posthumanity; f_1 to the fraction of PHCs interested in constructing these simulations; and $\bar{N}_1$ to the average number of ancestor-simulations run. Sixth, based on the claim that $\bar{N}_1$ is very large—given that PHCs would be able to run hugely many ancestral-simulations—Bostrom argues that there are only three possibilities: (1) $f_p \approx 0$ (the fraction of PHCs is near zero); (2) $f_1 \approx 0$ (the fraction of PHCs that want to make ancestor simulations is near zero); or (3) $f_{sim} \approx 1$ (the majority of human-like beings with experiences like ours exist in a simulation). Finally, using a *bland indifference principle*—which holds that absent further evidence, we should apportion credences equally across all relevant possibilities—Bostrom infers that, 'We can take a further step and conclude that conditional on the truth of (3), one's credence in the hypothesis that one is in a simulation should be close to unity' [ibid.: 249–51].

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

Notice that Bostrom's argument hangs on the aforementioned substrate-independence assumption. For it to be possible that the vast majority of beings with experiences like ours are simulated beings, it must be possible for merely simulated beings *to have experiences like ours*. However, Bostrom does not defend substrate independence in detail. Instead, he writes, 'Arguments for this thesis have been given in the literature, and although it is not entirely uncontroversial, we shall take it here as given' [ibid.: 244]. He then adds:

> The argument we shall present does not … depend on any very strong version of functionalism or computationalism. … We need only the weaker assumption that it would suffice for the generation of subjective experiences that the computational processes of a human brain are structurally replicated in suitably fine-grained detail, such as on the level of individual synapses. This attenuated version of substrate-independence is quite widely accepted [ibid.].

This attenuated assumption may be widely accepted. Yet, is it true? Recent findings suggest grounds for doubt. First, whereas neural communication was once thought to be digital—as neural action-potentials were thought to be all or nothing affairs, with a spike representing '1' and absence of a spike '0'—it has been shown that different *physical magnitudes* of action potentials can have different neurological effects [Maley 2020]. Second, neurons have been recently found to communicate with each other *at a distance* through continuously varying the electromagnetic fields surrounding their cell bodies [Chiang et al. 2018].

In short, whereas brains were once thought to function digitally, it increasingly appears that they function as *analogue* computers [Maley 2018a, 2018b]. This is important because digital and analogue computation are different in kind [Maley 2011]. Although digital computers can emulate analogue computation, digital computation occurs through manipulating discrete values ('1' and '0'), whereas analogue computation involves continuous changes in physical magnitudes. To understand the difference, consider an

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

analogue thermometer, which uses mercury to measure temperature. Here, the measurement of temperature does not involve computing discrete values (0 through 100+ degrees). Rather, the thermometer measures temperature by mercury itself *physically expanding*. A digital computer can emulate this process by *simulating* it—that is, by assigning 'degrees of expansion' to numbers between 0 and 100, and representing different numbers between 0 and 100 through different digital series of 1's and 0's. However, such a simulation would be merely that: a *digital emulation* of mercury's expansion. A digitized thermometer does not *actually* involve mercury expanding. Rather, it merely computes a virtual facsimile of that physical phenomenon.

Here is why this matters. Once we distinguish analogue from digital computation, we can see that there are plausible grounds for doubting Bostrom's substrate independence assumption (See Arvan and Maley [in preparation]; cf. Koch [2019]). For example, consider standard arguments against functionalist theories of mind. The standard criticism of functionalism is that phenomenal qualities (such as experiential pain, redness, etc.) appear to be irreducible to functional states [Russell 1921; Chalmers 1996]. Whereas the nature of a functional state is purely relational (as a function simply is a relation between variables), phenomenal states appear to have intrinsic, non-relational properties: for example, *redness itself*. This line of argument has led numerous authors to defend *panpsychism*, the view that properties of phenomenal consciousness (or alternatively, 'proto-phenomenal properties') are fundamental properties of the physical world akin to fundamental physical properties such as mass, charge, or spin [Chalmers 1996; Strawson 2006, 2016; Goff 2017]. Chalmers has also recently defended an even more elaborate view: *panqualityism*, the view that qualities (such as greenness, redness, etc.) literally pervade the physical world, not as phenomenally conscious experiences, but instead as qualities *in* things (greenness being a quality in green grass) that can in turn *be* experienced by conscious beings [Chalmers 2013: sec. 7]. Notice,

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

finally, that qualities as such—whether they be essentially mental (as in panpsychism) or non-mental as well (as in panqualityism)—also appear to be analogue. What orange looks like, for example, is not merely an 'on/off' matter: orange appears to be a *phenomenal magnitude* that can differ continuously in all kinds of ways (hue, brightness, etc.). While digital computers may be able to emulate colours (assigning 'orange' to a series of 1's and 0's) and various magnitudes thereof (such as hue), if panpsychism or panqualityism is true then digital emulations of these physical processes are *merely emulating* what panpsychists and panqualityists take to be fundamentally qualitative features of the physical world.

We do not mean to commit ourselves to the truth of panpsychism or panqualityism. We also cannot purport to show just how likely it is that cognition, phenomenal consciousness, or qualities in nature are analogue. Although we believe what we outlined above—that emerging empirical evidence suggests that human brains may be analogue devices, and that the qualities we experience in phenomenal consciousness appear to be analogue as well—we do not think that *anyone* is in a good position to know how likely these things are, given the current states of fundamental physics, neuroscience, and persistent disagreement over how phenomenal consciousness relates to physical and functional states. Consequently, we maintain that theorists such as Bostrom are unjustified in simply assuming substrate independence. Might Bostrom merely assume that substrate independence is *more likely* than its denial? He might, of course. However, in that case, his Simulation Argument hangs on a controversial premise that, if we are correct, there are good reasons to doubt. In any case, we contend that (A) panpsychism and panqualitysm are serious philosophical hypotheses that some have argued to be the most plausible resolutions to the hard problem of consciousness, and (B) they both raise serious doubts about substrate independence. If panpsychism or panqualityism is true, then it may be that for creatures to have phenomenal experiences at all—let alone rich, coherent phenomenal experiences like yours or mine—their

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

brains must be *analogue*, manipulating fundamental physical magnitudes (mass, spin, etc.) and phenomenal magnitudes (redness, orangeness, etc.) that inhere in the fundamental level of physical reality *in the right way*. Finally, for reasons outlined above, digital simulations may be incapable of doing this, merely emulating physical-phenomenal processes in digital ways that leave 'simulated beings' with either no phenomenally conscious experiences at all or experiences that are incoherent.

Might it be possible for PHCs to create vast numbers of *analogue* ancestor simulations? If so, then even if substrate independence fails and phenomenal consciousness cannot be realized digitally, Bostrom's argument might still go through: the vast majority of people with experiences like ours might be simulated beings living in *analogue* simulations. However, there are several related problems here. First, in our actual world history, the use of analogue computers has vastly declined in favour of digital ones for reasons having to do with size, economic cost, precision, and 'nonideal' features of analogue computing, including but not limited to issues with temperature coefficients and parasitic effects in semiconductors. Second, although hybrid analogue-digital computers appear to have computational advantages over digital computers [Hardesty 2016], it is unclear whether fully analogue ancestral simulations (of the sort potentially necessary to reproduce human experiences like ours) would be feasible, or why the members of a PHC would go through the trouble to create them. Indeed, a third problem here is that existing analogue computers only manipulate a small number of physical parameters: specifically, DC and AC voltage, frequency, and phase. Existing analogue computers do not manipulate the vast majority of fundamental physical features of our world, such as mass, spin, weak or strong nuclear force, and so on. Consequently, it is unclear how analogue computers *could* 'emulate' our fleshy brains without simply *being* fleshy brains. After all, our brains just are bundles of cells, axons, dendrites, and synapses constructed out of fundamental particles. Now, we do not know the

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

extent to which human brains may or may not manipulate basic physical properties such as gravity, spin, weak and strong nuclear force, and so on, in generating phenomenally conscious experience. Yet, the particles constituting human brains clearly *do* have mass, charge, spin, strong and weak nuclear force in ways that (if panpsychism or panqualityism is true) *may* make an important difference in phenomenal experience. Consequently, constructing an 'analogue simulator' of a human brain—one that reproduces *all* of our experiences—may only be possible in one way: by *creating a literal brain*, not a 'simulation.' Finally, it may be impossible for members of PHCs to ever understand the true relationships between phenomenal consciousness and fundamental physics [McGinn 1989]. Given that panpsychism and panqualityism are 'dual-aspect' theories holding that the physical world has 'two sides'—a side of physical properties (mass, charge, etc.) and another side including phenomenal properties (redness, greenness, etc.)—it could well be prohibitively difficult or even impossible for PHC scientists to understand how the physical and phenomenal relate to each other in a fine-grained enough fashion to create an analogue ancestral-simulation that reproduces experiences like ours.

If our argument above is correct—if, that is, substrate independence is false—then there may be only one possible way for our world to be a simulation: namely, for analogue brains to be hooked up to a digital simulation, as in the film *The Matrix*. However, if this is the case, then Bostrom's third disjunct is false: even if PHCs created vastly many ancestor-simulations, *next to none* of those simulations would contain beings with experiences like our own. Only those where PHCs hooked up as *brains-in-vats* to ancestor-simulations would have such experiences. Yet, there are plausible grounds for thinking that PHCs would be highly unlikely to set up vastly many brains-in-vats featuring experiences like ours (*vis-à-vis* a high value for $\bar{N}_1$). First, in order to become PHCs, civilizations would have to survive long enough to develop vast technological resources—as well as successful social-political

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

systems—that would plausibly make their world a better one to live in than our own present-day. Second, as we see in videogames and cultural representations of 'brain-in-vat'-like scenarios—such as *The Matrix* and *Ready Player One*—the vast majority of simulated realities that people are interested in experiencing are unlike our actual reality, tending to instead to involve fantasy elements (such as magic, superpowers, etc.) and 'fun story-lines' (saving the world) that are vastly more entertaining than the often-banal world in which we live. Why, exactly, would a member of a PHC want to *experience a life like mine*, rather than (say) the life of a rock-star, wizard, or action-hero? Many people's lives are, after all, not only replete with banal repetition, but also with apparently-meaningless suffering: some people are murdered, others suffer or die from horrible diseases or accidents; we may witness the premature deaths of loved ones, endure long-term incarceration, suffer from depression, endure abuse or injustice, or simply fail to realize any of our hopes or dreams. While one can conceive of *possible* reasons why PHC-inhabitants might be hooked up as brains-in-vats to have experiences like ours—such as the desire to experience the lives of their ancestors, or to enslave large numbers of their inhabitants in a simulated world (as in the original *Matrix* film)—the fact that so many human lives are a decidedly mixed bag (or worse) provides reasons to doubt whether very PHCs are *likely* to hook up large numbers of people as 'brains-in-vats' to have experiences like ours. This seems particularly plausible for two related reasons: (1) to become a PHC rather than go extinct, a civilization would plausibly have to progress morally, politically, and technologically to a point where it seems unlikely that they would need to or want to virtually enslave vast numbers of people (cf. Bostrom [2013: sec. 6]); and (2) unlike in videogames, where one can 'quit' cost-free if one finds the game boring or unenjoyable, in *our* world the only evident way to 'quit' is through suicide (which, for all we know, is the end of our life altogether).

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

Finally, similar considerations cast doubt on simulation arguments concerning fundamental physics. If the only way that we can live in a simulation is as brains-in-vats, then we have independent grounds for thinking that it is unlikely that we would be brains-in-vats *and* experience a world with a physics like our own. Once again, people in our world primarily pursue virtual reality experiences to play out exciting fantasies (of the sort that occur in videogames or in *The Matrix*). Consequently, even if some version of the simulation hypothesis can 'explain' features of fundamental physics, we have plausible grounds for thinking that it is unlikely to be the true explanation. Our general psychological evidence suggests that if we were PHC-era brains-in-vats, chances are that we would experience a world with (1) a very different kind of physics (one where things like magic, superpowers, etc. exist), in which (2) we would also experience more exciting and satisfying life-stories. Again, one can imagine *possible* scenarios where members of PHCs might hook people up in large numbers to experience a simulated world with lives and physics like ours. However, in evaluating the likelihood of this, we should apportion our credences based upon all of our evidence, including evidence that it seems *unlikely* that many members of PHCs would want to spend significant amounts of time (let alone entire 'lifetimes') envatted in a world with experiences and physics like our own. Finally, even if some version of the simulation hypothesis can 'explain' observed physics, a simpler hypothesis is arguably more likely: that we live in the 'ground-floor' (non-simulated) reality, and it is functionally analogous to a simulation without *being* one [Arvan 2013: 45-6].

Now, for all we have shown, perhaps panpsychism and panqualityism are both false. Further, if panpsychism or panqualityism is true, then perhaps substrate independence somehow still holds. Third, if either panpsychism or panqualityism is true, then perhaps inhabitants of PHCs would know how conscious experiences like our own can be recreated in digitally simulated beings, or how to create analogue ancestral-simulations that would do so.

We cannot definitively rule out of any of these possibilities. The point is simply that we currently have no clear grounds for thinking that any of them are true. We thus submit that there are new reasons to doubt Bostrom's Simulation Argument, as well as simulation arguments concerning fundamental physics. There may be exactly *one* way to 'live in a simulation': namely, for us to be brains-in-vats. But, in that case, as we have seen, it is unclear why members of PHCs would create large numbers of brains-in-vats with experiences *or* physics like ours—and, as Huemer argues, the probability that we are brains-in-vats appears to be vanishingly small [Huemer 2016].

3. If Panpsychism or Panqualityism, then the Matrix is a Sceptical Hypothesis

Now consider Chalmers' argument against the idea that the Matrix is a sceptical hypothesis. Chalmers argues that it is plausible—even before considering the notion of a simulation—that our reality had a Creator, that microphysical processes are computational, and that the mind and body are distinct. Chalmers terms the conjunction of these three claims 'The Metaphysical Hypothesis' about the nature of reality [Chalmers 2005: 141-2]. Chalmers then argues that The Metaphysical Hypothesis is *indistinguishable* from The Matrix Hypothesis (that we are brains-in-vats hooked up to a simulation). After all, The Matrix would have a Creator, its nature is computational, and our minds would be distinct from the simulation. Consequently, Chalmers concludes that The Matrix Hypothesis is not a sceptical one calling into doubt our knowledge of the external world. Instead, it is a metaphysical hypothesis about the nature of our reality (cf. Chalmers [2017]).

Hanley [2017] contends that Chalmers' argument here is at most partly successful. Hanley gives three situations: one where he is not envatted and not made of computational bits, one in which he *is* envatted but not made of bits, and a third scenario in which he is envatted *and* made of bits. Hanley calls this a *Cartesian predicament* to point out the obvious ways in which these different scenarios appear—much as Descartes' original evil demon

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

argument—to be scenarios one cannot rule out as false. Finally, Hanley argues that these scenarios are sceptical scenarios, in that nothing in one's present experience can enable one to tell them apart.

Our argument in Section 2, however, reveals that Chalmers is in a worse situation than this. As Stoljar [2020] argues, Chalmers' structuralism—his epistemological view that we come to have knowledge of the world through experiencing *its structure*—appears to be inconsistent with Chalmers' own views on consciousness. For Chalmers, consciousness is a *non*-structural and fundamentally qualitative side of reality. In the past, Chalmers [1996] has endorsed panpsychism (or 'panprotopsychism'), and in more recent work [Chalmers 2013] he has proposed *panqualityism* as a Hegelian synthesis that can resolve apparent problems with panpsychism (principally the 'combination problem'). Let us now think about the epistemic implications of these views, specifically in light of our argument in Section 2 that, if either view is correct, there may be only *one* possible way to live in a simulation (as brains-in-vats).

If panpsychism is true, then if I am a brain-in-vat, it is entirely possible *that I am the only person consciously experiencing the simulated reality around me.* For all I know, every other human and non-human creature I interact with is merely a digital emulation: a non-conscious facsimile of a person or animal. But, in that case, many of my most central factual beliefs about the 'world' around me—that there are other humans and nonhuman animals in it who think, feel, and suffer—are systematically false. Further, it would then arguably be the case that most of my ethical beliefs would be false as well, as consciousness is plausibly a necessary condition for moral status [Shepherd 2018]. If there is nothing 'it is like' to be any of the 'people' or 'animals' around me, then it is hard to see how I have any moral duties to them. At most, my moral duties would merely be *apparent* duties, as I might never be in a position to know that I am a brain-in-a-vat and that everyone around me lacks consciousness. The point then is this: if panpsychism is true and our argument in Section 2 is correct, then

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

the *only* simulation hypothesis that might be true of us—the scenario where we are brains-in-vats—is a sceptical hypothesis well beyond the respects that Hanley presents: it is a sceptical hypothesis regarding the existence of *other minds*, and consequently, a sceptical hypothesis regarding our *ethical duties* to others.

Now consider panqualityism. If panqualityism is true, then veridical perception would intuitively have to go roughly as follows. The physical world around us is suffused with *fundamental qualities*, such as colours. Grass, on this account, does not merely give off wavelengths in the 'green spectrum'. If panqualityism is true, then grass itself has the *quality* of being green, where this is something more than 'surface reflective properties' or wavelengths of light: it is literally *being green*. For the panqualityist, then, I perceive the world around me veridically if and only if I *perceive those actual qualities*: the greenness of the grass around me, or the redness of the apple I am about to eat. Now, however, suppose that for the reasons given in Section 2, substrate independence is false for mental entities. Suppose, that is, that phenomenal consciousness depends in some deep way on analogue features of fundamental physics that a digital simulation would not instantiate. If this is the case, then by parity of reasoning, if panqualityism is true, we have grounds for doubting the substrate independence of *non-mental* qualities such as the greenness of grass. It may well be that only *actual* grass has that quality, whereas digitally-simulated grass has no such quality. In that case, if I were a brain in a vat hooked up to a simulation, then I might believe that the grass around me is qualitatively green. Yet, I would simply be wrong. The grass would not be green at all. Instead, the simulation would be merely sending my brain digital information giving rise to experience *as if* the digital grass around me is green when it is not. Notice that this is not unlike the situation at the end of *The Matrix*, where the protagonist 'Neo' is finally able to see past the façade of the simulation, perceiving that everything around him is just computer code his brain is fed, no more and no less. If panqualityism is true, then before

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

coming to this realization, Neo was deceived: he believed that the grass around him in the Matrix was in fact green, and his belief was false. Thus, if panqualityism is true and our argument in Section 2 is correct, then the only kind of simulation hypothesis that we might live in—as brains-in-vats—is a sceptical scenario where we are *systematically deceived* about qualities of the world around us.

4. Replies to Objections

We anticipate two primary objections. The first is that for all we know we may live in a *digital* simulation, and every argument for panpsychism and panqualityism has been based upon our lives as digital agents—in which case digital simulations may well reproduce phenomenal consciousness or phenomenal qualities. A final objection is that if panqualityism is true, then there is still an externalist sense in which brains-in-vats perceive the world around them veridically (*contra* our argument in Section 3). We now elaborate upon and respond to both objections.

Interestingly, it appears to be an open question whether our world is analogue or digital. Quantum mechanics and relativity are currently our best theories of fundamental physics, having been confirmed in every experimental test to date. However, they appear to be incompatible. Whereas relativity is a classical theory according to which space-time and gravity are continuous variables, quantum mechanics holds that fundamental particles and forces are *quantized*, or discrete [Wigner 1979]. This suggests that one theory or the other may have to be false. Consequently, physicists are seeking ways to reconcile the theories. One theory is that 'gravitons' may be a fundamental physical particle that make gravity (and by extension space-time) discrete, 'digital' entities [Witten 1993]. A second theory, the holographic principle, holds that spacetime may be an emergent property from quantized information on the Universe's cosmological horizon, thus making the 'physical' world around us a holographic, digital projection [Susskind 1995]. However, if our world *is* digital,

then our argument would seem to fail at the first step. Existing arguments for panpsychism and panqualityism would provide no grounds for believing that digital simulations would fail to give rise to phenomenally conscious experiences like ours or a fail to generate a world full of qualities. Instead, we might be digital agents whose phenomenally conscious experiences *just are* digital information.

We believe that if contemporary physics discovered that our world is digital, then that might undermine our argument. However, we demur at two points. First, physics has not found any such thing yet. There is not only currently no experimental evidence for gravitons [Rothman and Boughn 2006] or the holographic principle [European Space Agency 2011]. It is also unclear what might confirm the holographic principle [Motl 2012] and whether gravitons could be confirmed with any physically-realistic experiment [Rothman and Boughn 2006]. Second, even if science found that the 'world' around us is digital, it does not follow that *consciousness* is digital. Instead, arguments for panpsychism (*vis-à-vis* the irreducibility of phenomenal properties to functional ones) may show that consciousness is essentially analogue, and we may be hooked up to a *digital simulation* that makes the world appear as a digital projection [Arvan 2013]—in which case substrate independence may still fail to hold. Our point, again, is that we simply do not currently have enough evidence to determine these matters either way. Substrate independence could well be the case. However, at least at present (and perhaps forever, if we are not in an epistemic position to ever settle these matters), we have multiple grounds for doubting whether it is true, and hence the new grounds for doubting simulation arguments that we defend in Section 2.

One final objection is that even if panqualityism is true, it may be argued (in defence of Chalmers' Metaphysical Hypothesis) that there would still be an *externalist* sense in which our beliefs as brains-in-vats would be veridical. To see how, consider what 'red' would *mean* if we were brains-in-vats. Every use of the term 'red' would refer to things in the simulated

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

world around us—to, for example, simulated red apples. In that case, Chalmers might argue that even if digitally simulated apples do not instantiate the quality of redness (since, again, if panqualityism is true, redness may be a primitive analogue quality), our beliefs that apples are red would still be true. This is because in our language within the simulation, 'red' would merely refer to the digitally-rendered surface-reflective properties of simulated apples, not any (analogue) qualities that they may have (or not have). Similarly, a semantic externalist might argue that even if none of the people or animals around us *actually* have phenomenal consciousness, our words in the simulation—such as 'pain', 'joy', and 'phenomenal consciousness'—have only been used throughout our lives in the simulation to refer to *digital* features of other beings. In that case, it will turn out to be true in our (simulated-world) language that other beings experience pain, joy, and other phenomenally conscious states, and that simulated grass is green—in which case Chalmers would still be right: the Matrix is a metaphysical hypothesis, not an epistemic one.

We believe that this line of argument reveals an under-recognized and important implication of semantic externalism. As we see in the above example and in Putnam's original Twin Earth case, semantic externalism holds that the meaning of words depends on the surrounding environment within which they are used [Putnam 1974]. As Putnam puts it, prior to the year 1750, me and my twin on Twin Earth might think that 'water' means the same thing on our respective planets. However, once Earth scientists discover that water is H2O and Twin-Earth scientists discover the stuff they call 'water' is XYZ, we will have discovered that our words meant different things all along (that 'water_{E}' on Earth always meant H2O and 'water_{TE}' on Twin Earth meant XYZ). Bearing this is mind, suppose that you and I are analogue agents possessing phenomenally conscious experiences (that is, experiences of the *quality of redness*) hooked up as brains-in-vats to a digital world in which *no one else* has phenomenal experiences (as merely digital emulations of people). In that

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

case, when you and I use the word 'red', we will plausibly use it (at least in part) to refer to our own phenomenally conscious experiences. 'Redness', we say, is *that* (the experience I have when looking at a red apple—note: we discuss scepticism about private linguistic reference below). Notice, next, that when you and I use the term 'red', we will also presumably assume that others have *that* too (experience of phenomenal redness). However, if panpsychism is true and digital beings do not have phenomenally conscious experiences at all, then this assumption is false. Other 'people' around you and I will not experience *that* (phenomenal redness) at all. When you and I apply the word 'red' to other people (as in 'Jones sees the red sign'), we will (unbeknownst to us) merely use it to refer to *digital simulacra* of phenomenal experiences, much as Twin-Me merely uses 'water', unbeknownst to him, to refer to XYZ. By a similar token, if panqualityism is true, then when you and I use the word 'green', we may assume on the basis of our conscious experience (of phenomenal greenness) that it refers to *genuine qualities* in the world around us—that is, to the greenness of grass. However, if we are brains-in-vats hooked up to a digital simulation and our argument in Section 2 is sound, then that assumption will be false: the digitized grass around us is *not qualitatively green.*

Here is why this is important. If panpsychism or panqualityism is true, then many of our words—specifically, those referring to qualities (redness, greenness, loudness, etc.)—will have *multiple* semantic interpretations. On the one hand, when you and I use the term 'red' in a digital simulation, there is sense in which the word will refer to purely structural features of the simulation: specifically, to the way in which 'apples' in the simulation 'give off red wavelengths', and so on. Relative to this structural way of interpreting quality-denoting words, all of our ordinary-language statements about things in a simulation may be true: it may be entirely right to say that there are red apples, green grass, and so on—since, on this interpretation, all it is to *be* a red apple is to be a certain kind of digitized object. However, if

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

panpsychism or panqualityism is true, then our quality-denoting words will also intuitively have *another* semantic interpretation—one where these words refer to non-structural *qualities* of our phenomenal experience ('Redness looks *like this*') and/or real, metaphysically primitive qualities of objects ('*That* is the greenness of grass'). On this interpretation, if you and I are brains-in-vats, then most of our ordinarily language statements—such as, 'Other people experience what red is like' and 'Grass is green'—may turn out to be false.

The point here then is this: if panpsychism or panqualityism is true, then all of our words referring to qualities will plausibly have two readings or semantic interpretations. First, they will have a *sceptical reading*. If you and I are brains-in-vats hooked up to a simulation where no one else experiences phenomenal redness, then there is a clear sense in which 'other people experience red' will be *false*. On the other hand, these same words will also have a *metaphysical* reading according to which the same sentences are true. For again, even if no one else does experience phenomenal redness, you and I will have also used 'red' our entire lives to *refer* to purely structural features of the simulation (such as the manner in which digitized beings' 'retinas' detect 'red' wavelengths).

If this is right, then Chalmers is half-right and half-wrong. The Matrix Hypothesis is both a metaphysical hypothesis *and* a sceptical one—and which of the two that it is depends entirely upon which semantic interpretation of our words we adopt, that is, upon which *propositions* we take our words to express: propositions referring merely to structural features of the world, or propositions referring to phenomenal and/or non-phenomenal qualities. Yet, this sounds exactly right to us. If you and I are hooked up to the Matrix and everyone and everything around is merely simulated, there is a *purely structural sense* in which other people exist (as digital people) and digitized objects 'have colours'. However, there is also a clear *qualitative* sense in which you and I are deeply and systematically deceived about the

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

world around us—since again, the 'people' and things around us may not have any of the *qualities* we thought they did.

One final worry is that in making this argument, we have run afoul of Wittgenstein's private language argument—Wittgenstein's argument that there is no sense in which 'red' can mean *this*, where 'this' is a merely private phenomenal state [Wittgenstein 1958: 91^{e}-102, sec. 256–307]. Although we cannot offer a full refutation of Wittgenstein's argument here, we want to note Wittgenstein's argument implicitly relies on an *internalist* assumption: namely, that in order for 'red' to mean a private experience, we would need to be able to tell 'from the inside' whether we are following the same rule over time. Wittgenstein writes:

> I want to keep a diary about the recurrence of a certain sensation. To this end I associate it with the sign "S" and write this sign in a calendar for every day on which I have the sensation…But what is this ceremony for?...A definition surely serves to establish the meaning of a sign.—Well, that is done precisely by the concentrating of my attention; for in this way I impress on myself the connexion between the sign and the sensation.—But "I impress it on myself" can only mean: this process brings it about that I remember the connexion *right* in the future. But in the present case I have no criterion of correctness. One would like to say: whatever is going to seem right to me is right. And that only means that here we can't talk about 'right' [Ibid.: 92*, sec. 258].

However, we believe that if semantic externalism is correct, then Wittgenstein's argument here is a *non sequitur*. For externalists, the meaning of a word is *not* a definition that we can reliably introspect 'from the inside.' After all, from the inside before 1750, neither me nor Twin Me could know whether 'water' refers to H2O or to XYZ. For the externalist, the meaning of a term is not a matter of whether we can *introspect* whether we are following a rule consistently, but rather whether—as a matter of fact—we *are* following a rule

The Version of Record of this manuscript has been published and is available in *Australasian Journal of Philosophy,* Published online 02 May 2021, http://www.tandfonline.com/doi/full/10.1080/00048402.2021.1913621

consistently. Nothing in Wittgenstein's argument rules this out. I may not have a purely private 'criterion of correctness' that I can ever check to make sure that I *am* following the same rule over time (such as using 'red' to refer to *this* over and over again, where *this* really is a private experience of redness). But, for all that, I may *in fact* use 'red' to refer to the same thing over time—in much the same way that (unbeknownst to them from the inside), Twin Earthlings were always calling *one* thing 'water' (XYZ) rather than, say, multiple things (both H2O *and* XYZ). We conclude not that Wittgenstein's private language argument is unsound, but rather that *if* semantic externalism is true, then his argument does not go through. Because the second objection to our argument above was based on semantic externalism—and because we have shown that if externalism is true, words in the Matrix simultaneously have metaphysical and epistemic meanings—we contend that our response succeeds: the Matrix hypothesis is, *contra* Chalmers, *both* a metaphysical and epistemic hypothesis, on different ways of disambiguating word-meaning.

5. Conclusion

We have argued that the nature of consciousness has far-reaching implications for simulation arguments. There are influential arguments in the literature for panpsychism and panqualityism. If either theory is true, then Bostrom's substrate independence assumption may be false. If substrate independence is false, then Bostrom's Simulation Argument is unsound. For all he has shown, there is only one way to live in a simulation: namely, as a brain-in-a-vat. But, if this is true, then panpsychism and panqualityism also call into question arguments that some version of the simulation hypothesis is a true explanation of fundamental physics—as there are significant reasons to doubt whether members of PHCs are likely to hook up many people to simulated worlds with physics like our own. Finally, if

panpsychism or panqualityism is true, then the simulation (or Matrix) hypothesis is a genuinely sceptical scenario calling into doubt much of what we think that we know.[1]

The University of Tampa

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No funding was reported by the author(s).

[1] Thanks to Stephen Hetherington and two reviewers at this journal for helpful comments on an earlier version of this manuscript. Many thanks also to Corey Maley for discussions clarifying the digital-analogue distinction.

REFERENCES

Arvan, Marcus 2013. A New Theory of Free Will, *Philosophical Forum* 44/1: 1–48.

Arvan, Marcus 2014. A Unified Explanation of Quantum Phenomena? The Case for the Peer-to-Peer Simulation Hypothesis as an Interdisciplinary Research Program, *Philosophical Forum* 45/4: 433–46.

Arvan, Marcus and Corey J. Maley In preparation. If Panpsychism, then Digital AI may be Phenomenally Scrambled.

Birch, Jonathan 2013. On the 'Simulation Argument' and Selective Scepticism, *Erkenntnis* 78/1: 95–107.

Bostrom, Nick 2003. Are We Living in a Computer Simulation? *The Philosophical Quarterly* 53/211: 243–55.

Bostrom, Nick 2013. Existential Risk Prevention as Global Priority, *Global Policy* 4/1: 15–31.

Brueckner, Anthony 2008. The Simulation Argument Again, *Analysis* 68/3: 224–6.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.

Chalmers, David J. 2005. The Matrix as Metaphysics, in *Philosophers Explore the Matrix*, ed. Christopher Grau, New York: Oxford University Press: 132–76.

Chalmers, David J. 2013. Panpsychism and Panprotopsychism, in *Panpsychism: Contemporary Perspectives*, ed. Godehard Brüntrup and Ludwig Jaskolla, New York: Oxford University Press: 19–47.

Chalmers, David J. 2017. The Virtual and the Real, *Disputatio* 9/46: 309–52.

Chiang, Chia-Chu, Rajat S. Shivacharan, Xile Wei, Luis E. Gonzales-Reyes, and Dominique M. Durand 2018. Slow Periodic Activity in the Longitudinal Hippocampal Slice Can

Self-Propagate Non-Synaptically by a Mechanism Consistent with Ephatic Coupling, *The Journal of Physiology* 597/1: 249–69.

European Space Agency 2011. Integral Challenges Physics beyond Einstein, URL = <http://www.esa.int/Science_Exploration/Space_Science/Integral_challenges_physics_beyond_Einstein>, retrieved 12 March 2021.

Goff, Philip 2017. *Consciousness and Fundamental Reality*, New York: Oxford University Press.

Hanley, Richard 2017. Skepticism Revisited: Chalmers on *The Matrix* and Brains-in-Vats, *Cognitive Systems Research* 41/March 2017: 93–8.

Hardesty, Larry 2016. Analog Computing Returns, *MIT News*, URL = <https://news.mit.edu/2016/analog-computing-organs-organisms-0620>, retrieved 12 March 2021.

Huemer, Michael 2016. Serious Theories and Skeptical Theories: Why you are Probably Not a Brain in a Vat, *Philosophical Studies* 173/4: 1031–52.

Koch, Christof 2019. *The Feeling of Life Itself: Why Consciousness is Widespread but cannot be Computed*, Cambridge, MA: MIT Press.

Maley, Corey J. 2011. Analog and Digital, Continuous and Discrete, *Philosophical Studies* 155/1: 117-31.

Maley, Corey J. 2018a. Brains as Analog Computers, *Medium*, URL = <https://medium.com/the-spike/brains-as-analog-computers-fa297021f935>, retrieved on 12 March 2021.

Maley, Corey J. 2018b. Toward Analog Neural Computation, *Minds and Machines* 28/1: 77–91.

Maley, Corey J. 2020. Continuous Neural Spikes and Information Theory, *Review of Philosophy and Psychology* 11/3: 647–67.

McGinn, Colin 1989. Can We Solve the Mind-Body Problem? *Mind* 98/391: 349–66.

Mizrahi, Moti 2017. The Fine-Tuning Argument and the Simulation Hypothesis, *Think* 16/46: 93–102.

Motl, Luboš 2012. Hogan's Holographic Noise Doesn't Exist, *The Reference Frame*, URL = <https://motls.blogspot.com/2012/02/hogans-holographic-noise-doesnt-exist.html>, retrieved 21 March 2021.

Putnam, Hilary 1974. Meaning and Reference, *The Journal of Philosophy* 70/19: 699–711.

Rothman, Tony and Stephen Boughn 2006. Can Gravitons be Detected? *Foundations of Physics* 36/12: 1801–25.

Russell, Bertrand 1921. *The Analysis of Mind*, New York: The Macmillan Company.

Shepherd, Joshua 2018. *Consciousness and Moral Status*, New York: Routledge.

Stoljar, Daniel 2020. Chalmers v Chalmers, *Noûs* 54/2: 469–87.

Strawson, Galen 2006. Realistic Monism: Why Physicalism Entails Panpsychism, *Journal of Consciousness Studies* 13/10-11: 3–31.

Strawson, Galen 2016. Mind and Being: The Primacy of Panpsychism, in *Panpsychism: Contemporary Perspectives*, ed. Godehard Brüntrup and Ludwig Jaskolla, New York: Oxford University Press: 75–112.

Susskind, Leonard 1995. The World as a Hologram, *Journal of Mathematical Physics* 36/11: 6377–96.

Weatherson, Brian 2003. Are you a Sim? *Philosophical Quarterly* 53/212: 425–31.

Whitworth, Brian 2008. The Physical World as a Virtual Reality. *arXiv preprint* arXiv:0801.0337.

Whitworth, Brian 2010. The Emergence of the Physical World from Information Processing. *Quantum Biosystems* 2/1: 221–49.

Wigner, Eugene Paul 1979. The Basic Conflict Between the Concepts of General Relativity and of Quantum Mechanics, in *The Collected Works of Eugene Paul Wigner - Part I: Particles and Fields, Part II: Foundations of Quantum Mechanics*, ed. Arthur Wightman, Berlin, Heidelberg: Springer-Verlag, 1997: 350.

Wittgenstein, Ludwig 1958. *Philosophical Investigations*, trans. G. E. M. Anscombe, Oxford: Basil Blackwell.

Witten, Edward 1993. Quantum Background Independence in String Theory. *arXiv preprint hep-th/9306122*.