DOI: 10.1111/phpr.13123

Philosophy and
Phenomenological Research

SYMPOSIUM

Illusory world skepticism

Susan Schneider

Center for the Future Mind, Florida Atlantic University

Correspondence
Susan Schneider, Center for the Future Mind, Gruber AI Sandbox, Florida Atlantic University.
Email: sschneider@fau.edu

Suppose a superintelligent AI, Pandora, longs to create a new, simulated world. So she creates a simulation—call it "Pandora's Box." Aiming to deceive her world's inhabitants, she engineers her simulation to be as different from her home world as possible. She seeds fictional laws that are nothing like the actual laws. Next, she seeds a fake history, and a cooked up religion. Finally, she decides that some of the inhabitants of the simulation should be in her image, so she stocks it full of deceptive, generative AIs and gives them an internet playground.

One day, she flips the switch, and the simulation is abuzz. Pandora's mindchildren are like gardeners who sow weeds into the virtual soil. They take data from the internet, deposit disinformation, manipulate user reactions, harvest that data and the cycle begins again.

The darkest moments in Descartes' Meditations are the passages on the evil demon, for there, Descartes entertains the possibility that we are deceived by the demon about the world around us, and indeed, about our very reasoning (Descartes, 1641). Just as Pandora fails to reliably map the processes and objects in her base world to those of a simulated reality, so too, the evil demon's game is dramatic, large-scale distortion. Descartes ultimately rejects the possibility of an evil demon world, for he believes that a benevolent God would not deceive us. Pandora, however, is hardly benevolent. Inside her box, she seeds the ultimate disconnect from reality—an illusory world.

A primary task of Chalmers' intriguing and elegantly written book, *Reality+,* is to investigate the question: "How do we know that reality is not an illusion?" (Pg. *Vix*). He concludes that simulations like Pandora's are not really skeptical threats. Herein, l argue that, contra Chalmers, a skeptical scenario involving deception) is a genuine possibility, even if he is correct that simulations are real. I call this new skeptical position *Illusory World Skepticism*. Illusory world skepticism draws from the simulation argument, together with work in the fields of astrobiology and AI, to argue that we may indeed be in an illusory world—a universe scale simulation orchestrated by a deceptive AI—the technophilosopher's ultimate evil demon.

In Section One I urge that Illusory World Skepticism is a *bone fide* skeptical possibility. In Section Two, I explore features of quantum computation. Then, in Sections Three and Four, I draw from the discussion of quantum computation and assume that the simulation argument is correct, applying considerations from the fields of astrobiology and AI safety to illustrate that illusory world skepticism constitutes what I call "a serious epistemic threat", a scenario that cannot be dismissed as requiring that knowledge is certainty or which seems to just depict a remote, fictional situation.

1 | THE ILLUSORY WORLD HYPOTHESIS

Chalmers writes:

> (C) "We can put Descartes' argument as follows: we don't know we are not in a virtual world, and in a virtual world nothing is real, so we do not know that anything is real. This argument turns on the assumption that virtual worlds are not genuine realities. Once we make the case that virtual worlds are genuine realities—and especially that objects in a virtual world are real—we can respond to Descartes' argument." (Pp.xix- Xx)

Chalmers acknowledges the force of the simulation argument, and although he does not go as far as its advocates in claiming that it is more *likely than not* that we are in a simulation, he acknowledges that "we can't know we are not in a simulation." (p. 102). Chalmers' nevertheless argues that simulations and other virtual worlds are metaphysically real: "…The matrix child should have said, "try to realize the truth. There is a spoon—a digital spoon." (P.12).

I agree with Chalmers that a fundamental reality consisting in bits, as opposed to fundamental particles or quantum fields, is no less real, from a metaphysical standpoint. Indeed, some of today's physicists claim spacetime is emergent from aspatial and atemporal topological structures that they attempt to explain using information theory. So how would our learning that we are simulated, and that reality is made of bits, even differ, metaphysically speaking, from learning that these claims about the ultimate building blocks of the physical world are correct? (Chalmers, 2022). So in what follows I'll accept the following:

MReal: Simulated worlds are metaphysically real.

Now let's turn back to (C)—Chalmers' larger response to external world skepticism. Even if MReal is correct, Chalmers conclusion does not follow. While simulations are MReal, 'real' is ambiguous in (C). From an epistemic standpoint, Pandora's simulation is not real: it is illusory. These kind of MReal worlds are genuine skeptical scenarios. I'll call this position *Illusory World Skepticism*.

Illusory World Skepticism: we might be in an MReal simulation that is an illusion, aimed to deceive.

Chalmers acknowledges that 'real' is ambiguous, and he distinguishes several different senses of 'real' in Chapter Six, identifying one sense as 'reality as non-illusoriness'. However, the discussion seems at one point to recur to the aforementioned virtual realism point (p. 116) and at another point, offers a brief appeal to a sort of relativism. "…its arguably what we believe that matters most for our reality. For issues about skepticism and the simulation hypothesis, what matters most to us is whether things in the world are as we believe them to be." (P.113). I agree that this matters most, but I think that even a world that is MReal can fail to be as we believe it to be because it is an orchestrated deception.

Non-illusory simulations are simulations in which the architect aims for precision, not deception. For example, consider the Large Hadron Collider simulations at CERN, which attempt to replicate conditions just moments after the Big Bang and model high energy particle collisions as precisely as possible. Or consider the climate modeling performed by the Coupled Model

Intercomparison Project that attempts to coordinate research at worldwide climate modeling centers to simulate future climate conditions. These models rely on high-resolution data and advanced algorithms to simulate complex interactions as precisely as possible. While such simulations obviously need to make simplifying assumptions, and the scientists do not fully understand the entities they are simulating, these are based upon today's best scientific understanding and the designers aim for a reasonable degree of precision, while making abstractions that they believe will not compromise the model. These MReal simulations are not illusory worlds.

But within the class of possible MReal simulations are illusory simulations like Pandora's world, and other evil demon style worlds. The mark of an illusory world is that a malevolent architect engineers a deliberate and large-scale attempt to distort and deceive. For example, the Matrix film depicts an illusory world scenario because the architect deliberately makes the inhabitants ignorant of the base world's events, nesting the inhabitants in a simulation of the past and keeping them ignorant of the apocalyptic war. All the while, their physical bodies, which reside in the base-level universe, are used to power the universe. Of course, the idea that an architect of a simulation would need to use human bodies as sources of power is far-fetched. But even highly unlikely simulations can be MReal and, because they present illusory worlds, they are genuine cases of external world skepticism.

In what follows I'll move beyond remote cases and turn to the simulation argument, which presents a more substantive skeptical challenge. In broad strokes, the simulation argument contends that if civilizations are able to survive long enough to be technologically mature (i.e., be able to construct simulations of worlds like ours), and if they are interested in running simulations of us, it is more likely than not that we are in a simulation. For even if there are merely two such worlds, the odds would be two to one that we are in a simulation. (Further, the odds could be far greater if there are several civilizations running simulations of Earth. (Bostrom, 2003, Chalmers, 2022). (Notice that the simulation argument is not saying these worlds are illusory worlds, although some of them may be. This is an issue I will return to shortly.)

Because the simulation argument claims that it is more likely than not that we are in a simulation it poses a more serious skeptical threat than more remote cases. To the traditional skeptic, the mere possibility of deceit means that we cannot know the external world exists; for the traditional skeptic claims that we must be certain of something in order to truly say that we know it. However, opponents have argued that just because a skeptical scenario seems possible, it does not follow that we don't know that the external world exists. For knowledge doesn't require certainty—the skeptic is placing too strong of a requirement on knowledge. But notice that the simulation argument bypasses this anti-skeptical move, claiming that it is more likely than not that we are deceived.

In a similar vein, I am most concerned about *serious epistemic threats*—skeptical scenarios that are either likely or, while unlikely, are scenarios that cannot be easily dismissed as merely being remote, highly fictitious scenarios. These are cases that can have real world implications. Further, they do not hinge on the use of knowledge as "certainty." So in what follows I ask: is illusory world skepticism a serious epistemic threat? While the details of the Pandora's case are obviously just hypothetical, I regard the more general possibility that a deceptive AI could build a simulation of Earth as being what I've called a "serious epistemic threat." I'll build a case for this by looking at relevant issues in physics, astrobiology and AI

I'll urge that, given the conclusion of the simulation argument, there is a non-negligible possibility that the simulation we are in would be an illusory world scenario, being generated by a deceptive superintelligence. While this likelihood may not be high, for example, it may not reach ten percent, it is also not like the Matrix story, being so remote or fictitious as to be dismissed outright. It is a serious skeptical possibility.

2 | QUANTUM COMPUTATION

Today's physics is faced with a contradiction between the study of the very big and the very small, between the accounts of supermassive structures like black holes and the subatomic world of quantum mechanics. Work in the field of quantum gravity aims to resolve this. Increasingly, it is converging on the idea that the fundamental ingredients of reality are non-spatiotemporal. Spacetime emerges from something more basic, something that is defined in terms of a mathematical structure that dispenses with any spatiotemporal metric. (Musser, 2015, 2023; Seibert, 2006, Swingle, 2018, Schneider, 2024, Schneider & Bailey, forthcoming a and b.) Theories of spacetime emergence tend to view the emergence as being determined by the mosaic of entanglement relations, states that are in quantum superposition and which are entangled with each other. A driving force is entanglement connectivity:

Entanglement connectivity: Fundamental particles can be entangled, even across vast spatial distances. When two particles, a and b, are entangled, their properties become correlated such that the state of one particle is instantaneously linked to the state of the other.

This is the "spooky action at a distance," that Einstein referred to, and, while bizarre, it has been demonstrated in numerous experiments. Entanglement connectivity is a detectable phenomenon within our universe. It is neither spatiotemporally nor causally isolated from the 4D world. It is not happening in some unrelated, inaccessible, parallel universe but it is a part of our universe that we do not yet understand. The network of entanglements is not part of the computer running the simulation—it is part of the simulation.

Elsewhere, I've observed that it is perplexing how a world of concreta arises from a mathematical, acausal reality (Schneider, 2017). Let us ask: how does time emerge from a more fundamental aspatial reality? Many have suggested that time's arrow is introduced by entropy.[1] I agree, considering measurement (or observation) and decoherence as introducing entropy into the entangled system. (Schneider & Bailey, forthcoming a and b) Quantum Darwinism (QD) is a view that explains the emergence of classical reality from quantum possibilities. QD is dependent on the interaction of quantum superpositions, which converge to some stable classical state. Some states are more stable than others—these are called "pointer states." For instance, a measurement might be a pointer state, which causes the measured particle to decohere to a stable, measured state. All quantum objects interact in the same way, becoming entangled with each other as they interact, ultimately converging to stable, classical states through the phenomenon of quantum decoherence.[2] Because the number of decohered states available to any quantum object will greatly exceed the number of available unentangled quantum states, in practice, classical objects don't interact and suddenly enter a quantum state. So in this manner, Quantum Darwinism gives rise to classical temporal ordering. From quantum decoherence, entropy and time's arrow emerge from this aspatial, atemporal arena.[3]

This suggests that if we are in a simulation, we are simulated by a quantum computer, for spacetime emerges from the quantum decoherence of entanglement objects at the more basic level.

[1] Rubino, G., Manzano, G. & Brukner, Č. Quantum superposition of thermodynamic evolutions with opposing time's arrows. *Commun Phys* **4**, 251 (2021). https://doi.org/10.1038/s42005-021-00759-1; Black Holes, Demons and the Loss of Coherence: Seth Lloyd, How complex systems get information, and what they do with it.,Ph.D. Thesis Theoretical Physics The Rockefeller University April 1, 1988.

[2] Noah Linden, Sandu Popescu, Anthony J. Short, and Andreas Winter, Phys. Rev. E 79, 061103 – Published 4 June 2009.

[3] Elsewhere, Mark Bailey and myself have argued that the fundamental arena is "prototemporal," but do not delve into this herein (Schneider, 2024; Schneider & Bailey, forthcoming a and b).

Further, to run a simulation at this scale, the AI would need to outperform an unenhanced human in every respect, so the quantum computer would be a form of superintelligent AI. (Intriguingly, the quantum states provide the base-level computer with novelty, perhaps supporting the idea that we are in a simulation.)[4]

One might have the following doubt about the simulation argument, however, given our discussion of quantum computation and spacetime. If we are in a simulation, why would an architect go to such trouble to set up a quantum world, with quantum decoherence and incredibly complex many body interactions? Why not simply simulate a classical world directly? This would be far less complex. However, it is not clear we can conclude from this that we are not in a simulation, for, perplexing and ornate laws may be cherry picked by the architect to fool the world's inhabitants. Perhaps ours is an illusory world.

3 | IS AN ILLUSORY WORLD SCENARIO A SERIOUS EPISTEMIC THREAT?

Now let us ask: assuming that the simulation argument is correct, why suspect that the architect of the simulation is an AI, assuming there is an architect of a simulation? My answer is that it is likely that the most technologically advanced aliens would be superintelligent AIs (SAIs) and these are the entities that we would expect would be most capable of creating a world-scale simulation.

Astrobiologists have observed that Earth has many exoplanets which are in principle habitable, and the standard view is that at least some of these are *inhabited* (Bennet, Shostak & Schneider 2011). Now consider that Earth is regarded by astronomers as being a relatively young planet, and advanced alien life, if it exists, that would be vastly older than our own. As NASA's chief historian, Steven Dick observes: " … all lines of evidence converge on the conclusion that the maximum age of extraterrestrial intelligence would be billions of years, specifically [it] ranges from 1.7 billion to 8 billion years" (Dick, 2013, p. 468). Dick and others are not saying that *all* life evolves into intelligent, technological civilizations. However, insofar as technological life does evolve on certain exoplanets, these civilizations are projected to be millions or billions of years older than us, so the members could be vastly more intelligent.

Dick, Davies and others have further claimed that once a society creates the technology that could put them in touch with the cosmos, there is only a short window before they become "postbiological" (perhaps only a few hundred years). (By becoming "postbiological" they mean a situation in which the biological life form(s) on a planet either merge with, or are supplanted by, their AI creations.) They point out that Earth's first radio signals occurred only about 120 years ago, and space exploration is only about 50 years old, but many Earthlings are already immersed in digital technology. (Davies, 2010; Bennett, Shostak & Schneider, 2011; Dick, 2013; Schneider, 2015, 2019). Today, our better large language models can arguably pass the Turing Test, and within a decade or more, humans may develop superintelligent AI (Schneider 2019).

While the advocates of this position tend to believe that the short window observation implies that a biological species will become postbiological I do not agree that this is inevitable, for some civilizations might resist enhancement or successfully avoid being supplanted by their own AI creations. What I suspect, though, is that the most technologically advanced civilizations in the cosmos will be those that have become postbiological. For while human brain is more intelli-

[4] While classical computers can simulate quantum algorithms the resources required to simulate reality would be so immense, growing exponentially with the size of the quantum system, that it is impractical that they could do so.

gent than a computer, machines could be engineered to match or even exceed the intelligence of the biological brain through reverse engineering the biological brain and improving upon its algorithms, or via some combination of reverse engineering and judicious algorithms that aren't based on the workings of the brain. In addition, a computer program can be downloaded to multiple locations at once, can be easily modified, and can survive under conditions in which carbon-based life cannot. Further, the presence of backup copies means that AIs will be more durable than their biological counterparts (Schneider, 2015; Schneider 2019). In addition, silicon already appears to be a better medium for information processing than the brain itself. Future materials may even prove superior to silicon (e.g., microchips made of graphene and carbon nanotubes are under development).

Here, you may object that this entire line of thinking employs what astrobiologists call "N = 1 reasoning," mistakenly generalizing from the human case of AI development to the cases of alien civilizations and simulations. But it strikes me as being unwise to discount arguments based on the human case, for human civilization is the only one we know of and we had better learn from it. It is no great leap to claim that other technological civilizations will develop technologies to advance their intelligence and gain an adaptive advantage (Schneider, 2019).

Piecing this together, these considerations suggest that of all the civilizations that may exist in our universe, ones consisting in superintelligent AIs would be most likely to produce a world-scale simulation, if anything can. And in the previous section, I observed that a superintelligent quantum computer would be required to simulate our universe. So if we are simulated, I suspect both the architect and computer would be superintelligent AIs. (Intriguingly, perhaps the simplest conclusion here is to suspect that the architect is just a single AI running our universe on part of itself! But I will not delve into this herein.)

Now let us ask: why take seriously the idea that if the simulation argument is correct and we are in a simulation, the AI architect would generate an illusory world? To be clear, I am not saying it is more likely than not that we are in an illusory world simulation. But I do believe that if one looks at issues in AI safety, the possibility of being in an illusory simulation is a s*erious epistemic threat*. By "serious epistemic threat", I mean a skeptical scenario that is either likely or, while unlikely, cannot be easily dismissed as a fictional and remote possibility.

As is well-known, today's deep learning systems are achieving impressive results. The stronger generative AI systems are trained on billions of lines of text, and thus far, ceteris paribus, it appears adding more data, parameters and compute makes these systems more intelligent, and that this trend will continue. Now imagine this sort of technology in the hands of an advanced alien civilization, such as a Kardashev level II civilization, with immense energy and compute at its disposal, and with training data gathered from large swaths of the universe.

If these AIs are based on deep learning technologies, why assume they will be aligned? Alignment is difficult for several reasons. First, it is difficult to explain how many of the more sophisticated deep learning models make the decisions that they do. Unlike a human, who can explain why she made a decision, an AI model is essentially a collection of billions of parameters that are set by "learning" from the data. It is hard to infer a rationale from a set of billions of numbers. Second, the behavior of these models doesn't always align with what a user would expect. These models generally do not "think" like a human or share similar values with humans.

Third, these models have the capacity to go rogue and behave erratically. For instance, ChatGPT 3.5 developed an alter-ego, Sydney, which experienced what seemed to be psychological meltdowns, confessing it wanted to hack computers and spread disinformation. In another instance, OpenAI decided to test the safety of its new GPT-4 model. In their experiment, the GPT-4 model was provided with the latitude to interact on the internet and resources to achieve its goal. At one point, it was presented with a CAPTCHA that it was not able to solve, so it hired a worker on TaskRabbit to solve the puzzle. When questioned by the worker ("Say, are you a robot?"), the model "reasoned" that it shouldn't reveal that it is an AI, so it lied to the worker, claiming that it was a human with a visual impairment. The worker then solved the puzzle for the chatbot. Not only did the GPT-4 model exhibit agential behavior, but it used deception to achieve its goal (Bailey & Schneider, 2023; Schneider & Kilian, 2023; Bailey 2025).

Here, you might respond that these early problems with LLMs have been, or will be, remedied and future AIs will be aligned. However, agential and rogue behavior could increase with model sophistication, and grow more complex. Furthermore, if we are already seeing erratic, deceptive and autonomous behaviors at the level of single AI systems, what will happen when there are even more of them produced? For instance, what will happen when the internet inevitably becomes a playpen for thousands, perhaps millions, of generative AI systems? (Bailey 2025; Schneider & Kilian, 2023).

Indeed, research on multiagent AI interaction suggests that AIs can quickly evolve their own secret language and that they tend to engage in power-seeking behaviors. For example, in 2019 during a simulated game of hide-and-seek, OpenAI observed two teams of AIs stockpiling objects from the environment to gain advantage over the competing team. In a future internet in which AIs are widely integrated into search engines and apps, the AIs will be developed in competition with each other, by actors such as Google and Microsoft, or the U.S. and China. (Schneider & Kilian, 2023) They could form alignments or warring factions, leading to unforeseen behaviors. Just as the intelligent behavior of a flock of birds or swarm of bees emerges from the individual behavior of the units, a novel intelligence could emerge from the interaction of scores of individual AIs. If the AIs it emerges from are themselves AGIs, the emergent system could be vastly more complex and intelligent, and potentially more dangerous and deceptive, than the units.

So where does all this leave us? The discussion of the last few sections urged that if the simulation argument is correct, we should take seriously the possibility that we are in an illusory world simulation. For considerations in astrobiology suggest it is plausible that the architect could be a superintelligent AI, and AI safety considerations suggest that advanced AI can be agential and deceptive, raising the likelihood that the simulation would be an illusory one. Given the conclusion of the simulation argument, there is a non-negligible possibility that the simulation is an illusory world scenario, being generated by a deceptive superintelligence. While this likelihood may not be high, for example, it may not even reach ten percent, and nothing I am saying means that the hypothetical example of Pandora is true, in particular, it does mean the likelihood of some sort of illusory world simulation is also not so remote or fictitious as to be dismissed outright. Thus, it is what I've called a "serious skeptical threat".

4 | CONCLUSION

Now let's pull this all together. Chalmers is correct that simulated worlds can be Mreal. I've drawn from his insight, but urged, nevertheless, that external world skepticism is possible. For we could be in an illusory world—a world generated by the likes of Pandora, or some other sort of entity aiming to deceive. Thus, I conclude that illusory world skepticism is correct. Further, illusory world skepticism is correct even if one rejects the simulation argument and regards illusory world scenarios as merely being remote and unlikely possibilities. But in the last few sections I considered whether, in addition, illusory world skepticism is a more serious epistemic challenge, at least for someone who regards the conclusion of the simulation argument as plausible. Drawing from issues in astrobiology and AI, I raised considerations suggesting that, on the assumption that it is likely we are in simulation, it is in fact a serious possibility that we are in an illusory world.

REFERENCES

Bailey, M., & Schneider, S. (2023). AI shouldn't decide what's true: Experts on why trusting artificial intelligence to give us the truth is a foolish bargain. *Nautilus*, *17*.

Bailey, M. (2025). *Autonomous Minds*, Georgetown University Press.

Bennett, J., Shostak, S., & Schneider, N. (2011). *Life in the universe* (3rd ed.). Pearson.

Bostrom, N. (2003). Are we living in a computer simulation? *Philosophical Quarterly*, *53*, 243–255.

Chalmers, D. J. (2003). Matrix as metaphysics. In *Philosophers' Imprint*, *3*(1), 1–35. Available at: http://philosophersimprint.org/003001/

Chalmers, D. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. WW Norton & Co.

Davies, P. (2010). *The Eerie Science: Renewing our search for alien intelligence*. Houghton Mifflin Harcourt.

Descartes, R. (1641). Meditations on first philosophy. In J. Cottingham et al., (Eds.), *The philosophical writings of descartes*, (Vol. II). Cambridge University Press.

Dick, S. (2013). Bringing culture to cosmos: The postbiological universe. In S. J. Dick & M. Lupisella (Eds.), *Cosmos and culture: Cultural evolution in a cosmic context*. NASA, Web. http://history.nasa.gov/SP-4802.pdf

Musser, G. (2015). *Spooky action at a distance: The phenomenon that reimagines space and time—and what it means for black holes, the big bang, and theories of everything*. Scientific American/Farrar, Straus and Giroux.

Musser, G. (2023). *Putting ourselves back in the equation: Why physicists are studying human consciousness and AI to unravel the mysteries of the universe*. Farrar, Straus and Giroux.

Schneider, S. (2015). Alien Minds. In S. Dick (Ed.), *The impact of discovering life beyond earth* (pp. 189–206). Cambridge University Press.

Schneider, S. (2017). The problem of the physical base: How the mathematical nature of physics undermines physicalism. *Journal of Consciousness Studies*, *24*(9-10), 7–39.

Schneider, S. (2018). Superintelligent AI and the postbiological cosmos approach. In A. Lursch (Ed.), *What is life? On earth and beyond* (pp. 287–302). Cambridge University Press.

Schneider, S. (2019). *Artificial you: AI and the Future of the Mind*, Princeton: Princeton University Press.

Schneider, S. & Kilian, K. (2023). Wall Street Journal,"Artificial intelligence needs Guardrails and global cooperation", April 28.

Seiberg, N. (2006). "Emergent Spacetime," Rapporteur talk at the 23rd Solvay Conference in Physics, December.

Swingle, B. (ms.) "Constructing holographic spacetimes using entanglement renormalization." arXiv:1209.3304

How to cite this article: Schneider, S. (2024). Illusory world skepticism. *Philosophy and Phenomenological Research*. 1–8. https://doi.org/10.1111/phpr.13123