17

The Simulation Hypothesis

Metaphysics, Epistemology, Value

David J. Chalmers

I'd like to thank Grace Helton, Terry Horgan, and Christopher Peacocke for their rich commentaries on my book *Reality+*. As it happens, all three of them focus on the simulation hypothesis: the hypothesis that we are living in a lifelong computer simulation. Where the simulation hypothesis is concerned, I have three main theses in the book, one concerning metaphysics, one concerning epistemology, and one concerning value.

Metaphysics: If we're in a simulation, the objects around us are real.

Epistemology: We can't know we're not in a simulation.

Value: We can live a good life in a simulation.

These three theses about simulation are very similar to the three main theses in the book concerning virtual reality, except that they concern only simulation scenarios and not ordinary virtual reality scenarios. The theses are interconnected. The core thesis is the metaphysical thesis, simulation realism, which has consequences for both epistemology and value. Simulation realism blocks the inference from the epistemological thesis to external world skepticism. Simulation realism also blocks one key argument against the value thesis: that life in a simulation is not valuable because it is illusory.

The three commentators address all three of these theses. Horgan argues against the metaphysical thesis. Peacocke argues against both the metaphysical thesis and the epistemological thesis. Helton uses epistemological considerations to argue against a version of the value thesis. I'll address the three commentaries in this order.

1. Horgan on Simulation Realism

Horgan rejects simulation realism. He holds that a being in a simulation, such as a philosopher's brain in a vat (BIV) has systematically nonveridical beliefs about its world. His case for this view is grounded in what he calls a Cartesian intuition that a BIV has largely nonveridical beliefs, combined with a phenomenological examination that tends to reinforce these intuitions.

On Horgan's view, we have phenomenal acquaintance with a number of properties and relations, including especially spatial properties and relations, which we attribute to entities in our environment. A BIV is phenomenally identical to an ordinary non-BIV, is acquainted with the same properties, and attributes them in a similar pattern. But where the non-BIV may attribute these properties veridically, a BIV's attribution of them is intuitively not veridical.

It is true that simulation realism is counterintuitive for many people. That is why I argued for it at considerable length, especially in chapters 9 and 22. Three central arguments for simulation realism from these chapters are outlined in the precis. These arguments have as their key premises: "If the simulation hypothesis is true, the it-from-bit-creation hypothesis is true" (p. 171), "Photons are whatever play the photon role" (p. 176), and "Physical theories are structural theories" (p. 413).

Horgan does not directly address any of my arguments. I suspect that he would deny the three key premises I have outlined. He might well reject the structuralist view of the content of physical theories, probably by holding that spatial claims in our physical theories have some more-than-structural content. Since he thinks we are directly acquainted with spatial properties and relations, then perhaps this acquaintance might ground non-structural content in our theories of the world.

Still, structuralism (or structural realism) is a very popular view of physical theories, and I suspect that many physicists and philosophers of physics would reject Horgan's view that physical theories have more-than-structural content. After all, on Horgan's view it's quite possible that the properties we have phenomenal acquaintance with are not instantiated in our environment (since phenomenology is constitutively independent of the environment).

Horgan will have to say that if so, our physical theories involving space are false, even if they capture the mathematical structure of the world and are predictively successful. I think the view that our physical theories would still be true is in many ways more attractive.

What about the Cartesian intuitions that Horgan thinks undercut simulation realism? Unlike some structuralists, I am inclined to give these intuitions some weight but to relocate them. As Horgan notes, I take these intuitions to concern the Edenic content of perception. At one level, our experience presents objects in the world as having Colors (Edenic colors: primitive qualities of Redness and Greenness, as in the garden of Eden) and as being located in Space (Edenic space: a non-relative, Euclidean space that contains everything). If we are in a simulation, objects do not have the Colors or Spatial locations that we take them to have, so the Edenic content of our experiences and beliefs will be nonveridical.

However, I don't think the falsity of Edenic contents entails that our ordinary experiences and beliefs are nonveridical. In the case of color, we discovered long ago that apples are not (Edenically) Red. But apples are (non-Edenically) red all the same. Likewise, physics strongly suggests that our world does not have Edenic space. But Michael Jordan is over six feet tall all the same. The truth-conditions of our ordinary experiences and beliefs are not given by their Edenic contents but their non-Edenic contents.

So, I accommodate Horgan's Cartesian intuitions by agreeing that if we are in a simulation, the Edenic content of our beliefs and experiences is false, but denying that this entails that our beliefs and experiences are false in the ordinary sense. In this respect simulations are no worse off than our ordinary post-Fall relativistic world. The world lacks Space and Color, but it still has space and color. The same goes for a simulation.

Where color is concerned, Horgan himself endorses a closely related two-tiered picture. He thinks color experiences represent Edenic colors and are always nonveridical, but that color judgments represent non-Edenic colors (physical properties that play the causal roles of colors) and are often veridical. This framework allows Horgan to respect both phenomenological intuition and the correctness of our ordinary judgments that apples are red. He also indicates openness to a corresponding framework involving perception and judgments about solidity. If we do the same for space, we can likewise respect both phenomenology and the correctness of ordinary judgments, and the same framework would allow our beliefs to be true in a simulation.

Now, Horgan insists that his two-tiered picture with nonveridical Edenic perception and veridical non-Edenic judgment does not extend to veridical non-Edenic judgments in a simulation. He says there are "reference-eligibility constraints" on counting even as non-Edenic space that a simulation does not meet. At this point, however, Horgan's judgment about the constraints

seems somewhat theoretical and not a matter of clear phenomenological intuition. Those intuitions concerned Edenic color which has been accommodated. I think that once we have fallen from Eden, it is not clear why non-Edenic digital properties that play the roles of color and space should not be able to make our judgments veridical.

In the key case of space, Horgan rejects the two-tiered picture, holding that both perception and judgment have Edenic spatial contents. He holds that unlike Edenic colors, Edenic space is instantiated in the actual world: space is Edenic space.

My view is that it is implausible that space is Edenic space. Edenic space—space as presented in the manifest image—is Euclidean, non-relative, and fundamental. Space as presented in the scientific image is non-Euclidean, relative, and quite possibly nonfundamental. Horgan doesn't consider these three differences. When he asks how the scientific worldview might undermine Edenic space, he instead considers three different factors: space's role in motion, interaction, and perception. I allow that all of these are present in both Edenic and non-Edenic space, so these are the wrong factors to consider in the fall from Eden.

Now, Horgan might argue that space in the manifest image is not committed to being Euclidean, non-relative, and fundamental. If Edenic space were neutral on these things, then Edenic space would be consistent with modern science. Perhaps one could make the case that the phenomenology is not entirely committed to Euclidean space or fundamental space (though space does seem to be presented as a basic container for everything). I do think it is hard to reconcile Edenic space with relativity, however. I think that Edenic space is presented to us as non-relative and we can't really grasp the idea of it being relative to a reference frame. We can grasp the idea of space being relative, but that requires a fall from Eden.

Furthermore, even if we allowed that there has been no spatial fall from Eden due to science in the actual world, Horgan's general framework presumably allows that certain more radical changes could lead to such a fall. Then paralleling the case of color, we would have Edenic spatial contents for perception, non-Edenic contents for judgment, and our spatial judgments would be largely true despite the fall. Under that framework, if we were to find that we were in a simulation, the digital quasi-spatial relations that play the space role in a simulation would qualify naturally as non-Edenic space.

Horgan may again say that digital properties do not meet his "reference-eligibility constraints" on being contents of (non-Edenic) spatial judgments, but again these constraints and their justification are far from clear. At this

point, Horgan cannot derive much support from Cartesian intuitions, which largely concern Edenic space. He does suggest that one constraint is that the relevant non-Edenic properties must be instantiated by objects we perceive, and that in a simulation, we do not perceive genuine objects at all. But I've argued at length that this is wrong: when in virtual reality, we perceive real digital objects. If Horgan has an argument against that thesis, it would be interesting to hear it.

To sum up: even on Horgan's own preferred semantic framework, the simulation hypothesis can naturally be understood as a fall from Eden that renders the Edenic contents of perception false, but allows our ordinary judgments to be true. If so, we have a form of simulation realism.

2. Peacocke on Simulation Realism and Simulation Knowledge

Peacocke first concentrates on metaphysics, arguing against simulation realism. He then focuses on epistemology, arguing against the thesis that we don't know we're not in a simulation.

Peacocke's case against simulation realism (unlike Horgan's) focuses squarely on my structuralism about the physical world, and especially on my spatial functionalism. This is roughly the view that space is picked out by its functional role: in slogan form, space is what plays the space role. Like other sorts of realizer functionalism (for example about color), the idea is that the slogan is a priori but not necessary.

Spatial functionalism provides a natural case for simulation realism via the idea that if we are in a simulation, a certain cluster of digital properties and relations play the space role and thereby qualify as spatial properties and relations. These spatial properties and relations are genuinely instantiated in the simulation roughly where they seem to be. If so, our spatial experiences and spatial beliefs are veridical, which helps to vindicate simulation realism.

Peacocke starts by distinguishing functionalism about space from functionalism about the contents of spatial experience. For what it's worth, I am not a metaphysical functionalist (sometimes called a role functionalist) who holds that spatial properties are identical to functional properties. Rather, I am a realizer functionalist who holds that (non-Edenic) spatial properties are picked out as whatever properties play the relevant functional role. I also think there are Edenic spatial properties (uninstantiated in our world) that are not functional properties.

As for spatial experience, I think it has Edenic content which is not functional (involving primitive Edenic properties), Fregean content which is functional (holding e.g. that the property that plays the space role is so-distributed), and Russellian content that need not be functional (involving non-Edenic spatial properties).

Peacocke homes in on my realizer functionalism about space, asking whether the space role (the condition that spatial properties must satisfy to count as spatial) is truly independent of space itself (is extricable), or involves space itself (is inextricable). The thought is that extricable conditions might allow a reductive identification of spatial properties, whereas inextricable conditions will have an element of circularity.

My own account of the space role appeals partly to space's role in motion (roughly, location varies continuously with time) and interaction (roughly, closer things interact more), and partly to its role in perception (roughly, things that are a certain distance away tend to cause experiences as of being that distance away). Peacocke thinks that the motion and interaction conditions alone are inadequate to pick out the spatial. He holds that the perception condition may be adequate, but thinks it is inextricably spatial, as spatial experience is fundamentally characterized in terms of space. This leads to a form of circularity and cannot ground a reductive identification of space.

On my view, we escape from circularity by distinguishing Edenic and non-Edenic space. Spatial functionalism is a view of non-Edenic space. Non-Edenic space is picked out in part as what causes certain spatial experiences. Spatial experiences consist fundamentally in relations to Edenic spatial properties. As a result, the perceptual part of the space role is inextricable from Edenic space, but is extricable from non-Edenic space. Since the space role is used only to pick out non-Edenic space, circularity is avoided.

Even if one eschews the Edenic apparatus, I am not sure how bad the resulting inextricability is for simulation realism. Let's say that spatial properties are partly picked out as the causes of spatial experiences, and that spatial experiences consist partly in relations to spatial properties. In my view this inextricably spatial role can nevertheless give us some guidance as to the nature of spatial properties, as long as we have a prior grasp of the category of spatial experiences. I think we plausibly have a phenomenological grasp of what counts as a spatial experience, one that does not depend much on metaphysical disputes about the nature of spatial properties. And we can know that if we are in a simulation, it is certain digital properties that bring about spatial experiences. That can help us to identify

these digital properties with spatial properties. Realizer functionalism does not need to have an entirely reductive characterization of the role in order to helpfully characterize the realizer.

Peacocke also sets out an intriguing thought experiment in which pitch plays some of the roles of space, in order to cause trouble for spatial functionalism. He writes:

> Consider a two-dimensional genuinely spatial world in which there are sound events in the space. In this world, we stipulate, a sound event causally interacts only with other sound events that are close to it in pitch. Pitch then meets Chalmers' interaction condition for being a spatial dimension. There is also a relation that realizes a notion of motion across the two spatial dimensions and across the dimension of pitch. (A sound event can move as other events, such as hurricanes, can move.) But none of this makes pitch into a spatial dimension. Pitch here plays the role Chalmers specifies, but it is not a spatial dimension. Nor is experience of pitch in itself spatial perception.

Peacocke uses this thought experiment to argue that the nonexperiential aspects of the space role (in terms of motion and interaction) are inadequate to pick out space, so that the role in perception is required to do real work (leading into the problem of inextricability discussed above). Where my defense of simulation realism is concerned, having to invoke the combined role would be OK, as I think I have an answer to the inextricability problem.

At the same time, I am interested in whether one could develop a wholly *nonphenomenal* spatial functionalism where the roles that pin down space do not involve experience in any way. This might be especially relevant for the purposes of understanding "emergent spacetime" (that is, the grounding of spatiotemporal properties in more primitive non-spatiotemporal properties) in the philosophy of physics, where physicists will prefer a reduction that gives no special role to conscious experience. (See my "Finding Space in a Nonspatial World" for discussion.) Peacocke's argument threatens to rule out any sort of nonphenomenal spatial functionalism. So I am interested to evaluate it.

As it stands, Peacocke's description of the scenario is perhaps somewhat tendentious. He stipulates a two-dimensional spatial world with a third pitch dimension that plays a certain role. This is more or less to stipulate that the pitch dimension is nonspatial. Still, it isn't hard to describe the

situation in a more neutral way, and many will still have Peacocke's intuition that the pitch dimension is nonspatial.

Another problem with the case is that Peacocke is describing the case as a counterfactual possible world in which pitch plays the space role without being spatial. But recall that realizer functionalism is not committed to the thesis that necessarily, space is whatever plays the space role. On this view, it is metaphysically possible that something nonspatial could play the space role without being spatial. So this version of Peacocke's case does little to rule out (nonphenomenal, realizer) spatial functionalism. To connect better with the sort of spatial functionalism at issue here, we need to think of the case as an epistemic possibility (at least one not ruled out a priori).

Consider the following as an epistemically possible hypothesis about our world: there are two spatial dimensions and a third pitch dimension, where pitch plays a spatial role. To start with, we can stipulate that pitch plays both the experiential and nonexperiential space roles: for example, it causes experiences of spatial location along the relevant dimension. And we may as well suppose that experiences in this world are (and always have been) indistinguishable from those in an ordinary spatial world. Is this a scenario where pitch is entirely nonspatial, or is it one where a third dimension of space is constituted by pitch? The latter does not seem at all implausible to me. Certainly, if we discovered that we lived in this world, I think that this is what we would say. We would take ourselves to have made a discovery about the nature of three-dimensional space in our world: that it is partly constituted by pitch.

What if we drop the stipulation that pitch plays the experiential roles of space, and hold that it plays only the nonexperiential roles in motion and interaction. Then matters are less clear. If pitch plays no role in our spatial perception (and presumably we have no spatial perception along the relevant dimension), there is perhaps some grounds for not identifying it with space. But once we have allowed that pitch can be spatial in the previous case where it plays the experiential role, we have at least blocked the intuition that pitch is simply the wrong sort of thing to constitute a spatial dimension.

In the second part of his commentary, Peacocke takes issue with my claim that we cannot know we are not in a simulation. He argues that factive perceptual states such as my perception of a cube, can serve as evidence, enabling me to know that this is a cube. Furthermore, he thinks we can know (via philosophical reflection supporting simulation irrealism) that if

this is a cube, we are not in a simulation. This allows us to know that we are not in a simulation.

Around this point, Peacocke observes in a footnote that factive views of perceptual evidence are common, and notes that my discussion of epistemology in *Reality+* is not as extensive as my discussion of simulation. Here Peacocke seems understandably to have missed the fairly extensive epistemological discussion in the online appendices to the book. Among other things, these include a substantial discussion of externalist views of evidence (under the heading "Sims won't have our evidence!"), arguing that these do not allow us to know that we are not in a simulation (at least given the simulation-argument claim that most beings with experiences phenomenally like ours are in simulations). I don't think I can improve on that discussion, so to enter it into the standard publication record, I will include an excerpt from the online appendices as the paragraphs that follow.

Some philosophers hold that our evidence about the world goes well beyond our conscious experience to include elements of the external world. If so, we may have evidence about the world that a perfect simulation does not (see Weatherson 2003). For example, I am seeing a wooden desk in front of me. This desk is part of my evidence. A perfect sim simulating me is not really seeing a wooden desk in front of it. There is no wooden desk in the simulation at all. At best there is a simulation of a wooden desk. So the sim does not have my evidence. Even if most people with conscious experiences like mine are sims, most people with evidence like mine are not. So given my evidence, I can be confident that I am not a sim.

This line is somewhat reminiscent of Moore's line that his hands are proof that the external world exists, although with the weaker notion of evidence replacing the stronger notion of proof. One reply is that I cannot know I have the evidence of my desk (or my hands). That's part of what we're trying to determine. But for these philosophers (so-called externalists about evidence), what matters for me to know I am not in a simulation is that I have the evidence of a nonsimulated world, not that I know that I have it.

Another reply is that if I am right about the Reality Question [that is, about simulation realism], then if I am simulated I too really see a wooden desk in front of me. If so, my evidence about a wooden desk does not really cut against the simulation hypothesis. But an opponent might reject my line on the Reality Question, and at this stage I do not want to presuppose it.

More importantly: once I know that most people with my conscious experiences are sims, my external evidence can no longer justify my belief that I am not a sim. We can bring this out with a series of analogous cases.

Suppose I'm told by a reliable authority that half the people in the world (selected randomly) have just been imperceptibly given a drug so that they are falsely hallucinating a normal-seeming environment in front of them, while the other half are perceiving normally. I have an experience as of a cat in front of me. Suppose that in fact I am one of the lucky ones perceiving normally, though I have no special indication of this. How confident should I be that I am really seeing a cat? An externalist could suggest that I have the evidence provided by the real cat, so I should be very confident that this is a cat. But this seems clearly wrong. In this circumstance I should be only 50% confident that I am perceiving accurately, and correspondingly 50% confident that I am seeing a cat. In a similar way, if I know that 50% of people with experiences like mine are sims, I should be 50% confident that I am a sim.

Likewise, suppose I know that nine out of ten "zebras" in zoos are holograms that look exactly like real zebras. Suppose that on one occasion I happen to be seeing a real zebra. An externalist may say that in this case I have the real zebra as evidence, so I can know I am not seeing a hologram. But it seems clear that I do not and cannot know this. My knowledge that holograms are common prevents the zebra from justifying my belief that this is a hologram. In fact, I should be 90% confident that I am seeing a hologram.

Moving closer to the sim case, suppose I'm told that in nine out of ten countries in the world, all apparent zebras in zoos are holograms. Absent any indication that my own country is special, then I can't know that what I'm seeing is not a hologram. Even if I'm actually seeing a zebra, it would be rational to be 90% confident that we're seeing a hologram.

Now moving to the sim case: suppose I know that in nine out of ten worlds, all apparent tables are simulations. Absent any indication that there's anything special about my own world, then I can't know that I'm seeing an unsimulated table. Even if I happen to be a nonsim, it would be rational to be 90% confident I'm seeing a simulated table, and 90% confident that I am a sim.

Furthermore, it is quite straightforward for externalists about evidence to accept these verdicts. Even most externalists allow that perceptual evidence (e.g. seeing a zebra) can be defeated by other evidence (e.g. knowing that most zoos contain holograms). When we grant that 90% of beings with evidence like ours are sims, this in effect overwhelms any evidence provided by our being nonsims, so that we should be 90% confident that we are sims. An externalist of this sort can endorse the key indifference principles that we have been working with. I think that reflection on the cases we have discussed recommends this view.

In the philosophical literature, some related cases are pressed against the externalist by Roger White (2014) and Jonathan Vogel (2008). I don't know of explicit discussion of these cases by externalists. As I've noted, many externalists allow that a subject's external evidence can be defeated by other evidence, which when applied to the simulation cases will tend to lead to the conclusions in line with the original indifference principle.

At least one externalist, Maria Lasonen-Aarnio (2010), takes what she calls the "radical option" of holding that knowledge is not undermined by potential defeating evidence. On this line, someone seeing a zebra might continue to know that they are seeing a zebra in a case like this even in light of the evidence about holograms—although their believing they are seeing a zebra would be unreasonable. This radical externalism (combined with the view that sims undergo illusions) might lead to a view where we might be able to know that we are nonsims (if in fact we are), even though we know that 90% of beings with experiences like ours are sims. Even on Lasonen-Aarnio's view, this defeating knowledge would make it unreasonable for us to believe we are nonsims, however. It seems that it would be most reasonable for us to have a high credence that we are sims.

Returning to my reply to Peacocke: One thing this discussion brings out is that the Bostrom-style simulation argument (construed as an argument that there is a significant probability that we are in a simulation) plays a crucial role in my full argument that we cannot know we are not in a simulation. As I put things at the end of chapter 5 of the main text: the simulation argument makes it a serious possibility that we are in a simulation; and once it is a serious possibility, these [anti-skeptical] arguments cannot rule it out. Peacocke's discussion addresses my initial prima facie case (in chapter 2 of the book) that we can't know that we are not in a simulation. It would be very interesting to know his response to my use of the simulation argument (in chapter 5 and the associated appendices) to defang anti-skeptical responses such as his own.

3. Helton on Solipsism and the Value of Social Knowledge

Grace Helton focuses especially on issues about value. These issues build on Helton's earlier epistemological work, in which she argues that skepticism about other minds should be taken seriously. In her earlier work, Helton holds that there is a significant possibility that we live in a *solipsistic*

simulation, in which (apparent) other people lack minds altogether. The solipsistic simulation hypothesis seems at least to be a coherent hypothesis. Helton argues that my structuralist path to simulation realism, which suggests that tables will be present in a simulation in virtue of its structure, does not work where other minds are concerned. Further, there is reason to think that many solipsistic simulations may be built, in part because they may be more efficient or more ethical for simulators to run. If so, then an adaptation of the simulation argument suggests that we should give significant credence to the solipsistic simulation hypothesis.

All this has consequences for value. Intuitively, the solipsistic simulation hypothesis is very bad. If our loved ones don't love us back, and if no one ever understands us, that is bad. Helton makes the case in more detail by arguing that if we are in a solipsistic simulation, we lack social knowledge, and that for most of us, social knowledge is extremely valuable, perhaps so valuable that it swamps the value of most other sorts of knowledge. If so, we should not be sanguine about the possibility that we are in a simulation. If we are in a simulation, we may well be in a very bad one.

I am open to Helton's conclusions. Her main theses are consistent with the letter of what I say in the book: I say that we can live a good life in a virtual world, but I don't say that all virtual worlds will support good lives. Her theses may cut against the spirit of the book, though, which tends to suggest (without quite saying it) that simulated worlds may be about as good as nonsimulated worlds in the most important respects. If most actual simulated worlds are solipsistic and most actual unsimulated worlds are not, then at least in actuality, simulated worlds are typically much worse in an important respect than nonsimulated worlds. So it is worth going over Helton's reasoning to see which parts I accept.

I agree with Helton that the problem of other minds is a serious problem. More precisely, I think that the problem of other *conscious* minds is a serious problem. I hold that consciousness cannot be analyzed in functional terms, and a scenario where others are zombies without consciousness cannot be excluded a priori. Now, there may be some mental states and properties that can be functionally analyzed. For those mental states, I think the problem of other minds is much less pressing. These functional mental states will be present at least in a perfect simulation. Helton suggests that even for these mental states, there will be behaviorally normal beings ("faux-folk") that lack the mental states entirely, but this is not obvious to me. There will be extreme cases like Ned Block's Blockhead, but this can't be realistically created (and certainly can't be efficiently created) due to combinatorial explosion.

There are AI systems such as the GPT models, but at least to date these systems fall short of fully humanlike behavior, and it's unclear just what sort of mental states they might have if and when they get there.

Still, the problem of other conscious minds is a problem enough. If everyone aside from me is a zombie, that is a bad thing. And I agree that structuralism doesn't do too much to rule out zombies. I am a structuralist about the physical world but not about consciousness. I do think there are other routes that work better to establishing other conscious minds: for example, abductive arguments (inferring psychophysical laws from our own case) get at least some purchase.

Helton's version of the simulation argument applies to zombies as well as to faux-folk. Ethical simulators might well prefer to build simulations in which everyone or almost everyone is a zombie. Doing so avoids needless suffering and minimizes playing god with conscious beings. Of course, for simulators to do this, zombies must be nomologically possible. My view is that physically or functionally identical zombies are not nomologically possible—but I can't be certain of this, so there remains a skeptical possibility here. We could also run Helton's argument with behaviorally identical zombies, perhaps along the lines of sophisticated large language models. Behaviorally identical zombies are not obviously possible but are hard to rule out.

There is one respect in which the solipsistic simulation argument is not as strong as the original simulation argument. This concerns the numbers. Construed as an argument that we are probably in a solipsistic simulation, the argument requires roughly that most conscious beings (or better, most beings with consciousness like ours) are in solipsistic simulations. But even if most simulations are solipsistic simulations, it does not follow that most conscious beings are in solipsistic simulations. Suppose there is one nonsolipsistic simulation (with one billion conscious beings) and one million solipsistic simulations (with one conscious being and 999,999,999 zombies in each). Then despite there being a million times as many solipsistic simulations, one thousand times as many conscious beings will be in nonsolipsistic simulations. This weighting of the numbers makes it *much* less antecedently probable that most conscious beings will be in solipsistic simulations than that most conscious beings will be in simulations simpliciter. In effect, this weighting serves as a "solipstic sim blocker" (analogous to the sim blockers of chapter 5) that potentially explains why relatively few beings are in solipsistic simulations, in a way that may defang the solipsistic simulation argument to some extent.

This factor will also tend to greatly reduce the conditional probability that we are in a solipsistic simulation, conditional on being in a simulation. That in turn will greatly reduce the probability that we are in a bad simulation, conditional on our being in a simulation. This reduction will help block the value-theoretic objection that if we're in a simulation, it is probably a bad one.

What about Helton's social swamping view: that for many people, the value of social knowledge massively swamps the value of non-social knowledge? One preliminary point is that I am not sure that the value of knowledge is the key issue here. Intuitively, what is really bad about a scenario in which others are zombies is not that I don't *know* they have minds. It's that they don't have minds. We can diagnose this by considering a situation in which I have an unjustified but true belief (or a Gettiered justified true belief) that others have minds. Here, others have minds although I don't know it. My intuition is that this situation is far better than the situation in which others lack minds, and is only a little worse than a situation in which others have minds and in which my true belief is knowledge. For similar reasons, I think even the belief isn't crucial here. The dominant source of value in these cases is not knowledge of or belief in other minds, but other minds themselves.

More generally: I think that if there's something whose value swamps other factors here, it's not social knowledge per se, or social belief, but social reality. Helton's social swamping claims can naturally be understood as claims about the value of social reality (compared to the value of non-social reality), and I'll henceforth understand them this way.

Does the value of social reality swamp the value of non-social reality? I think it depends on the person. For many people, non-social reality is very important. For a gardener or a mathematician or an explorer, a huge amount of the value of their lives may come from non-social sources. In some cases this may exceed value from social sources, in other cases not. But certainly this non-social value need not be swamped by social sources. Still, Helton's core thesis claims only that for many people, social value swamps non-social value, and that claim seems quite plausible.

What follows for simulations? Let's start with mindless simulations: those in which others lack any mental states. It's at least arguable that if others lack any mental states, then there's little or no social reality. If the value of social reality is the dominant source of value for many people, then for these people, mindless simulations will be lacking this dominant source of value, and will thereby be much worse (other things equal) than nonsimulated worlds with minds.

Now, I am not sure that mindless simulations (indistinguishable behavior, no mental states) are possible. Zombie simulations (no conscious states) may be possible, but it is less clear that these simulations lack social reality. It's at least arguable that if others have non-conscious mental states, this can support considerable social reality. Still: social reality or not, it seems plausible that for many or most people, a situation in which others lack consciousness entirely is a very bad one.

What follows for the value of simulations more generally? One thesis I am tempted by is that life in a simulated world is roughly as good as life in a corresponding nonsimulated world, where the two worlds have the same structure and importantly have corresponding minds. (I don't say exactly as good—nonsimulated worlds may gain some added value from nature or fundamentality or some other factor.) That thesis is not threatened by solipsistic simulations. For all we've said here, a solipsistic simulation may be roughly as valuable as a corresponding solipstic non-simulation, with both being much worse than non-solipsistic counterparts.

What about the thesis: if we're in a simulation, things are roughly as good as if we're not? This thesis is threatened by Helton's argument, since she holds that if we're in a simulation, a form of solipsism is more likely to be true than if we're not, and she thinks that solipsism is very bad. I agree that solipsism is very bad. I'm open to the idea that many simulations are solipsistic, perhaps for reasons of ethics or efficiency. I don't think it follows that solipsism is probable, for reasons I discussed earlier. But even if this increases the probability of solipsism by a little, it will likewise reduce the expected value of life in a simulation by a little. Perhaps this gives each of us some reason to hope that we're not in a simulation.

References

Chalmers, D. J. 2021. Finding Space in a Nonspatial World. In (C. Wüthrich, B. Le Bihan & N. Huggett, eds.) *Philosophy Beyond Spacetime* (Oxford University Press).

Helton, G. 2024. (this volume) The Simulation Hypothesis, Social Knowledge, and a Meaningful Life. *Oxford Studies in Philosophy of Mind*, volume 4.

Horgan, T. 2024. (this volume) Why Virtual Worlds Aren't Real: How Phenomenal Intentionality Constrains Mental Reference. *Oxford Studies in Philosophy of Mind*, volume 4.

Lasonen-Aarnio, M. 2010. Unreasonable Knowledge. *Philosophical Perspectives 24(1): 1–21*.

Peacocke, C. 2024. (this volume) Simulation: Its Metaphysics and Epistemology. *Oxford Studies in Philosophy of Mind*, volume 4.

Vogel, J. 2008. Internalist Responses to Skepticism. In (John Greco, ed.), *The Oxford Handbook of Skepticism*. Oxford University Press.

Weatherson, B. 2003. Are You a Sim? *Philosophical Quarterly* 53: 425–431.

White, R. 2014. What Is My Evidence That I Have Hands? In (Dylan Dodd and Elia Zardini, eds) *Scepticism and Perceptual Justification*. Oxford University Press.

David J. Chalmers, *The Simulation Hypothesis: Metaphysics, Epistemology, Value* In: *Oxford Studies in Philosophy of Mind Volume 4*. Edited by: Uriah Kriegel, Oxford University Press. © David J. Chalmers 2024.
DOI: 10.1093/9780198924159.003.0017