Online Appendices for *Reality+: Virtual Worlds and the Problems of Philosophy*

[still in progress!]

Contents:

Bostrom on the simulation argument [chapter 5]

I'm broadly sympathetic with Bostrom's argument in "Are You Living in a Simulation?", but I disagree on a number of details in the argument and the conclusion.

I'll divide my discussion into two main parts: Bostrom's argument (especially his formula for the fraction of beings that are in simulations), and his conclusion.

Bostrom's argument. Bostrom's argument rests mainly on a formula for the fraction of all observers with human-type experiences who live in simulations, as follows:

$$f_{sim} = f_P\overline{N}\overline{H}/(f_P\overline{N}\overline{H} + \overline{H})$$

Bostrom defines f_P as the fraction of all human-level technological civilizations that survive to reach a posthuman stage, $\overline{N}$ as the average number of ancestor simulations run by a posthuman civilization, and $\overline{H}$ as the average number of individuals that have lived in a civilization before it reaches a posthuman stage (call these pre-posthumans). He notes that $\overline{N} = f_I N_I$ where f_I is the fraction of posthuman populations that are interested in creating simulations and N_I is the average number of simulations that these interested populations create, so that $f_{sim} = f_P f_I N_I / (f_P f_I N_I + 1)$. He then notes that N_I is very large (as there is so much available computer power), from which it follows that either (1) f_P is very low (very few human-level populations survive to become posthuman), (2) f_I is very low (very few posthuman populations are interested in creating simulations), or (3) f_{sim} is close to 1 (most observers in human-type experiences live in simulations).

Bostrom doesn't explain the reasoning behind the formula, but the idea seems to be that (1) $\overline{H}$ is the number of humans with human-type experiences per civilization (because all and only pre-posthumans have human-type experiences); (2) $f_P\overline{N}\overline{H}$ is the number of sims with human-type experiences per civilization (because all sims with human-type experiences are in ancestor simulations that simulate all pre-posthumans), and (3) f_{sim} can be calculated by dividing the number of sims with human-type experiences by the number of humans and sims with human-type experiences (because all observers with human-type experiences are humans or sims, and there are a finite number overall).

There are a number of things in Bostrom's formula to quibble with. I'll suggest two potential objections to each of (1), (2), and (3).

(1) The assumption that all and only pre-posthumans have human-type experiences is very likely false. (a) Many pre-posthumans (e.g., those in the distant past and the medium-term future) won't have human-type experiences, given that human-type experiences are experiences broadly like ours. (b) Furthermore, many humans after the posthuman stage may have human-type experiences (e.g., if posthumans create physical simulations of human history using biological organisms on terraformed planets).

(2) (a) This thesis assumes that every ancestor simulation simulates all pre-posthuman humans, whereas it seems likely that many or most simulations will be more local than this. (b) Furthermore, (2) assumes that all sims with human-type experiences are in ancestor simulations, but there may well be human-type experiences in non-ancestor simulations.

(3) (a) This calculation of f_{sim} ignores the possibility of creatures with human-type experiences who are neither humans nor sims—robots in a nonsimulated environment, for example. (b) The fraction will also be undefined for infinite populations.

Some of these issues are minor. Problems (1a) and (2a) arguably affect the figures by only a couple of orders of magnitude, so that an appropriately adjusted version of the argument should still go through, at least if the original argument had a large enough margin for error. Problem (2b) can be fixed by broadening from ancestor simulations to human-type simulations (simulations including creatures with human-type experiences) more generally, which will increase only the resulting figure for f_{sim}. Problem (3b) can perhaps be handled by invoking limit proportions, as suggested earlier. (Bostrom addresses problems (2b) and (3b) in roughly these ways in the "Simulation Argument FAQ".)

Problems (1b) and (3a) run deeper. It requires a substantive additional assumption to ensure that sims with human-type experiences greatly outnumber posthuman humans and robots in nonsimulated environments with human-type experiences. As we've seen, that assumption can be reasonably motivated by claims about physical resources, but it remains a substantive issue. So this at least opens up a new sim blocker: *More nonsims than sims will be created.*

Some other minor differences with Bostrom's argument include: (1) I don't think the simulation argument is best formulated in terms of ancestor simulations (see the main text) and (2) I don't rely on the reasoning that once we create simulations of ourselves, we then can't rule out that we are the created simulations (see the discussion of *We know we're not the sims we create* in the appendix on further objections).

Bostrom's conclusion. Translating Bostrom's three-way disjunctive conclusion into my terms, it takes the form *Either there are sim blockers or we are probably sims*--where

Bostrom invokes two specific sim blockers: *Nonsims will die before creating sims* and *Nonsims will choose not to create sims*.

An initial issue concerns the form of his conclusion. Strictly speaking, this conclusion doesn't follow from his previous theses, which are of the form (1) *Either there are sim blockers or most beings are sims* and (2) we should have a high confidence in *We are sims* conditional on *Most beings are sims*. From these premises, one cannot derive that *Either there are sim blockers or we should have high confidence that we are sims*. All that really follows is that we should have a high confidence in *We are sims* conditional on *There are no sim blockers*, and that we should have high confidence in *Either there are sim blockers or we are sims*. As a result, Bostrom hasn't really established his disjunctive conclusion, though he has made a case that we should have high confidence in it.

(Formally: from (1) A or B (and even $p(A \text{ or } B)=1$), and (2) $p(C \mid B)$ is high, one can't infer (3) A or ($p(C)$ is high). But one can infer (3') $p(C \mid {\sim}A)$ is high) and (3") $p(A \text{ or } C)$ is high.)

A second issue concerns the disjuncts in Bostrom's conclusion. Bostrom includes two sim blockers among his three disjuncts: in my terms, these are *Nonsims will die before creating sims* and *Nonsims will choose not to create sims*. It is far from clear why other sim blockers are not included. We've seen above that on mathematical grounds alone, something like *Nonsims will create more nonsims than sims* needs to be included (or perhaps excluded by argument or assumption).

As discussed in the main text, I would also add *Intelligent sims are impossible*, *Conscious sims are impossible*, *Sims take too much computer power,* and *Simulators will* *avoid creating conscious sims.* (In Bostrom's framework, these all correspond to reasons why $\tilde{N}$, construed as the average number of human-type ancestor simulations created by interested populations, may be low.). Bostrom says that substrate-independence is an "assumption" of his argument (ruling out the second of these sim blockers), as is an assumption about computer power in the universe (ruling out the third). Perhaps further assumptions could be made to rule out the first and the fourth, though such assumptions would be far from obviously correct. In any case, Bostrom's division between assumptions of the argument and disjuncts in the conclusion seems fairly arbitrary, especially where non-obvious assumptions such as substrate-independence are concerned. So I think it makes sense to draw a broader conclusion that includes all six of these sim blockers are disjuncts.

A seventh potential sim blocker is mentioned earlier in the note: *We are alone.* Here the idea is that we are the only humanlike nonsims in the galaxy, and that as it happens we never create any humanlike sims. This sim blocker could perhaps be subsumed under *Nonsims won't create sims* (either due to extinction or by choice), but it's worth noting that unlike Bostrom's familiar sim blockers along those lines, this version doesn't require that many populations inevitably die out or choose not to create sims. It just requires that one population goes in this direction, which seems more plausible *a priori.*

A more general way to put the conclusion is that we should be highly confident that either we are sims or that there are sim blockers. If we divide up the sim blockers as in the main text, we arrive at my own preferred formulation of the conclusion: we should be highly confident that either (1) we are sims, or (2) humanlike sims are impossible, or (3) humanlike sims are possible but few humanlike nonsims will create them. In my view this version is the most general and forceful version of the simulation argument.

Further Objections to the Simulation Argument [chapter 5]

(1) *Should we be indifferent between nonsims and sims?* Premise 2 of my reformulated argument is *If most humanlike beings are sims, we are probably sims.* One natural way to make a step like this is to assume that I am equally likely to be each of the beings with experience like mine. If so, it follows that if most beings with experience like mine are sims, I am probably a sim. As noted in the main text, this sort of principle is sometimes known as a Principle of Indifference, because it means that probabilistically I am indifferent between the hypothesis that I am this candidate or that one.

More formally, let us say that a me-like being is a being whose experiences are exactly the same as my actual experiences, qualitatively speaking. What it is like to be a me-like being is exactly the same as what it is like to be me. Our key indifference principle then says that if I am certain that a specific fraction x of me-like beings are sims, then I should have confidence x that I am a sim. (We can put this formally in terms of conditional credence: Cr(phi | f_{phi} = x) = x, where Cr is rational credence, phi is any property and f_{phi} is the fraction of me-like beings who are phi.) It follows from this that my unconditional confidence that I am a sim should be identical to the expected fraction of me-like beings who are sims. (Cr(phi)=E(f_{phi}).)

This principle concerns me-like beings (beings with experiences like mine) rather than humanlike beings (beings with experiences roughly like mine at least in sharing all major sim signs and nonsim signs). However, we can move to the latter by noting that the expected fraction of me-like beings who are sims is the same as the expected fraction of humanlike beings who are sims. The reason for this is that any difference between these fractions would yield a sim sign or a nonsim sign in me-like experience, and all sim signs and nonsim signs in me-like experience have been built into humanlikeness. If that's right, my unconditional credence that I am a sim should be identical to the expected fraction of humanlike beings who are sims. (Cr(sim) = E(f^*_{sim}), where f^*_{sim} is the fraction of humanlike beings who are sims.). This derivative indifference principle is enough to support premise 2 of the argument: *If most humanlike beings are sims, we are probably sims*.

Philosophers sometimes reject indifference principles such as the key principle above. In "Are You a Sim?" (*Philosophical Quarterly* 53: 425-31. 2003), Brian Weatherson raises a number of questions about what sort of indifference principle can support the argument and whether they are true. Some of Weatherson's issues do not apply to our formulation in terms of me-like experiences, but others do. For example, it is arguable that some beings with me-like experiences have evidence that other beings with me-like experiences lack, at least if we understand evidence as being partly external (e.g. really seeing a table) whereas experience is internal (e.g. having an experience as of a table). If so, then my external evidence of a table may give me reason to favor some of these me-like beings (those seeing tables) over others (those merely hallucinating tables). If so, the indifference principle is false.

Another test case was devised by the philosopher Adam Elga, who wrote a classic article called "Defeating Dr. Evil with self-locating belief". In Elga's thought experiment, Dr. Evil is about to press a button that will destroy the Earth. In response, the Philosophy Defense Force on Earth creates ten exact duplicates of Dr. Elga, each experiencing the same thing. They tell Dr. Evil that they have arranged that if any of the duplicates press the button, he will be tortured. According to the indifference principle, Dr. Evil should reason that he is ten times more likely to me one of the duplicates than the original Dr. Evil, so he should refrain from pressing the button. This reasoning is controversial. Many philosophers hold that in these cases, the original has reason to believe he is the original and not the duplicates. If so, the indifference principle is false.

I am sympathetic with the indifference principle, but I don't I need a strong version of it here. All I need is a weak and presumptive version saying that I should be indifferent among me-like beings (and endorse the key indifference principle as above) unless there is some special reason to favor some beings over others. I can then entertain possible reasons why I should favor my being a nonsim as opposed to being one of the more plentiful sims. In what follows, I'll entertain possible reasons deriving from the extra evidence that nonsims might have (under *Sims won't have our evidence*) and from the special position that simulators may be in (under *We know we're not the sims that we create*). I'll argue that upon examination, these considerations are not strong enough to greatly change our confidence that we are sims. As a result, these objections to indifference principles do not undermine the simulation argument.

(2) *Sims won't have our evidence!* 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, "Are You A Sim?"). 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 my 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 exist, 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, 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% percent 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 in "What is My Evidence that I Have Hands?" in Dylan Dodd and Elia Zardini, eds., *Scepticism and Perceptual Justification* (Oxford University Press, 2014) and Jonathan Vogel in "Internalist Responses to Skepticism", in John Greco, ed., *The Oxford Handbook of Skepticism* (Oxford University Press, 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 in "Unreasonable Knowledge" (*Philosophical Perspectives,* 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 percent 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.).

(3) *We know we're not the sims we create!* Here the objection is that there are at least some sims that we know we are not: namely, those we create. Suppose I have just created a trillion sims. If I know I have created them, then I know I am not them, since I can't create myself. So even if I know that sims greatly outnumber nonsims, I can't conclude from this that I'm probably a sim. In effect, being created by me is a sort of sim sign that I know those sims have and I lack.

Some people don't accept this objection. Nick Bostrom thinks that if I create a trillion sims that perfectly simulate me (so each of them also has the experience of creating a trillion sims), then after the moment of creation, I should then be much more confident that I am one of the created sims than that I am the creator.

Bostrom's reasoning parallels reasoning Elga's reasoning in the Dr. Evil thought experiment, in which Dr. Evil creates ten duplicates of himself. According to Elga, after this act of creation, Dr. Evil should reason that he is ten times more likely to me one of the duplicates than the original Dr. Evil.

As we've seen, this reasoning is controversial. Many philosophers hold that in these cases, the original has reason to believe he is the original. If so, then the original objection stands. Furthermore, the Bostrom-Elga reasoning doesn't obviously apply to creating sims that are different from me, which will presumably be common. If I experience creating a trillion sims, none of which experience creating any sims, then I can know that I am not one of the created sims. So the *We know we're not the simulations we create* objection still stands in cases like this.

(Some cases of creating sims different from me are tricky, however. If I experience pressing a red simulation button to create a trillion sims who each experience pressing a green simulation button, then my experience of red enables me to know that I am not one of those trillion sims experiencing green. On the other hand, for all I know I might be one of a trillion sims experiencing red created by a single nonsim experiencing green. Antecedently I don't have much reason to think that experiencing red correlates with being a nonsim. As a result, I might reasonably become confident that I am a sim—even though I know I'm not one of the green-button sims that I experienced creating. Something similar applies to any case of our creating humanlike sims, as I've discussed under objection (1).)

Even if we reject the Bostrom-Elga reasoning, however, this doesn't do much to undermine the main argument. To avoid the objection, we need only exclude sims that we create (or that are created by sims we create, and so on) from the reference class of humanlike beings in the argument. For example, we could say an *exogenous humanlike being* is a humanlike being that is not a sim created by us (directly or indirectly). We can then modify premises 1 and 2 by replacing *humanlike beings* by *exogenous humanlike beings.* (1. If there are no sim blockers, most exogenous humanlike beings are sims. 2. If most exogenous humanlike beings are sims, we are probably sims.) Both premises remain plausible, and premise 2 is no longer vulnerable to the objection under consideration.

That said, this version of the reasoning does make various potential sim blockers salient: not only familiar sim blockers such as *Nonsims will die before creating sims* and

Nonsims will choose not to create sims, but also the *We are alone* sim blocker (discussed in the notes) according to which we are the only nonsim population (or the only humanlike nonsim population) in the universe.

(4) *We don't know the physics of the next universe up!* A number of people have suggested that the simulation argument illicitly argues from premises about our universe (that we'll create simulations) to conclusions about what happens in the next universe up (that we were created). As the physicist Max Tegmark puts it in *Our Mathematical Universe* (chapter 12), "the computational resources of your own (simulated) universe are irrelevant: what matters are the computational resources in the universe where the simulation is taking place, about which you know essentially nothing."

In light of our discussion of the previous objection, we can now see that this isn't quite right. As I've formulated it, the argument turns only on a general claim about human-level intelligences: that they'll have the capacity to create sims, and that if they choose to, they'll create them. This isn't a claim specifically about our world. It's true that the evidence for this claim brings in a claim about our world: that our world has the capacity to support simulations, and that humans have the capacity to create them. But in this case the reasoning from our world to higher-level worlds seems fine. If our world has the capacity to support simulations, then any world in which our world is embedded also has that capacity. And if humans have the capacity to create simulations, so do other humanlike beings. So I think my formulation of the simulation argument is off the hook here.

It's true that there's a version of the simulation that turns directly on claims about us. (1) If there are no sim blockers, we'll create many sims. (2) If we create many sims, most humanlike beings will be sims. (3) If most humanlike beings are sims, we are probably sims. Therefore (4) If there are no sim blockers, we are probably sims.

We might call this the *first-person simulation argument*. It contrasts with the *impersonal simulation argument* that I have laid out, in which the first-person "we" in premises (1) and (2) is replaced by an impersonal "sims" (or "humanlike sims"). The first-person simulation argument is perhaps suggested by some remarks of Bostrom's: especially the conclusion of "Are We Living in a Simulation?", which says: "Unless we are now living in a simulation, our descendants will almost certainly never create an ancestor simulation", though other elements of the article (including Bostrom's formula for the fraction of human-type beings in a simulation) suggest the impersonal version.

The first-person simulation argument is at least potentially open to the charge of moving from premises about our universe (We'll create many sims) to conclusions about the next universe up (We're probably sims). This charge is perhaps strongest when combined with the previous objection: *We know we're not the sims we create.* If the sims we create are not even candidates to be us, it is not obvious how their existence increases the likelihood that we are sims. In response, a proponent of the first-person simulation argument can justify the inference by appealing to the Bostrom-Elga thesis that all beings that have experiences like ours are equally likely to be us. Given this thesis, the moment we experience creating a hundred sims, all of whom have the same experience as us, then we can reasonably infer that we are more likely to be a created sim ourselves.

Still, the Bostrom-Elga reasoning is highly controversial, and without it the inference in the first-person simulation argument may indeed be illicit as Tegmark suggests. The impersonal simulation argument provides an alternative way to run the argument that avoids this inference as outlined above.

(5) *We should expect to live in an impoverished world!* Physicist Sean Carroll has suggested that the complexity of our world undermines the original simulation argument. He first argues (along the lines of the previous paragraph) that if we accept standard assumptions of the simulation argument, most observers will live in impoverished universes where building many simulations is impossible. He observes that an analog of the simulation argument now suggests that we almost certainly live in an impoverished universe. But we do not! We appear to live in a complex universe where building many simulations is possible. So the simulation-argument reasoning must go wrong somewhere.

I think Carroll's argument isn't really analogous to the simulation argument. At best, it's analogous to the simple version of the simulation argument before we considered sim indicators, resting only on the premise that most conscious beings are sims. But our version of the simulation argument does more—it argues that most conscious beings *with experiences like ours* are sims. Carroll argues that most conscious beings live in impoverished worlds, but he doesn't argue that most beings with experiences like ours live in impoverished worlds. Our very experiences act as a rich-world indicator that rules out impoverished worlds. Most beings with experiences like ours live in rich worlds, not impoverished worlds. So in light of our experience, it is probable that we live in a rich world, not in an impoverished world. And as always, it is probabilities in light of our experience that matters.

Carroll’s argument is really a different sort of argument. His argument says that simulation assumptions should be rejected because they make our experiences *atypical* among conscious beings. It’s analogous to an intriguing argument that ants can’t be conscious. Suppose that ants are conscious. Then given the number of ants, the great majority of conscious beings will be ants (or simpler beings still). So probabilistically, we should expect to be ants (or simpler beings). But we are not ants! We have complex experiences. So something has gone wrong, and we should reject the assumption that ants are conscious.

The ant argument and Carroll’s argument are intriguing, but they are quite different from the simulation argument. They only work if we accept a *typicality thesis*, saying (roughly) that we should expect to be typical among conscious beings and that we should reject hypotheses that make us atypical. (The Doomsday argument we considered earlier does something similar.) I think the typicality thesis is far from obvious, not least because there are many reasons to think we are atypical in many ways. Importantly, the simulation argument doesn’t need anything like it. Even if we are atypical among conscious beings, the argument goes through. So Carroll’s interesting point doesn’t undermine the simulation argument.

(6) *If we’re in a simulation, evidence about our computer power may be misleading:* I’ve discussed this objection (due to Fabien Besnard, “Refutations of the Simulation Argument,” http://fabien.besnard.pagesperso-orange.fr/pdfrefut.pdf, 2004; and Jonathan Birch, “On the ‘Simulation Argument’ and Selective Scepticism,” *Erkenntnis* 78 (2013):

95−107) briefly in the main text, by noting that the objection only arises if we're already in a simulation. To put the point somewhat differently, we can reason: (1) either our evidence about computer power is heavily misleading, or it is not, (2) if our evidence about computer power is heavily misleading, we're probably in a simulation (as that's the most likely way for this evidence to be misleading), (3) if our evidence about computer power is not heavily misleading, we're probably in a simulation (by the original argument), so (4) we're probably in a simulation.

Still, I think *Sims take too much computer power* should be acknowledged as a potential sim blocker. Furthermore, the likelihood that simulations will be misleading does bring out that the simulation argument can easily be turned into an argument for skepticism about certain sorts of scientific knowledge, even if (as I will argue) it doesn't lead to global skepticism about the external world.

Shortcut Simulations [chapters 2, 5, 24]

To what extent can simplified models be used to simulate the behavior of macroscopic objects in a way consistent with all of our observations? To handle every *possible* observation of a system, simplified models won't be enough; a simulation in full detail will be required. But most actual systems are observed less closely than this. For example, if a bowl of ice slowly melts into a bowl of water with no-one watching for a day or so, a simple model specifying the water temperature a day later may suffice. If someone is watching as the ice melts, a more detailed model of the melting process will be required. If images are recorded for possible later examination and scientific analysis, a far more detailed model will be required.

Simulators seeking efficiency in modeling worlds like ours will presumably use models at different levels, depending on the level of observation involved. But if a system leaves many observable traces on systems around it (which may be the typical case), and those traces can be analyzed, similar issues will arise. It will be risky to use a simplified model to simulate a hurricane, for reasons like this. The effects of the simplified model will differ in subtle ways from the effects of a genuine complex hurricane, and these effects will in principle be analyzable in a way that could give away shortcuts in the simulation. If simulators have control over what sort of observations are made when, then this will give them much more leeway to use simplified models.

Julian Togelius has suggested to me that for related reasons, quantum mechanics may be a sim sign. There are versions of quantum mechanics suggest that reality only becomes determinate when we are conscious of it (see e.g. chapter 14). This is what one would expect in a just-in-time simulation where simulators only simulate what is necessary to explain sim's conscious observations. On the other hand, simulating an uncollapsed quantum wave-function may not be any easier than simulating a collapsed version.

Other Replies to External-World Skepticism [chapter 4]

A question in the 2020 PhilPapers survey asked: "Which is the strongest response to external-world skepticism?" The leading answers were pragmatic (23%), abductive (22%), epistemic externalist (19%), dogmatist (13%), contextualist (11%), semantic externalism (8%).

In chapter 4, I discuss three of these six responses: abductive responses (Russell), dogmatist responses (Moore), and semantic externalist responses (Putnam). I also discuss theist responses (Descartes), idealist responses (Berkeley), and verificationist responses (Carnap), which were not included in the PhilPapers Survey as they're not currently popular. My guess is that all would have popularity under 10%. (In a separate question, 7% endorsed idealism about the external world.)

The other three responses on the survey (but not discussed in chapter 4 for reasons of space) are pragmatic, contextualist, and epistemic externalist responses.

Pragmatic responses to skepticism (e.g., William James in "The Will to Believe," Susanna Rinard in "Pragmatic Skepticism") focus on the pragmatic impossibility of adopting skepticism in living one's life: One needs to adopt beliefs about the world in order to act. Pragmatic responses are often highly concessive to the skeptic, by conceding that we may not have good evidence or justification for our beliefs and may not have knowledge in the strict sense, while holding that we have pragmatic reasons to hold these beliefs anyway.

Contextualist responses to skepticism (e.g., Keith DeRose in "Solving the Skeptical Problem," David Lewis in "Elusive Knowledge") argue that *know* has different meanings in different contexts, so that (for example) "I know that I am not in a simulation" is true in ordinary contexts and but false in philosophical contexts. This line is also often concessive to the skeptic by granting that their view as expressed in philosophical contexts is true. In addition, the argument of chapter 5 makes a reasonable case that "I know that I am not in a simulation" is false even by the standards of ordinary contexts.

Epistemic externalist responses to skepticism (e.g., Alvin Goldman in *Epistemology and Cognition* (Harvard University Press, 1986) and Timothy Williamson in *Knowledge and its*

Limits (Oxford University Press, 2000)) focus on what's required for knowledge, arguing that this turns on external factors such as a reliable method of forming true beliefs, or appropriate evidence in one's own external environment. Our belief that we're not in a simulation may be reliably true and may be appropriately grounded in external evidence, at least if we're not in fact in a simulation. According to this line, people who aren't in a simulation may be able to know that they're not in a simulation. However, the arguments toward the end of the next chapter make a strong case against this line (see especially the discussion of externalism about evidence in the online appendix on further objections to the simulation argument).

We could perhaps try to run the Simulation Riposte against some of these lines. Sim James says, "I pragmatically have to believe that I'm not in a simulation." Sim DeRose says, "In the ordinary sense, I know that I'm not in a simulation." Sim Goldman says, "I have a reliable belief about simulations, so I know I'm not in a simulation." Sim Williamson says, "I have the spoon as evidence, so I know I'm not in a simulation." I don't know whether the Riposte really causes problems for these lines. However, I think the arguments of chapter 5 suggest that none of these lines can establish that we know we're not in a simulation.

Michael Heim and Philip Zhai on virtual realism [chapter 10]

As noted in the main text, the phrase *virtual realism* first appeared as the title of the American philosopher Michael Heim's important 1998 book on the ramifications of VR.

At the start of *Virtual Realism* (Oxford University Press, 1998), which is itself a sequel to his *The Metaphysics of Virtual Reality* (Oxford University Press, 1994), Heim says:

> Virtual realism is an art form, a sensibility, and a way of living with new technology. ... On one side are network idealists who promote virtual communities and global information flow. On the other side are naïve realists who blame electronic culture for criminal violence and unemployment. Between them runs the narrow path of virtual realism.

and later:

> Realism begins as a sober criticism of overblown, high-flown ideals. Yet at the core of realism is an affirmation of what is real, reliable, functional. Today we must be realistic about virtual reality, untiringly suspicious of the airy idealism and commercialism surrounding it, and we must keep an eye on the weeds of fiction and fantasy that threaten to stifle the blossom. At the same time, we have to affirm those entities that VR presents as our culture begins to inhabit cyber-space. Virtual entities are indeed real, functional, and even central to life in coming eras. Part of work and leisure life will transpire in virtual environments. So it is important to find a balance that swings neither to the idealistic blue sky where primary reality disappears, nor to the mundane indifference that sees in VR just another tool, something that can be picked up or put down at will.

As we can see, Heim used the label primarily for a broad social and political view of virtual reality, invoking realism as a label for social and political moderation.

However, he also associates the label with the view that “Virtual entities are indeed real, functional, and even central to life in coming eras.” I’m using the label *virtual realism* in the latter sense.

In his 1998 book *Get Real,* the Chinese-American philosopher Philip Zhai (now publishing as Zhai Zhenming) argues for a more metaphysical version of virtual realism and a version of the no-illusion view.

Quoting an early interview in which Jaron Lanier calls VR an illusion, Zhai says:

> What I am showing in this chapter is, however, exactly the opposite, that is, the virtual is no more illusory than the actual, since they are *reciprocal* in their relation to the core of personhood as the center of sensory perception. (p. 33; italics are Zhai’s.)

Zhai’s arguments rest on a “Principle of Reciprocity between Alternative Sensory Frameworks,” which says, “All possible sensory frameworks that support a certain degree of coherence and stability of perception have equal ontological status for organizing our experiences.” (p. 2)

As I understand it, Zhai’s principle is a broadly idealist (and more specifically phenomenalist) principle according to which the reality of a perceived object is determined by its coherence and stability with respect to other perceptual experiences. According to this sort of idealism, as I noted in chapter 6, we might say *stable and coherent appearance is reality.* Zhai outlines seven such rules for using stability and coherence to define reality (coherence within a single modality, coherence across modalities, temporal regularity, and others) and argues that all can be satisfied in VR. I reject this sort of idealism for reasons broadly similar to my reasons for rejecting

Berkeley's idealism in chapter 4. (See also my article "Idealism and the Mind-Body Problem", in William Seager, ed., *The Routledge Handbook of* Panpsychism (Routledge, 2019). Especially important is that we need reality beyond appearances in order to explain the stable and coherent appearances themselves. I think the case for the no-illusion view of virtual reality can be made even without idealist principles.

Apart from Heim and Zhai, some other authors whose work contains elements of virtual realism include David Deutsch (discussed in chapter 6) and Philip Brey (discussed in chapter 10). Elements of simulation realism are endorsed by Douglas Hofstadter (discussed in chapter 20) as well as in the articles by Andy Clark and Hubert Dreyfus in *Philosophers Explore the Matrix*. In addition, O. K. Bouwsma (chapter 6) and Hilary Putnam (chapter 20) show sympathy for a view akin to simulation realism without explicitly discussing simulations per se.

Structural, semantic, and symbolic information [chapter 8]

I'm using "semantic information" in closest to Carnap and Bar-Hillel's sense, where semantic information is a *content* (of a sentence, for example) and is thereby something in the vicinity of a fact or a proposition. Floridi defines semantic information as "well-formed meaningful data" or "well-formed truthful meaningful data." If well-formed data include structures of bits and the like, Floridi's semantic information may also include what I'm calling symbolic information, which we'll see can also be regarded in my taxonomy as a variety of concrete semantic information.

"Structural information" is sometimes used quite differently to mean "information about structure," where the structure in question is physical or geometric. See e.g., Emanuel

Leeuwenberg & Peter A. van der Helm, *Structural Information Theory: The Simplicity of Visual Form* (Cambridge UK: Cambridge University Press, 2013) and Mark Burgin & Rainer Feistel, "Structural and Symbolic Information in the Context of the General Theory of Information," *Information,* 8:4, 139 (2017). In this usage, structural information is a specific variety of semantic information.

"Symbolic information" is used in many different ways, but usually for something like *symbols encoding meanings*, which is at least somewhat close to my usage. Symbols are often understood to go well beyond structures of bits (as when an eagle is a symbol of liberty, for example), and there's also a use in "symbolic AI" where not all structures of bits that encode facts count as symbolic (distributed representations in artificial neural networks are subsymbolic, for example). In my usage, these distributed representations are still symbolic information.

Early uses of 'virtual world' and 'virtual reality' [chapter 10]

The first published use I've found of "virtual reality" in the current sense is in an interview with Lanier by Robert Wright in *The Sciences,* November-December 1987. Lanier tells me that there should be earlier uses.

The Australian science fiction writer Damien Broderick uses the expression "virtual realities" in something not far from its current sense in his novel *The Judas Mandala* (New York: Pocket Books, 1982), which involves a *Matrix*-like computer-generated environment, also called a "virtual matrix" in the novel: "Basically, we're the only dysentropic probability vector in these 'virtual realities': the ontology's plastic. There's a sort of consensual cocoon around us modifying our immediate environment synchronistically."

Broderick tells me that his use of VR in *The Judas Mandala* was somewhat different from the now-standard use:

> "Most Homo Sapiens dream away their collective lives in imaginary worlds under the somewhat sentimental stewardship of blazingly fast AGI/sapiens upload hybrids, which is exemplary VR. But while I describe it in terms not unlike Arthur Clarke's *The City and the Stars*, I reserved "VR" for a sort of alternative reality that only gifted organic sapiens can enter and, unobserved by the computers, plot to set humanity free again." [email from Damien Broderick, January 2021].

The American music theorist Richard Norton discusses a more distantly related notion of "virtual reality" in the context of Susanne Langer's theory of virtuality in art in his essay "What is Virtuality?" *The Journal of Aesthetics & Art Criticism*, 30:4, pp. 499-505 (1972).

At least in *Feeling and Form*, Langer uses "virtual world" primarily for the worlds depicted in fiction and poetry, but many others have used it more generally. Jaron Lanier says that Ivan Sutherland got the expression "virtual world" from Langer, so there's a clear line of descent from Langer's usage to the usage in contemporary VR. I haven't seen independent evidence of this, though, and so far I can't locate "virtual world" in Sutherland's writings. Sutherland is often quoted as saying, "The screen is a window through which one sees a virtual world. The challenge is to make that world look real, act real, sound real, feel real," in his 1965 essay "The Ultimate Display." In fact, these sentences come from a much later "paraphrase" by Frederick Brooks in "What's Real About Virtual Reality?" (*IEEE Computer Graphics & Applications*, November/December 1999). The closest that can be found in Sutherland's essay is "A display connected to a digital computer gives us a chance to gain familiarity with concepts not realizable in the physical world. It is a looking glass into a mathematical wonderland."

Free will in the experience machine and in virtual reality? [chapter 17]

The source of many reservations about the experience machine was that the experience machine is preprogrammed. Everything is scripted in advance. As a result, in the experience machine we seem to lack a certain sort of free will. We are not really living our life. Instead it is being lived for us.

These questions are worth raising in their own right. First, do we have free will in the experience machine? Second, do we have free will in virtual reality?

The problem of free will is often posed by asking: how could we have free will in a deterministic universe? On many physical theories, the laws of nature are deterministic. What happens at any moment in time is fully determined by the previous state of the world and the laws of physics. Everything that ever happens is determined in advance! The same applies to human brains and to human behavior. In a deterministic universe, everything we do is determined in advance. So how could we have free will?

Some physical theories, such as quantum mechanics, are nondeterministic. They have a probabilistic element. In the famous two-slit experiment, a particle may go through one slit and it may go through another, and which way it goes is not determined in advance. The same may be true for human brains and behavior. Some have looked to quantum mechanics for a way to save free will, but it is not clear that it helps. It just adds a random element on top of deterministic processes, akin to rolling quantum dice, and it is not clear why random processes should be any better than deterministic processes at supporting genuine free will.

There are at least three different philosophical reactions here. *Libertarians* hold that we have a special sort of free will that is inconsistent with determinism. *Hard determinists* hold that the universe is deterministic, and that as a result we do not have free will. *Compatibilists* hold that free will and determinism are compatible, so that even in a deterministic universe (or a quantum universe with determinism plus randomness), we can have free will all the same.

I can't settle which of these views is right here. Compatibilism is the most popular view among philosophers, but all three views have many supporters. My own view is that it depends on exactly what you mean by free will. If you mean something fairly weak (like the ability to do what one wants to do), then free will is certainly compatible with determinism. If one means something much stronger (like the ability to choose one's own nature), than free will may not be compatible with determinism. For present purposes, I'll stay neutral on this. Instead, I'll argue that whichever of these views is true, then *if* we have free will in ordinary physical reality, we have free will in virtual reality.

The easiest way to see this is to observe that at least with existing VR devices, our actions in VR are always brought about by physical actions in the physical world. One moves in the physical world, or one presses a button, and a virtual action results. *If* the physical action is brought about by free will, then the virtual action is too. Suppose we face a choice about whether to go left or go right in the virtual world. We decide to go left, and press a button on a controller to make this happen. If the button press (the physical action) results from free will, then going left (the virtual action) results from free will do.

Even for future VR involving brain-computer interfaces, something similar applies. Brain-computer action interfaces might work by reading brain activity in the decision or action areas of the brain, and producing a corresponding action in the virtual world. If the ordinary processes by which we come to a decision involve free will, then the processes by which we come to a decision in VR will presumably involve free will as well.

It is true that some VR environments offer less opportunity to exert free will than others. For example, in some video games, there are only limited opportunities to make choices. For long periods between choices, things may unfold without the user giving any input. Furthermore, many videogame virtual worlds have a highly constrained progression through levels. One may have choices within levels, but there is no escaping the pre-determined progression.

One might say that in VR environments of this sort, one has less *freedom*. But freedom is not exactly the same as free will. A prisoner in a jail cell lacks freedom (they cannot do as they please, and they have many limitations on their options), but they still have free will (they can choose their own actions, at least if anyone does). Their actions are still under their control. Likewise, even in constrained video game environments, we exert free will.

Importantly, many virtual worlds are not like these constrained video game environments. Virtual worlds like *Second Life* are open-ended, without any strongly constrained pathways. People can make their own lives in these virtual worlds. There may still be limits on what one can do, but there are also limits in the nonvirtual world, imposed both by the laws of nature and by the laws of society. Some virtual worlds may even avoid the limits of the nonvirtual world: perhaps one can fly (breaking the nonvirtual laws of nature) or one can ride in a self-driving car (breaking the nonvirtual laws of society). In principle, one can have as much freedom in a nonvirtual world as in a virtual world.

Of course, someone might say that the brain is deterministic and all our actions are determined in advance, and that this is incompatible with free will. If this is right, both our physical actions and our virtual actions will be determined in advance, and we won't have free will in either domain. But at least virtual reality will be no worse off than physical reality.

What about the experience machine? Can one have free will in there? We have seen that it is preprogrammed; but is this any worse than the world being deterministic?

At this point, much depends on just how the experience machine works. Nozick left its inner workings quite unclear, and it is not at all easy to see how it could work as described. The basic worry is that the experience machine is scripted in advance, but one still has the feeling of making choices. What happens if the user decides to go right, and then the script says to go left? In many cases the user will experience dissonance, as if they are not in control of their actions, and the experience will be quite unlike what is intended. To avoid this, the user's decision and the script must always be in perfect alignment.

One route to alignment is through *advance testing*: fine-tuning the script in advance based on knowledge of the subject's brain. It is not easy to see how this would work. If the brain's processing is indeterministic, it seems that there is no way to avoid some mismatch between its decisions and the script. If the brain is deterministic, however, perhaps simulation technology could help.

What I call a *Replay Experience Machine* works by replaying previous simulations (see also the two-brain scenario in chapter 24). Perhaps one first asks the subject what sort of experiences they would like, and then one scans their brain (and temporarily freezes it) and runs millions of simulations of how the brain will react in different environments. (Let's conveniently ignore the question of whether these simulations would themselves be conscious.) Eventually one will find an environment where things go just as the subject wants. One then unfreezes the brain and connects it to the chosen virtual environment, replaying the environmental simulation. If the brain is deterministic and the simulation was accurate, the brain should have just the experiences that the subject wanted.

An alternative route is *brain manipulation*: controlling the subject's brain so that it is aligned with what happens in the script. A relatively simple route may be a *Choice Blindness Experience Machine*, where we suspend the subject's critical faculties so that she never notices when her decisions are put into effect. Choice blindness is already a known phenomenon (P. Johansson, L. Hall, S. Sikstrom, "From change blindness to choice blindness", *Psychologia*, 2008). Sam is asked to choose which of two people (Robin or Sydney) she finds more attractive. If she chooses Robin, experimenters manipulate the results as if she had chosen Sydney, and ask her to explain why she chose Sydney. Much of the time, the subject does not notice, and goes on to explain the choice they did not make. Perhaps an extension of this phenomenon -- a drug that greatly enhances choice blindness? -- could be used to build a Choice Blindness Experience Machine.

Alternatively, a *Brain Control Experience Machine* could exert direct control over the subject's brain, in effect controlling the subject's beliefs, desires, and decisions. If done well enough, the subject would still have the feeling of being in control, even if they are not. These last two options involve brain interference that goes well beyond standard virtual reality, but they might at least make for a seamless scripted experience machine.

Do these versions of the experience machine eliminate free will? It's arguable that in the Replay machine, the subject still has free will, at least if she does in ordinary reality. Her brain is reacting to the environment much as it would in ordinary reality. It is true that her environment is very carefully selected, and others may know in advance what she will do. Perhaps she is not truly free, because everything is so carefully planned and controlled.

Nevertheless it seems as if she is still making decisions in the same way she makes ordinary decisions, involving the same amount of free will.

In the Brain Control machine, on the other hand, the subject's decisions are manipulated. In the Choice Blindness machine, many of the subject's actions are not under the subject's control at all. If someone else chooses what one decides, then this seems a serious breach of free will. If someone else chooses what actions one's decisions lead to, that is also a breach of free will. So I think these brain-manipulation experience machines do involve a serious threat to free will.

These preprogrammed experience machines may make an excellent form of entertainment, but they do not have everything we value in life. Most of us value the ability to make our own decisions, and we value coping with a world that has not been entirely planned for us in advance. So I think Nozick is right that we should reject a life in these experience machines. At the same time, none of this is reason to reject life in virtual reality.

Donald Hoffman's case against reality [chapter 23]

[First two and last two paragraphs are from chapter 23.]

[In his recent book *The Case against Reality: Why Evolution Hid the Truth from Our Eyes*, the cognitive scientist Donald Hoffman makes an evolutionary case for skepticism. He argues that evolution doesn't care about whether our beliefs about the world are true. It just cares about whether we're fit—whether we survive and leave offspring. He also argues that there are many more ways for our beliefs to be massively false than for them to be true, so we should expect most of our beliefs to be false: The world is almost certainly not as it appears.

Hoffman's argument assumes something like an Edenic model of perception. It's true that the Edenic content of our beliefs is likely to be false. We can't know that an apple is Red or a ball is Spherical. But once we move to a structuralist conception of perception and reality, our model of reality is robust. We can be much more confident that an apple is red or that a ball is spherical. We can no longer conclude that reality is not as it appears.]

Hoffman starts by observing that there are many different ways to connect properties in the world to different sorts of perception. If there are ten different colors and ten different color perceptions, there will be a vast number of ways that a visual system might hook up the colors to the color perceptions. Crucially, Hoffman assumes that just one of these many ways leads to correct perception and the others are all incorrect. He argues that the incorrect methods are just as likely to be evolutionarily fit as the correct method. As a result, evolution is much more likely to produce incorrect perceivers than correct perceivers. He draws the conclusion that there's only a tiny chance that our overall perception of the world is correct.

This is a fascinating argument, but I don't think it works. It goes wrong at the point where Hoffman assumes that only one way of connecting colors to color perceptions leads to correct perception. This assumption would be correct if colors were objective Edenic properties, out there in the world. But it's incorrect if colors are understood in terms of their functional role.

Suppose we think of redness as the power to normally cause reddish experiences. Then whatever objects in the environment that get connected to reddish experiences will have the power to normally cause reddish experiences, so they'll count as red. As a result, all ten different strategies for connecting objects in the world to reddish experiences will lead to correct experiences and true beliefs.

The same goes for space. My former Ph.D. student Brad Thompson devised a thought experiment about Doubled Earth, where everything is twice as large as on Earth. When Brad sees an object one meter tall, Doubled Brad sees an object that we would say is two meters tall. But Doubled Brad is isomorphic to Brad, so when he sees this object he has the same sort of experience that Brad has when seeing a one-meter-tall object. Is Doubled Brad suffering from an illusion, seeing the object as half as large as it actually is? Perhaps Hoffman will say yes. On an Edenic model with absolute sizes, one would say yes. But most people find it far more intuitive to say that both Brad and Doubled Brad are correctly perceiving the world. And certainly this is what spatial functionalism will predict. Once again, Hoffman's assumption that there's only one objectively correct mapping from world to mind seems implausible, at least after we have fallen from Eden.

Now, Hoffman might argue at this point that we have given up on objective reality. I don't agree—we have merely given up on the Edenic model of objective reality. There is still an objective world of structures out there. Objects in the world really do have colors and sizes. It's just that what makes them the colors and sizes they are is partly the roles they play and not some absolute, intrinsic nature.

Hoffman might try to argue that we can't even know about structure. Maybe he could run his rewiring argument for numbers: When there are two balls in the world, we could experience three balls, and vice versa. But it's easy to see this can't work. We'll remove one ball from what looks like a pile of two balls and it will look like there are now three balls! So Hoffman's arguments pose no obstacle to knowing about structural aspects of the world.

[That said, I agree with Hoffman where our perception of an Edenic world is concerned. This perception doesn't latch on to reality. There are no Colors and Sizes in the external world. I can even happily endorse Hoffman's idea that Edenic qualities serve as a sort of "interface" in perception. In effect, we're presented with an Edenic world that serves as a useful guide to the structure of the true external world, even though the true external world is not itself Edenic.

I diverge from Hoffman on the idea that perception doesn't tell us anything about the true nature of the external world. It tells us about the colors and sizes of things just fine. Knowing about colors and sizes may not tell us about the intrinsic Colors and Sizes of things, but it still tells us a great deal about the structure of external reality.]

Novels, fictions, and experience worlds [chapter 24]

What about novels and other fictions? Do events in these really take place in the head of the author or reader? I would say usually not. A reader's mind will not usually have anything like an interactive world-model. An author's mind may contain more of a model, but in many cases the model may often be more like a script building toward an outcome than a genuine open-ended and interactive world. For some authors in some cases, writing a novel may unfold as a full-scale interactive simulation. In that case, the events could have at least the limited mind-dependent reality of the events of a dream.

Interactive novels are a special case. In most existing interactive novels, the interaction is too intermittent for this to involve anything like a virtual world. However, a highly interactive novel would approach something like a text adventure game. *Colossal Cave Adventure* involves a genuine virtual world: It's interactive and computer-generated, with its state encoded in a database of virtual objects, even though it's not immersive. Someone playing *Colossal Cave Adventure* is genuinely interacting with a virtual world. The same goes for the virtual worlds involved in games such as *Dungeons and Dragons*, which are traditionally realized in the notebooks, props, and memories of participants; see Jon Cogburn & Mark Silcox, eds., *Dungeons and Dragons and Philosophy* (Chicago: Open Court, 2012). Even if there's no computer here, there's something akin to a virtual world.

Ordinary interactive novels and games don't really raise a skeptical issue. We can plausibly know that we're not in an ordinary *Dungeons and Dragons* game, since those games would not support our detailed perception. One could perhaps make a case that we're in an unusually rich version of the game that models much of our perception and the physical world. But this brings us back to a more standard version of the simulation hypothesis.

Experience World. Here's one more empty-world hypothesis. Let Ordinary World be a world like ours. Then let Experience World be a world containing only states of consciousness, with one law of nature: The states of consciousness in Experience World at time *t* are the same as those in Ordinary World at time *t* (where Ordinary World is specified by its laws of nature and initial conditions). Then beings in Experience World will have experiences just like Ordinary World, but there will be no external world there.

To respond: I'm not sure that there could be a law of nature like this. If there can be such laws, they're certainly more complex than the laws of Ordinary World, so there's a simplicity case against the hypothesis that we're in Experience World. I'd also argue that for this law to work, Experience World needs states that reflect the states of Ordinary World. Once we have those, Experience World is no longer a world with just conscious states; it's a world where conscious states interact with an external world.

Markus P. Müller describes a cousin of Experience World in “Law without law: From observer states to physics via algorithmic information theory,” *Quantum* 4, 301, (2020). In Mueller’s ingenious framework, observations evolve from other observations by a single law in algorithmic information theory. Roughly: The probability of the next observation being A is the algorithmic probability of A given earlier observations, which is determined by the length of the shortest algorithm that produces previous observations and A. It’s highly unlikely that Mueller’s framework would produce even the appearance of an external world, as opposed to a regular parade of internal experiences. In any case, perhaps one could argue along the same lines as for Experience World: Using algorithmic probabilities in a law of nature requires use of the relevant algorithms that will then support an external world.