Simulation Scenarios and Philosophy

Peter Godfrey-Smith

University of Sydney

Presented at a 2023 Pacific APA symposium on David Chalmers's book *Reality+*

This talk will mostly be about the epistemological side of *Reality+,* specifically about the idea that we should take seriously the possibility that we are, right now, living inside some sort of computer simulation.[1] I will argue against this view; I don't take the possibility seriously. I won't worry here about how many of our beliefs are true rather than false if we *are* inside a simulation. For Chalmers, "virtual reality is genuine reality" and chairs and tables still exist in them. I assume that simulation hypotheses are very surprising no matter how that issue is handled – our lives are observed and controlled by hidden powerful agents, and so on. Chalmers gives scenarios like this a probability of about 25%.

My discussion will have three parts. First, I will make some arguments more or less within Chalmers's background assumptions, about simulation hypotheses. I'll then discuss a topic in the philosophy of mind: the substrate independence of the mental, including conscious experience. This bears on the question of what kinds of simulations we might inhabit – what sort of hardware would be needed. I'll then go back to the epistemological side and put some questions about time and reasoning on the table. Here I'll look at memory, temporally extended reasoning, and the "Boltzmann's brain" idea.

1. Human Simulation Projects

2. Substrate Independence

3. Time

---

[1] Thanks to Eric Schwitzgebel for organizing the Pacific APA symposium. The speakers were Eric, Grace Helton, David Chalmers, and myself, with Anand Vaidya as chair. This file is the text as given, without references. I'll post a fuller version later on.

1. Human Simulation Projects

I take the central claim to be that we should take seriously a family of somewhat paranoid, partially skeptical, possibilities about our present situation because *our own picture of things tells us to.* Our own picture tells us some scenarios are plausible, technologically possible, will probably exist in the future if they do not exist now, etc. By "our picture," I mean mainstream science, also a background view of what intelligent agents are likely to want to do, and so on. All this we use to show that we should doubt our ordinary assumptions about what is going on around us.

That "our own picture tells us..." side is important. If, at some stage, the discussion lapses back into, "Who knows? Something weird might be going on" – then that is a different thing. At the end I'll discuss some of those issues.

There are roughly three things we might be, in these scenarios:

(i) At least partly embodied brain-in-vat or *Matrix* sorts of beings, with a biology, controlled by agents of some sort, or
(ii) Pure software beings in a computing artifact, also with a hidden controller of some kind, or at least a history of hidden control, or
(iii) Boltzmann brains – biological systems, or at least physical systems not tied to computing artifacts, not controlled but spontaneously arising and probably short-lived. (I guess we could exist within spontaneously arising Boltzmann computers, too – we could be Boltzmann software – but I won't discuss that one as I don't think it adds much to the other scenarios.)

Take first the ones with hidden agents involved - (i) and (ii). We are in a simulation that has been built for us to inhabit (or perhaps one simulation for each of us), by someone we don't know about, for their purposes. There is a lot of computing technology behind it – something like a fine-grained model of a 'world' we interact with, a model feeding us sensations and responding to our decisions, whether we turn our head, blink, drive 100 miles east, build a microscope, etc. I wonder whether Chalmers underestimates the difficulty of this task, but I am not going to worry about that. The threat to our familiar assumptions is to come from the idea that a large majority of minds like ours, a large

majority over some long span of time, will inhabit a simulation of some sort. The minds inhabiting these simulations won't be able to tell that this going on. Given that (probably) most minds will be in this situation, why not ours?

One question I then ask is: why would people (or other agents, Chat GPT-11 or whoever) bother to do this? In large numbers, what would be the point? It might be a research project. Why then make it "total," rather than very local in space and time? You might say: we don't know. As long as there's a *chance* that they might get up to this, the argument can be posed. We can introduce a multiplier, a probability-reducer, in the argument. The pressure might be reduced but it is still real.

I think the question of motive has more effect than this. Again, the point of the argument is supposed to be that within our own picture, we can see why it would make sense for things to look completely different, to most sentient agents, from how they really are. Relatively specific features of our overall picture *imply that we can't trust* a bunch of things that we usually trust. *There should be* lots of things like me, who are not living as they think they are. Whenever the argument includes a "who knows why they'd do it?" or "who knows how they'd manage this?" step, it takes a lot away.

Maybe it's not science but recreation? A rather sadistic form of it, if our virtual world is more or less as it appears, in terms of who experiences what. (A "problem of evil" arises for some versions of this view.) There's a lot of unexplained deception, anyway.

Might it be something like the scenario in *The Matrix*? Perhaps not science or recreation but an economic motive lies behind it, where we are a power supply? *The Matrix* was a good movie, but just a movie; the scenario made no energetic sense at all. They imagined humans being used as a kind of source for electricity. But we've got to be fed. So you're putting chemical energy in to get electricity out? That is the opposite of what it would make sense to try to do, as the renewable power industry has learned; the hard process is to get from electricity, via solar and wind, to a good storage medium, a fuel. Our picture says what they were doing in the movie makes no energetic-economic sense.

I assume that any biologically based simulation setup is expensive for these reasons. It has to be a "pure" sim, where people are just software, to make large numbers feasible.

Are they doing it for science or for recreation? I don't see either as likely. But is this possible anyway?

2. Substrate Independence

I doubt that experience like mine or yours right now could exist in a variety of non-biological realizations. If we are "pure sims," in a computer or being shuttled around between computers, then the mental, including the experiential, has to have a lot of "substrate independence" (SI), and I don't think it does.

The way I said it used the phrase, "a lot" of SI – that makes it sound like a matter of degree. That is how we should think about SI, and I think there's not much of it.

Other people think the mental just *is* substrate-independent, or at least that there's a great deal of it. Where does that leave us here, in a short commentary? I think I can give reasons for saying that some of the usual reasons people *believe* there is SI, or is a lot of it, are not good reasons.

Chalmers starts out with the idea of a "perfect simulation" of a brain such as his.

> I'll concentrate on just one type of machine: a perfect simulation of a brain, such as my own. The brain simulation is a digital system running on a computer. If we can establish that one digital system is conscious, then we know there's no general reason why digital systems cannot be conscious, and the floodgates will open. Compared to other machines, the simulated brain has the advantage of maximizing similarity to a human brain, which makes some of the reasoning simpler. For example, we don't need a full theory that tells us which systems are conscious, since we can start from the one case that we know is conscious: ourselves. Furthermore, brain simulation helps to illuminate simulated worlds and uploading, which are the main reasons we're pondering digital consciousness here.
>
> How would simulating a brain work? We can suppose that every neuron is simulated perfectly, as is every glial cell and other cells throughout the brain. The interactions between neurons are simulated perfectly, too. All the electrochemical activity is simulated, and so is other activity, such as blood flow. If there's a physical process in the brain that makes a difference in how the brain functions, it will be simulated. In what follows, I'll adopt the simplifying assumption that only neurons need to be simulated, but everything I say will apply without that assumption. (287, emphasis added)

Philosophers have gotten used to talking like this, but this "perfect" duplicate idea is problematic. We can reach my point by thinking about the details of timing. Some facts

about timing in a brain are certainly "functionally relevant" – both the time which things take to happen as a whole, also the simultaneity or non-simultaneity of parallel processes, and so on. Some other fine details of timing look like they probably don't matter much, but they're still *there*; the brain processes are still different if they are changed. Time gets ignored in classical functionalist discussions – you've got the machine table, or an imagined network. If it gets things done, it gets things done. But neurally, timing is a big deal.

How much fidelity of timing does there have to be in a "perfect" duplicate? That is an ill-posed question. Suppose something takes a millisecond longer, or a tenth of that, or the synchrony of two parallel processes is not quite the same.... What there is, in these cases, is functional *similarity,* to various degrees. Timing can make the point, and so can other details – a few more ions here or there. Chalmers says, "If there's a physical process in the brain that *makes a difference* in how the brain functions, it will be simulated" – but that raises the same question. Is another millisecond here or there a "difference"? Yes, it is a difference, but a tiny one. It's a difference to experience, too – if the process that is the basis of experience goes on another millisecond, then that is different.

You might say: millisecond differences are not part of experience. They don't exist in experience, in some sense. One physical system can be a "perfect duplicate" of another, not in all ways but in all ways that matter to experience, even if there are micro-functional differences of that kind. I don't see any reason to believe this – even if you can't compare two experienced events and judge that one took a millisecond longer than the other, they can still be different; one experience did go on for a bit longer than the other.

These points can be made in this in-principle way, but can be made more concrete as well. What is it within a brain that is the basis for felt experience? It might all be relatively discrete neuron-to-neuron effects that involve action potentials – this cell fires and that one does.... And it might not be. Suppose it's like this – my version of this picture draws on the work of Bruno van Swinderen, a fly researcher at University of Qld. Van Swinderen thinks that there is a duality of processes in brains (not just ours, but those of flies and others) that matters to both cognition and experience. There are spikes (action potentials), and there are large-scale, less local, electrical oscillations, the kinds of

things picked up in an EEG or (in the flies) LFP. These involve rhythmic movements of ions across cell membranes below the threshold that initiates a chain reaction and a spike. In van Swinderen's picture, each of these two processes affects the other. Sub-threshold oscillations prime or suppress spikes; spikes affect oscillations. A single neuron's firing can be part of the activity of several different ensembles of neurons, according to the details of its timing. If it fires at t1, and that is part of a pattern in one ensemble; if it fires at t2, it acts as part of another. Those timing details are affected by large-scale sub-threshold oscillations.

Exact timing matters, spatial details matter. There are the point-to-point local effects that are basis for computational models in neuroscience, and there are the more holistic phenomena that depend on physical details. Nervous systems bring these two together. When I said above, "suppose it's like this," in the brain, I don't mean: as a thought experiment that shows the possible sensitivity of brain processes to physical details. I mean: this is the bet I would make.

Let's think about neural replacement scenarios in this light. (This is discussed in the last part of my paper "Mind, Matter, and Metabolism" paper, 2016, though without the oscillatory details.) I say that as you change the makeup of the system, you change the micro-functional profile. And you will change the behavior. Initially only in tiny ways – a millisecond here, a millimeter difference in the movement of a hand there. A neural replacement argument - "you would not notice, no one would notice, if your neurons were slowly replaced" – depends on the idea that the behavior stays exactly the same. Whatever it means to say you would not notice "from the inside," there are going to be differences visible from the outside.

Some people would deny this – they would say: no, it is possible to have no differences when the physical substrate is changed. Do they mean really *no* differences? Not millisecond-scale differences? It is surely more likely that as you replace neurons, the system changes micro-functionally, and it will also change micro-behaviorally – in timing, fidelity of repetition, in all sorts of details. What is happening, inside and outside, is different. This means there's no reason to believe that experience is unaffected in a slow replacement. First a little, then a lot. This could include all sorts of transformation and fading.

Setting aside replacement scenarios: If the picture sketched above is right, then it is hard to get felt experience outside of physical systems of a specific kind. Someone insists in reply: "We can simulate all that." At this point we need to also keep an eye on the distinction between simulation in the sense of *representation* and simulation in the sense of *realization*. It can't just be a representation; you can't just compute the functions and go from one represented value of a variable to another. You can't handle the large-scale oscillatory phenomena by having a bunch of memory addresses that form a coordinate system and that you write values to very quickly. You need to have a version of the process, some materially different copy of it, *in* there – with the right kind of relations between parts, not just *some* relations that map to the neural ones. If someone says "We can simulate all that" in the realization sense, then... they can *say* it. I don't see why we should believe it.

If a lot of substrate independence is unlikely, that makes large numbers of simulated minds less likely. Simulated minds are only cheap if we're talking about solar-powered computers with conscious software. Otherwise they're expensive.

3. Time

In many discussions of skepticism about the external world, I've thought there's a fudge or evasion going on. Here I mean the sort of reflection where you weigh things up, go through arguments of various kinds, get coffee, go back to work... and the goal of it is to suspend belief in the ordinary stuff around you and see if you can work your way back. If you are doing this, you should also suspend belief in just about everything about the past, including the immediate past. You have no reason to believe that you spent the morning doing this, no reason to believe that you reasoned through, a few seconds ago, to the things that you find in your head now. We should not take seriously an external-world "spatial" suspension of belief without taking equally seriously the temporal side.

Santayana talked about how wholesale doubt leads to our being reduced to the "solipsism of the present moment" (*Scepticism and Animal Faith*). This is one reason I like Alex Proyas's movie *Dark City*, a paranoid film in which skepticism about the past is true. The ordinary people in the movie have their memories artificially reconfigured every 24 hours (? or half day?), as part of something like an experiment.

If we are willing to really doubt, then we have to suspend everything, from perhaps half a second or so ago. What do you then have? You find ourselves with half a second or so of impressions and memories, trains of thought underway. Who knows which memories are reliable, including whether you remembered to carry the "2"?

A couple of people have argued from this to an anti-skeptical conclusion of some sort. A paper by Susanna Rinard argues that external world skepticism leads to skepticism about "complex reasoning," and this leads to kind of reductio. (Crispin Wright also?) My attitude is different. It's not that we can reason our way out of skepticism. It's that a certain kind of project collapses. If you really do impose foundationalist standards, trying to work your way back to a common-sense picture from what is undeniably apparent to you, you will get nowhere. (Nowhere unless you can make a move like Descartes did - finding an idea of God within you, right now, that has only one possible etiology, and working from there.)

What effect does this have on *Reality+*? These themes come up mostly at the end of the book, in a discussion of "Boltzmann brains." Physics allegedly tells us that under some circumstances we can expect large numbers of brains to pop randomly into existence, with memory traces intact. Are you one of them? Chalmers endorses a reply given by Sean Carroll -

> As the theoretical physicist Sean Carroll has pointed out, however, the thesis that I am almost certainly a Boltzmann brain is "cognitively unstable." If it's true, then I cannot stably endorse it. If I endorse it, I then must endorse that my perception of the external world is almost certainly an illusion. But then I must reject all of my scientific reasoning that's based on my perception of the external world. In particular, I should reject the scientific reasoning that led to the physical theories on which the existence of Boltzmann brains was based in the first place. Those theories are the only reason to take the Boltzmann brain hypothesis seriously. Without those theories as support, we're back to the original situation: The hypothesis that I am a Boltzmann brain is extremely improbable.
>
> A bit later: "We can know that we're not Boltzmann brains."

I don't believe the last steps – the ones that get us back to normality, to "the original situation." We have been given (under these assumptions) new reason to doubt some

things. If we think Boltzmann brains are physically likely, then we think there are lots of minds out there going through these Carroll thoughts and doing so wrongly.

This also bears on the main discussion of simulations in the earlier chapters of the book. To suspect you are in a simulation is to suspect that all your memory traces could very likely be cooked up, that your background knowledge is no good, and so on. If there are human simulators, they seem as likely to be *Dark City* simulators, giving us misleading memories even of our own experiences, as any other kind. We'd also have to throw away whatever might be the basis for the assumptions we make about what the simulators might be up to and why. If this is undermining in the Boltzmann's brain context then it's undermining here, too.

How does this relate to the other arguments in my first section? I think: there's one thing we can do when we assume and make use of our present picture, and try to work out, using familiar epistemic standards, whether the universe will come to be populated with minds that are experientially like ours but misled about their situation. I don't think it will be. (In the Boltzmann case, I'd have to hear more to be convinced it's feasible.) And there's something else we can do, perhaps prompted by the first exercise or perhaps done from scratch, when we work out what we can conclude from what is presently available to us, without making the usual "external world" assumptions and also without assuming the integrity of memory. Then, epistemologically speaking, we run into a wall.
