DOI: 10.1111/phpr.13122

Philosophy and Phenomenological Research

SYMPOSIUM

Taking the simulation hypothesis seriously

David J. Chalmers

Department of Philosophy, New York University

Correspondence
David J. Chalmers, Department of Philosophy, New York University.
Email: chalmers@nyu.edu

Much of *Reality+* focuses on the simulation hypothesis: the thesis that we are living in a computer simulation. I argue that we should take the simulation hypothesis seriously, and that we cannot rule it out. I also argue that the simulation hypothesis is not a skeptical hypothesis where most of our beliefs are false. If we are in a perfect simulation, most of our beliefs are true. As a result, the simulation hypothesis does not lead to skepticism, and life in a simulation can be roughly as good as life in a non-simulated world.

All three commentators in this symposium focuses on broadly epistemological issues about the simulation hypothesis, often with issues about skepticism as well as issues about value in the background. Peter Godfrey-Smith argues that we should not take the simulation hypothesis seriously. Susan Schneider and Eric Schwitzgebel argue that while the perfect simulation hypothesis may not be a skeptical hypotheses, other versions of the hypothesis may be. As a result, they suggest that life in a simulation may not be as good as life in non-simulated reality, and that the simulation hypothesis may still lead to a degree of skepticism.

(See also a recent symposium on *Reality+* in *Oxford Studies in the Philosophy of Mind*, in which Terry Horgan, Christopher Peacocke, and Grace Helton all discuss epistemological about skepticism as well as metaphysical and value-theoretic issues.)

1 | IS THE SIMULATION HYPOTHESIS A SERIOUS HYPOTHESIS? (GODFREY-SMITH)

Godfrey-Smith takes the simulation hypothesis seriously enough to offer a number of reasons not to take it seriously. More specifically, he offers a number of reasons for denying that it's probable that there are many humanlike sims. Two reasons, offered very briefly, are tied to the feasibility of biosims and the motivations of simulators. Two other reasons, developed more extensively, use considerations about perfect simulation and substrate-neutrality to question the feasibility of humanlike brain simulations. Godfrey-Smith also draws a parallel with the Boltzmann brain hypothesis.

1. *Biosims*. Godfrey-Smith's first objection to the simulation idea is tied specifically to the biosim scenario in which we are biological systems hooked up to a simulation. Godfrey-Smith suggests that some experiences may be hard to simulate this way, mentioning hot showers and skateboard crashes. Perhaps the thought is that the body plays a special role in these experiences, and it affects the brain in a way that goes beyond standard sensory inputs and outputs. This thought might be fleshed out by asking: to undergo these experiences, does the brain have to be warmed up and jolted around, or is it enough for it to represent warmth and motion?

These are nontrivial questions that I suspect an advanced biosimulator could answer in multiple ways. But as Godfrey-Smith does, I'll set biosims aside to focus on the pure simulation hypothesis on which brains are part of the simulation. It's the pure simulation hypothesis that the Bostrom-style statistical argument tends to support, since pure sims are likely easier to construct in large numbers than biosims. Once we focus on the pure simulation hypothesis, the objection tied specifically to biosims falls away. For a feasible pure simulation in these scenarios, we just need a simulation of a warm brain in a hot shower, or of a brain being jolted around in a skateboard crash.

2 *Simulator motives*. Godfrey-Smith raises the question of motive: why would simulators create many cosmic simulations? There's an obvious answer: for the same sort of reasons that people create more mundane simulations today: science, engineering, prediction, and more. For example, if it becomes feasible to simulate whole universes at limited expense, then there will be strong scientific motive to do so, for example simulating many different universes with different laws will help understand the space of possible universes. Of course this motive could be defeated by other reasons: reasons of ethics, expense, or disinterest, for example. But the benefits of simulation already take us beyond "Who knows why they would do it?" to a positive reason grounded in reality.

In any case, the standard simulation argument is designed to accommodate our uncertainty about these matters. The conclusion is disjunctive. In Bostrom's version, it's roughly: we are probably sims *or* most nonsims won't choose to create sims *or* most nonsims will die first. In my version, the conclusion is "Either there are sim blockers or we are probably sims", with a list of sim blockers that includes Bostrom's disjuncts along with some others. Godfrey-Smith isn't rejecting these conclusions: he's simply focusing on one of the other disjuncts/sim blockers (the one about choice) and noting that it can't be excluded. Perhaps he thinks that given the disjunction, we should invest much more of our credence in the choice disjunct than in the simulation disjunct, but he hasn't really given us reason to do so here.

3 *Perfect simulations*. Godfrey-Smith's most serious objections to the simulation hypothesis concern the simulation of human brains: he suggests that the idea of a perfect brain simulation is problematic, and also that brain simulations will not replicate what is crucial for consciousness.

He illustrates the point about perfect simulation by asking: how much fidelity of timing must be present in a perfect duplicate? I suppose a truly *perfect* simulation would demand perfect fidelity, which would plausibly demand simulating a system down to the level of fundamental physics. Godfrey-Smith doesn't address this sort of physics simulation, but various questions could be raised for it: for example, is it possible to simulate continuous systems with infinite precision, and is it feasible to simulate brains down to the level of fundamental physics?

In any case, my arguments about the simulation hypothesis do not depend on simulations that are perfect in this sense. What matters is that simulations be *humanlike*. This does not requires perfectly simulating any existing human being. It just requires simulating a way that a human being might have been. This can plausibly be done without simulating every detail of a brain to infinite precision. One way to see this is to note that brain mechanisms themselves are subject to biological noise and so cannot rely on infinite precision for humanlike behavior.

Take Godfrey-Smith's case of timing, in which two neurons fire synchronously. How much does exact synchrony matter? Does it make a difference if one neuron fires a millisecond or a femtosecond later? We know that due to noise, the brain can never rely on exact timing and exact synchrony. Any neuron could easily have fired a femtosecond later than it actually did, due to noise. The result might have been somewhat difference from the actual result, but it would still have been reasonable behavior, and humanlike in the human case. This suggests that to replicate humanlike behavior, we need only replicate timing down to the level of precision permitted by biological noise, perhaps simulating the noise distribution and its effects if that turns out to be important. In any case one we recognize that perfect simulation is not required for the purposes of the simulation hypothesis, Godfrey Smith's worries about that notion fall away.

4 *Substrate-neutrality*. Perhaps Godfrey-Smith's largest concern about the simulation hypothesis concerns consciousness. He speculates that "the basis for felt experience" may be specifically biological and may not be replicated in a simulation. The specific view Godfrey-Smith likes gives a key role to the dynamics of electrical oscillations in neurons, involving ions that move across cell membranes. Godfrey-Smith does not deny that these processes could be simulated. Instead he denies that the simulation would have the relevant biological or chemical properties (such as electrical oscillations in neurons) that are required for consciousness: electrical oscillations in neurons, perhaps.

Now, the mere fact that electrical oscillations in neurons are the basis for consciousness in humans does not dictate that they are required for consciousness in all systems. A functionalist can reasonably hold that it's the structural dynamics of these oscillations that matter, rather than their realization, and that this structural dynamics could be present in a simulation. Godfrey-Smith suggests that a simulation might represent this dynamics but it would not realize it. But I think this misses the fact that genuine dynamics is present in computer simulation, with parts of a computer system affecting other parts in genuine patterns of oscillation. This dynamics will not just represent but replicate the structural dynamics among neurons, at least at a certain level of abstraction.

Around here the real disagreement is which level of abstraction in the dynamics matters for consciousness. I've used neuron replacement "fadng qualia" arguments (changing low-level processes while preserving high-level dynamics) to argued that low-level processes are unlikely to be required for consciousness. Godfrey-Smith is skeptical that this sort of replacement is possible without affecting behavior. Again, a perfect replication of behavior is not required for the argument to work. It suffices to have have humanlike behavior, without huge anomalies such as people reporting that their experiences are fading out. I suspect that neuron-replacement of this sort is possible, and may even become actual in coming decades. At that point we'll have empirical evidence to use in adjudicating these arguments.

To step back: the debate between biological views and functionalist views of consciousness, and the corresponding debate about whether simulations can be conscious, is likely to go on. I favor a functionalist view, while allowing that biological views are serious possibilities. Godfrey-Smith

favors a biological view, but he calls this "speculative" and presumably allows that functionalist views remain serious possibilities. Once we grant that both these views are serious possibilities, it must also be recognized as a serious possibility that simulations can be conscious, and this line of reasoning (at least in the absence of other sim blockers) cannot be used to dismiss the simulation hypothesis.

To step back: the reasons that Godfrey-Smith gives for not taking the simulation hypothesis seriously come down to two sim blockers: (1) *Conscious humanlike sims are impossible* (deriving from worries about brain simulation), and perhaps (2) *Nonsims will not create many conscious humanlike sims* (deriving from worries about simulator motives). I think these sim blockers are less likely than Godfrey-Smith does. But even he should allow that the denials of (1) and (2) are both serious and grounded hypotheses that deserve some substantial nonzero credence. If we give these denials even 10% credence each, then (assuming independence) this still leaves us with a 1% credence that most humanlike beings are sims. That would certainly be reason to take the simulation hypothesis seriously.

5 *Boltzmann brains*. Boltzmann brains are shortlived humanlike brains that some physical theories predict will form infinitely many times through random processes in an infinite space. These raise obvious skeptical issues: isn't it likely that we are Boltzmann brains, and that most of what we believe is false? In my brief discussion in *Reality+*, I follow Sean Carroll in holding that the belief that we are Boltzmann brains is cognitively unstable: if it's true, I cannot stably endorse it. If I accept that I'm a Boltzmann brains, I should reject my perception of the external world as illusory and my scientific beliefs as incorrect, which would eliminate the scientific case for believing I'm a Boltzmann brain in the first place. This diagnosis of instability is far from a full response to the complex skeptical worries raised by Boltzmann brains, but undermining the support for the hypothesis is at least a first step in defusing these worries.

Godfrey-Smith thinks there is a parallel between Boltzmann brains and the simulation hypothesis. He says that if the Boltzmann brain hypothesis is cognitively unstable, then so is the simulation hypothesis. Taking either of them seriously should lead us to doubt our own reasoning in general, and therefore to doubt the reasoning that led us to these hypotheses.

I don't think there is parity between the Boltzmann brain hypothesis and the simulation hypothesis here. Once one accepts simulation realism, as I do, then the simulation hypothesis (at least in its usual forms) should not lead us to reject our perceptions of the external world or our scientific theories, and so it should not lead us to rejecting the case for the simulation hypothesis. In paradigmatic simulations, unlike Boltzmann brains, our reasoning and most of our beliefs are still good guides to reality.

Godfrey-Smith says "To seriously suspect you are in a simulation is to suspect that many or all of your memory traces might be cooked up, and your background knowledge is no good at all". Of course I reject this claim. Once one accepts simulation realism, one can suspect that one is in a simulation without doubting one's memory or background knowledge. Now, perhaps Godfrey-Smith rejects simulation realism (though he says at the start that he won't worry about this issue, and he doesn't give reasons for rejection). Or perhaps he has in mind simulations where only some aspects of the world are simulated and many of our beliefs are false. In *Reality+*, I argue that even in these simulations, many of our everyday beliefs are fine. So there is still not parity with the self-undermining Boltzmann brain hypothesis. Still, simulations of this sort rise many issues of interest, some of which I discuss in the following sections.

2 | MIGHT WE BE IN A DECEPTIVE SIMULATION? (SCHNEIDER AND SCHWITZGEBEL)

In *Reality+*, I argue that the perfect simulation hypothesis, on which a whole universe is simulated with great fidelity, need not be a skeptical hypothesis. If we are in a perfect simulation, we are not deceived about most everyday matters (or at least, we are not deceived simply because we are in a simulation).

Schneider and Schwitzgebel both accept this sort of simulation realism, at least for the purposes of argument. But they both argue that there are *other* versions of the simulation hypothesis that I cannot exclude, and that should be regarded as skeptical hypotheses: if those hypotheses are true, most of our everyday beliefs are false.

Schneider focuses especially on *deceptive simulations*, devised by superintelligent AIs partly in order to deceive us. Schwitzgebel focuses on *local simulations* (or *small simulations*), which just simulate a very small part of the universe such as a single city. They argue that many or most simulations may be deceptive simulations or local simulations, and that if we are in a simulation of this sort, many or most of our everyday beliefs are false.

Now, in *Reality+*, I am mainly concerned to argue against *global skepticism*, on which we can't rule out Cartesian scenarios in which we are radically deceived about almost everything. Various more local forms of skepticism, turning on scenarios in which we are deceived about some things but not everything, are left on the table. The final chapter of the book focuses on residual skeptical hypotheses from local simulations to Boltzmann brains, and draws the overall verdict "[W]hile there are some important skeptical challenges here, none lead to global skepticism."

Schneider and Schwitzgebel's claims are compatible with the general line in *Reality+*. They don't argue for global skepticism. Instead, they argue that simulation scenarios should lead us to take more local skeptical hypotheses seriously. I could simply agree with this, but I think that these local skeptical scenarios are not quite as straightforward as Schneider and Schwitzgebel suggest, so I will take a closer look at them.

3 | SMALL SIMULATIONS (SCHWITZGEBEL)

Schwitzgebel asks:

> How confident ought we to be that if we inhabit a virtual reality the reality is large enough to be epistemically non-catastrophic – that the world contains more or less all of the things we care about, plus a reasonably deep past, plus a reasonably long future, and billions of people? Call this the Size Question. An optimist about the Size Question holds that we ought to be confident that if we are sims, we don't live in a catastrophically small simulation. The pessimist denies this.

Let's call Schwitzgebel's "epistemically non-catastrophic" simulations *large* simulations, and call the contrast class *small* simulations. Large simulations needn't simulate the whole universe, but they need to simulate most of the earth and its people, and all of the people and places we care about. In small simulations, much of what we care about isn't simulated and doesn't exist.

There's an immediate question about small simulations so defined. For people and places we care about, we typically have detailed memories of them, and other detailed records, such as

photos and text about them. In a small simulation, where do these memories and records come from? Even fake memories have to be generated somehow. The obvious answer is that they are generated by a simulation of those people and places. But if all of these are simulated, then our small simulation is now a large simulation.

Now, the small simulation theorist could say that the memories and records come from some other source—maybe from stock images and videos with actors or deepfake technology, for example? But it's not at all easy to see how use of these things without simulation could generate the vivid and detailed memories that we might have of our family and close friends. Perhaps there could be a multi-tiered approach where family and close friends are simulated in detail but where others (e.g. family members' close friends) are simulated in less detail, and people further out are just stock images and statistics. But it remains unclear just how interactions between, say, full-blown family members and the sketchily simulated friends will work.

Alternatively, a small simulation theorist could stipulate a scenario where most of these memories and records don't exist. Such a simulation might include the case where I'm just waking up and haven't begun thinking about these things, with just a vague sense that I've lived a life and have a world out there. Perhaps it will terminate the moment I go through memories or contact the outside world. We might call these *tiny simulations*. Perhaps we sometimes can't rule these out. But this method obviously won't scale to handle even a single day where I entertain many memories, read many texts, and interact with many people. For those purposes, we'll need many simulated people, who have memories of their own, and we'll be back toward the scenario above.

In *Reality+*, I suggest that it's hard to find a natural stopping-point between tiny simulations and large simulations such as simulations of the whole earth, in part because of all the interaction between people across the earth. Perhaps in historical eras we could simulate a fairly isolated population, but modern technology for travel and communication makes this much harder.

Schwitzgebel doesn't directly address this concern. He does suggest that it ought to be possible to simulate a whole city without simulating what's beyond:

> Might the simulation contain only one city? Stipulate that the city has existed for at least a hundred years, but nothing beyond it exists. Everyone in the city exists, and we have real conversations with each other. The room you are in exists, and the building, and the roads–but everything stops at the city edge. Anyone looking beyond the edge sees, presumably, some false screen. If they travel past the edge, they disappear from existence; and when they return, they pop back into existence with false memories of having been elsewhere. News from afar is all fake. Unless you grew up in the same city, your childhood is fake.

This passage leaves most of the key questions unanswered. Let's say I return to New York after a trip to San Francisco. How are my false memories of San Francisco generated? My detailed memories will require a detailed model of San Francisco and my travel though it. The same goes for most places on Earth, at least in the current era of widespread travel. Communications technology make things all the harder. Detailed and ongoing news reports of faraway places, and even webcams and the like, will require very detailed simulation. What happens when New Yorkers interact by phone or videoconference with others all over the world? Those others will presumably need to be simulated. All this suggests that large-scale simulation of the world outside the city will be needed.

Let's consider a case. Suppose I remember a long converation with Eric over Zoom yesterday, when I was apparently in New York and he was in California. How did this memory get

generated, on Schwitzgebel's picture where only New York is simulated? I was in New York, so my end of the conversation really happened. Officially the Eric simulation has gone out of existence (at least since the last time he was in New York). But some sort of sophisticated technology will be needed to generate Eric's end of the conversation. I'd think that something close to a full-blown Eric simulation would be needed. A low-grade model would be pretty easily detectable. But if a full-blown Eric simulation is needed, then a full-blown simulation will also be needed for most of the people with whom I remember having had rich interactions, and for most of the people other New Yorkers have interacted with. That will rapidly move us toward a full-blown simulation of billions of people.

Furthermore, these simulated people outside the city will report memories of interactions with many others, and those others will also frequently show up on video and the like, which will require further simulations in turn. At one point Schwitzgebel notes that standard social science models treat people as aggregates—but that doesn't work nearly as well once we have specific experiences and memories to deal with.

Of course simulators might try using advanced AI such as large language models and their extensions to stand in for people. Schwitzgebel himself has constructed a simulation of Daniel Dennett that some people couldn't differentiate from the original. This test used only small text samples in special contexts, but perhaps a full-blown multimodal language model of Dennett that's indistinguishable from the original might be built eventually. But these AI systems are themselves high-grade simulations of a sort (even if they simulate behavior more directly than they simulate brains), so they don't escape the need for extensive simulation.

It might be argue that these AI systems will not be conscious, or will lack other features of genuine minds, so that at least our beliefs about other minds are wrong when these systems are used. Grace Helton takes a line like this (arguing that the simulation hypothesis leads to solipsism) in a recent symposium piece on *Reality+*, and I reply there. It's far from clear that convincing human behavior without consciousness is possible, but even if it is, this casts more doubts on knowledge of other minds than on the physical world.

Another strategy invokes *mind control*, where simulators directly manipulate our minds. They can simply adjust our perception, our reasoning, our memories, or our actions whenever needed so we don't notice problems such as low-grade behavior. Mind control may help enable smaller simulations with the use of low-grade models and simulations, although a robust and high-grade city simulation (of the sort Schwitzgebel posits) will still require fairly detailed if lower-grade simulation of the outside world.

In another version of mind control, akin to Nozick's experience machine (and discussed in an online appendix to *Reality+*), I'm simply living out a script of my life in New York, perhaps with associated immersive movies. On this version no simulation is needed, but my actions and choices will have to be tightly controlled by the simulators to ensure that I do not veer off-script. Perhaps I can't rule out situations of this sort, but I'd count them as very different from the standard simulation hypothesis: life in a scripted immersive movie is not really life in a simulation.

In any case, I'm inclined to set aside mind-control skepticism as distinct from simulation skepticism. Mind control raises distinctive metacognitive skeptical issues of its own: recall Descartes' demon who tampers with our reasoning to make us doubt that two plus three is five. As I note in Reality+, this sort of metacognitive skepticism is particularly difficult to defeat. But mind control can be implemented with or without simulation, as we've see here. We can construct large, small, or tiny skeptical scenarios using mind control, whether or not we have simulation. At worst, simulation offers an especially natural route to mind control, as code may be easier to manipulate than brains. But it remains the mind control rather than the simulation that's really responsible

for the skeptical worries here. So I'll understand the simulation hypothesis as involving scenarios without mind-control.

Once mind-control scenarios are set aside, I am inclined to divide my credence in the simulation hypothesis mainly between large simulations (simulating much of the earth, at least) and tiny simulations (like simulating me in my room and thinking about philosophy for a few minutes). How do I apportion that credence? It's true that tiny simulations will be much cheaper and easier to set up than large simulations. On the other hand tiny simulations of creatures with few memories are somewhat odd and it's unclear why simulators will be motivated to create them. The scientific and predictive motives for creating large simulations are much clearer.

Perhaps there will be many fewer large simulations than tiny simulations, but that's not nearly enough to yield a low credence in the large simulations. Here the point about relative numbers comes in. A tiny simulation will typically contain just one person, a few at most, while a large simulation will contain billions of people. As a result, even if there are a million times as many tiny simulations as large simulations, there will be a thousand times as many people in large simulations as in tiny simulations. If that's what we expect, we should assign only one-thousandth of our credence in the simulation hypothesis to tiny simulations, and the rest to large simulations.

Schwitzgebel respnds to this point about the numbers by saying "cost considerations loom large", but those cost considerations have already been taken into account at the first stage, in allowing that there will be far more tiny simulations than large simulations. Of course it is hard to know what the right multiple is for "far more", but as long as it's under a billion or so, the billionfold boost for numbers in large simulations will lead to most of our credence going there.

Overall, I'm somewhat inclined to have a fairly high credence in the large simulation hypothesis, conditional on the simulation hypothesis in general. But this is quite tentative and open to revision. Given the tentativeness, I'd probably allow at least a conditional probability of one in ten for the small simulation hypothesis, which Schwitzgebel counts as "substantial". So I certainly wouldn't claim to have refuted the corresponding local forms of skepticism. Still, I'd hesitate to endorse Schwitzgebel's claim that we should hope that we aren't in a simulation.

4 | ILLUSORY SIMULATIONS (SCHNEIDER)

Schneider put forward an "Illusory World" hypothesis, suggests that we might be in an illusory simulation set up by an AI system that is trying to deceive us. She says the simulation might have fake laws, quantum chanciness, misinformation, deepfakes, and problematic biases.

Of course I think that illusions are possible in simulations just as they are possible in ordinary reality. So I think *partly* illusory worlds with numerous illusions are certainly possible. At the same time, I'm inclined to deny that *wholly* illusory worlds, where everything we perceive and believe is illusory, are possible.

It's not obvious whether Schneider is endorsing partly or wholly illusory worlds. If it's the former, then I agree with the conclusion. So to make things interesting, I'll consider her arguments as arguments for wholly illusory worlds and argue that they don't succeed.

Schneider says a number of things about her illusory worlds to support the claim that they are illusory. Perhaps the most central is that the AI *intends* to deceive us. Using the example of the Matrix, she says "A world that is MReal can fail to be as we believe it to be because it is an orchestrated deception."

Now, just because an AI intends to deceive us does not entail that it succeeds. The Matrix case brings this out. The machines may have intended to deceive humans by giving them the illusion

that tables exist. But if I am right, they did no such thing. Rather, humans in the Matrix have largely true beliefs about the world around them. They are not deceived, at least about everyday matters about whether tables exist (though they may be deceived about relatively esoteric matters such as whether tables are digital and so on). My view is that if the machines think they are deceiving us about these matters, they are making a philosophical mistake.

Of course, the machines may be philosophically sophisticated enough to avoid this mistake. But the point brings out that their mere intention to deceive does not entail that the world is illusory. If the world is illusory, it must be for a more substantial reason.

One substantial reason for global illusion is suggested by a recurring theme in Schneider's discussion: the machines may be using a simulation to entirely deceive us about the character of the "outer" world containing the simulation. For example, she suggests that a deceptive simulation may have "fake laws", which I interpret as meaning that its laws are very different from those in the outside world in which the simulation takes place. But again, I don't think that fake laws in this sense entail that people in the simulation are deceived. They may be ignorant of the laws of the outer world, but they need not have false beliefs about them. On my view, their world is the inner world of the simulation, its laws are the "fake" laws, and they may well have true beliefs about their world. Of course if they believed that these fake laws were the laws of the outer world, they would be wrong, but that belief does not seem to be central to being in this simulation. More generally, insofar as sims' beliefs are about the inner world (as I argue they mainly are) and not about the outer world, then the fact that the inner world is unlike the outer world does not entail that the simulation is deceptive.

Schneider also suggests that the simulation is likely to be full of misinformation. Here we can distinguish between *global misinformation*, which is roughly deceiving us about the entire world, and *local misinformation*, which is roughly deceiving us about many matters of particular fact without deceiving us across the board. If I am right, a deceptive simulation is much more likely to involve local misinformation than global misinformation. At least, I don't have a clear picture of how simulators could deceive us globally, while I have a clear picture of how simulators could deceive us locally. This would require just the same sort of mechanisms of misinformation that we find in the ordinary world. Of course local misinformation is still worrying, and if every belief is subject to hypotheses in which it involves misinformation, skepticism still threatens. But if the misinformation is local rather than global, then at least this skepticism will have limits.

5 | SIMULATORS AND CREATORS (SCHNEIDER AND SCHWITZGEBEL)

Schneider and Schwitzgebel also both raise issues about the character of the simulator, arguing that the simulator is likely to have bad character and that the simulation is correspondingly worse then ordinary reality.

Schneider argues that if our world is simulated, the simulator is very likely an AI system. This part seems quite plausible to me. And she argues that this AI system is likely to share many of the problems of existing AI systems: it may be riddled with biases, misinformation, and other sources of deception. The reasoning here is less obvious to me. An AI system that can support a universe simulation will be far more advanced than existing AI systems, and one can reasonably hope that part of this advancement will be progress on the problems of bias and misinformation.

On the other hand, Schneider has observation on her side. We do seem to be in a world riddled with bias and misinformation. All that suggests that if we are created by an AI, it has not done a good job of eliminating bias and misinformation from the world it created.

This connects to Schwitzgebel's worries about the problems in our world. He argues that if our world is simulated, then the simulator is responsible for suffering and evil from the plague to the Holcaust. If so, the simulator cannot be a good simulator. One could make a similar point using Schneider's observations about bias and associated oppression. Arguably, a simulator who would create such a problematic world cannot be a good simulator.

Of course this is a simple twist on the familiar problem of evil for theism: a benevolent god would not permit all this evil, so if there is a god, they are not benevolent. The simulation hypothesis just extends the issues from gods to creators more generally. There is a familiar raft of responses on behalf of a benevolent god. For example, perhaps this world has enough good that the good outweighs the evil, and creating it (perhaps alongside other good worlds) is better on balance than not creating it. Responses like these can be extended to a simulator, and are about as convincing or unconvincing here as in the original domain.

In *Reality+*, I argue that the simulation hypothesis is roughly equivalent to a combination of the creation hypothesis (the world is created) and the it-from-bit hypothesis (computation underlies the physical world). Schwitzgebel's worries about evil are clearly problems for the creation hypothesis generally, and not for the simulation hypothesis specifically. So when he says "Let's hope we're not in a simulation", the more general point here is "Let's hope there's not a creator". This point applies both inside and outside simulations.

Speaking for myself, I suspect that if our world is created, our creator is at best morally imperfect. But I'm not sure this means we should hope there's not a creator. We already know there is much suffering in our world, creator or no creator. I'm not sure that knowing there's an imperfect creator would make things much worse than we already know them to be. Schwitzgebel suggests an imperfect creator would make the world an ethically and axiologically worse place, but I'm not sure it would be worse for us. Perhaps this would reasonably lead us to worry more about bad things that might happen (the world ending suddenly, for example), but it might also raise our hopes for some good things (an afterlife, for example). The same goes if we're in a simulation.

REFERENCES

Godfrey-Smith, P. (2024). Simulation scenarios and philosophy. *Philosophy and Phenomenological Research*. DOI: 10.1111/phpr.13124

Schneider, S. (2024). Illusory world skepticism. *Philosophy and Phenomenological Research*. DOI: 10.1111/phpr.13123

Schwitzgebel, E. (2024). Let's hope we're not living in a simulation. *Philosophy and Phenomenological Research*. DOI: 10.1111/phpr.13125

How to cite this article: Chalmers, D. J. (2024). Taking the simulation hypothesis seriously. *Philosophy and Phenomenological Research*, 1–10. https://doi.org/10.1111/phpr.13122