sciendo

Kant Meets Cyberpunk

Eric Schwitzgebel
University of California at Riverside

DOI: 10.2478/disp-2019-0006 BIBLID [0873-626X (2019) 55; pp.411–435]

Abstract
I defend a how-possibly argument for Kantian (or Kant*-ian) transcendental idealism, drawing on concepts from David Chalmers, Nick Bostrom, and the cyberpunk subgenre of science fiction. If we are artificial intelligences living in a virtual reality instantiated on a giant computer, then the fundamental structure of reality might be very different than we suppose. Indeed, since computation does not require spatial properties, spatiality might not be a feature of things as they are in themselves but instead only the way that things necessarily appear to us. It might seem unlikely that we are living in a virtual reality instantiated on a non-spatial computer. However, understanding this possibility can help us appreciate the merits of transcendental idealism in general, as well as transcendental idealism's underappreciated skeptical consequences.

Keywords
Materialism, space, structuralism, artificial intelligence, computation.

1 Introduction

Transcendental idealism might be true. Transcendental idealism, as I intend the phrase, consists of two theses.

First, spatial properties depend on our minds. External things as they are in themselves, independently of us, do not have spatial properties. Things appear to us to be laid out in space, but that's only because our perception of external things necessarily construes them spatially, locating them in a spatial array. Differently constructed minds, no less intelligent and perceptive, might not experience or conceptualize reality in terms of spatially located objects.

Second, the fundamental nature of things as they are in themselves, independently of us, is unknowable to us. We cannot achieve positive

knowledge of things as they are in themselves through our empirical science, which is conditioned on our perceptual construal of objects as laid out in space; nor can we achieve positive knowledge of things as they are in themselves through a priori reflection or transcendental argument.

Transcendental idealism, famously associated with Kant, is a historically important alternative to materialism or physicalism. The view is *idealist* because it treats all of the spatial (and maybe also temporal and causal) properties of external objects as dependent on our minds, and because, contrary to *materialism* or *physicalism,* it does not regard the fundamental nature of things to be in principle discoverable by the physical sciences, which are limited by having to start from empirical evidence as it appears to our senses.[1]

I will present a how-possibly argument for transcendental idealism. I will argue not that transcendental idealism is true, or even that it's likely, but only that it might be true. I will do this by exploring an idea that was popularized by the cyberpunk movement in science fiction and philosophically by David Chalmers and Nick Bostrom in their classic papers "*The Matrix* as Metaphysics" (Chalmers 2003/2010[2]) and "Are We Living in a Computer Simulation?" (Bostrom 2003): the idea that we might be living in a virtual reality, or a Matrix, or a simulation. If so, the fundamental nature of reality might be unknowable to us and very different than we normally suppose. It might be, for example, non-spatial. Spatiality might merely be the way that a fundamentally non-spatial reality is experienced by our minds.

I will argue that if we entertain the specific epistemic possibility that we are living in a virtual reality implemented by a non-spatial system, that can help us understand transcendental idealism in

[1] Materialism or physicalism is notoriously difficult to define (Hempel 1980, Montero 1999, Chomsky 2009, Stoljar 2010). My characterization here only requires as a necessary (but not sufficient) condition that the fundamental nature of things be in principle discoverable by the physical sciences. This isn't quite adequate as a necessary condition, if materialism or physicalism is compatible with skepticism about what is in principle discoverable by the physical sciences. My specific characterization of "Angel" in Section 5 hopefully renders this nuance irrelevant.

[2] Chalmers hints at a Kantian interpretation of his work when he says that *The Matrix* "might be seen more fundamentally as an illustration of Kantian humility" (2003/2010: 489, note 2).

general. I will conclude with some thoughts about radical skepticism.

The virtual might not only be real (Chalmers 2017b), it might be the only reality we can know.

2 Kant, Kant*, and transcendental idealism

According to Immanuel Kant, space is nothing but the form of all appearances of outer sense. It does not represent any property of things as they are in themselves, independent of our minds (1781/1787/1965, A26/B42, p. 71). It is "transcendentally ideal" in the sense that it has no existence independently of our possible experience (1781/1787/1965, A28/B44, p. 72). We cannot know whether other thinking beings are bound by the same conditions that govern us; they might have a form of outer sense that does not involve experiencing objects as laid out in space (1781/1787/1965, A27/B43, p. 72).

Notoriously, these claims invite diverse interpretations. I will offer one interpretation, which I hope is broadly within the range of defensible interpretations. If it's not the historical Kant's view, perhaps it is the position of a merely possible philosopher *Kant**.[3]

On this view, things as they exist in their own right, independently of us, lack spatial properties. They do not have spatial positions relative to each other; or spatial dimensions like length, breadth, and depth; and they are not extended across spatial or spatiotemporal regions. Spatiality is something we bring to objects. However, we do bring it: Spatial properties are properties that belong to objects, not merely to our minds. Behind our patterns of spatial experience is a structured reality of some sort, which dependably gives rise to our spatial experiences, and because of this relational or dispositional fact, objects can be said really to have spatial properties. Since our empirical science bottoms out in what we can perceive, we cannot

[3] See Stang 2013 for discussion of the range of recent interpretations of Kant's metaphysics. I am broadly sympathetic with how Lucy Allais (2015) draws on work in philosophy of perception and the secondary-quality analogy to steer a middle course between strong phenomenalist or "two-world" approaches and deflationary epistemic interpretations. The transcendental idealism I present here might, however, be a bit more two-world and a bit more dispositionalist about perception than Allais' Kant.

use it to discover what lies fundamentally behind the empirically perceivable world of spatially given objects.

Kant denies that his view can be illustrated "by examples so altogether insufficient as colours, taste, etc." (1781/1787/1965, A29/B45, p. 73), but I believe it can be so illustrated, as long as we are careful not to draw too much from the illustration.[4] Consider sweetness. On one plausible understanding of sweetness, sweetness or unsweetness is not a feature of things as they are independently of us. Ice cream is sweet, and black coffee is unsweet, and milk tastes sweet to some but not to others; and this is a feature that we bring to those things due to the nature of our perception. An alien species might have no taste experiences or very different taste experiences. If they denied the reality of sweetness or asserted that very different things are sweet than we think are sweet, they would not be wrong or missing anything except insofar as they would be wrong or missing something concerning *us*. I am assuming here that sweetness does not reduce to any mind-independent property like proportion of sugar molecules (with which it correlates only roughly), but rather has something essentially to do with an object's tendency to evoke certain experiences in us.[5]

Not everything outside of us is perceived as sweet or unsweet. "Sweet" or "unsweet" cannot literally be applied to a gravitational field or a photon, since they are not potential objects of taste. This might be one reason Kant finds the illustration insufficient. Spatiality is a feature of our perception of *all* outside things, Kant says. It is the necessary form of outer sense. Also, "sweet" is insufficiently abstract for Kant's purposes ("square" would also be insufficiently abstract). A closer analogy might be having a location somewhere in the manifold of possible tastes (where one possible location might

[4] Kant himself illustrates the view by analogy to secondary qualities in his *Prolegomena* (1783/2004, 4:289, p. 40–1). See also Putnam 1981, Allais 2015, contra Van Cleve 1995.

[5] Although sweetness is a more intuitive case, color is more commonly discussed. I favor a dispositionalist approach similar to Locke 1689/1975, Peacocke 1984/1997, and Levin 2000, though I hope that other not-too-distant views of the nature of tastes and colors could also serve for the present argument. See also Chalmers 2017b: 321–2, especially the simple power view and the role functionalism view.

be sweetness +5, sourness +2, saltiness +3, bitterness 0, umami 0; see Chen, Gabitto, Pen, Ryba, and Zuker 2011). Furthermore, we might be able to explain sweetness in terms of something more scientifically fundamental, such as chemistry and brain structures. But breaking out of the box is not possible in the same way with spatiality, since, according to Kant, empirical science necessarily operates on objects laid out in space.

With those substantial caveats, then, we might bring spatiality to things in something like the way we, on this view, bring sweetness-unsweetness to things. As taste necessarily presents its objects in a taste-manifold that does not exist independently of possible experience, sensory perception in general presents its objects in a spatial manifold that does not exist independently of possible experience.

Cyberpunk can help us get a better grip on what this might amount to.

3 Cyberpunk, virtual reality, and empirical objects

Two classics of cyberpunk science fiction are William Gibson's 1986 book *Neuromancer* and the 1999 movie *The Matrix.* These works popularized the idea of "cyberspace" or "the Matrix"—a kind of virtual reality instantiated by networks of computers. ("Cyberspace" is now often used with a looser meaning, simply to refer to the Internet.) In *Neuromancer,* computer hackers can "jack in", creating a direct brain interface with the Internet. When jacked in, instead of experiencing the ordinary physical world around them, they visually experience computer programs as navigable visual spaces, and they can execute computer instructions by acting in those visual spaces. In *The Matrix,* people's bodies are stored in warehouses, and they are fed sensory input by high-tech computers. People experience that input as perceptions of the world, and when they act, the computers generate matching sensory input as though the actions were happening in the ordinary world. People can virtually chat with each other, go to dance parties, make love, and do (or seem to do) all the normal human things, while their biological bodies remain warehoused and motionless. Most people don't realize that this is their situation.

I will now introduce several concepts.

Following Chalmers (2017b), but adding "spatial" for explicitness,

an *immersive spatial environment* is an environment "that generates perceptual experience of the environment from a perspective within it, giving the user the sense of 'being there'" (2017b: 312); an *interactive* immersive spatial environment is an immersive spatial environment in which the user's actions can have significant effects; and a *virtual reality* is an interactive immersive spatial environment that is computer generated. So, for example, Neo when he is in the Matrix, and the computer hacker Case when he is in cyberspace, are in virtual realities: They are perceptually immersed in computer-generated spatial environments, and their actions affect the objects that they see. The same is true for typical players of current virtual reality games, like those for the Oculus Rift gear. You are also, right now, in an interactive immersive spatial environment, though perhaps not a computer generated one and so not a *virtual* reality. You see, maybe, a piece of paper as being a certain distance from you, a certain size and shape, laid out in space among other spatial things; you feel the chair in which you are sitting; you feel surrounded by a room; and you can interact with all these things, changing them through your actions.

Taking our cue from Kant, let's call the objects laid out around you in your immersive spatial environment *empirical objects*. In *Neuromancer,* the computer programs that the hackers see are the empirical objects. In *The Matrix,* the dance floor that the people experience is an empirical object—and the body-storage warehouse is not an empirical object, assuming that it's not accessible to them in their immersive environment. For you, reader, empirical objects are just the ordinary objects around you: your coffee mug, your desk, your computer. Our bodies as experienced in immersive spatial environments are also empirical objects: They are laid out in space among the other empirical objects. In *The Matrix,* there's a crucial difference between one's empirical body and one's biological body. If you are experiencing yourself as on a dance floor, your empirical body is dancing, while your biological body is resting motionless in the warehouse. Only if you break out of the Matrix will your empirical and biological bodies be doing the same things. Note that *empirical* is a relational concept. What is empirical for you depends on what environment you are spatially immersed in.

We can think of a *spatial manifold* as an immersive spatial

environment in which every part is spatially related to every other part. The dance floor of the ordinary people trapped in the Matrix is not part of the same spatial manifold as the body-storage warehouses. Suppose you are dancing in the Matrix and someone tells you that you have a biological body in a warehouse. You might ask in which direction the warehouse lies—north, south, east, west, up, down? You might point in various possible directions from the dance floor. Your conversation partner ought to deny the presupposition of your question. The warehouse is not in any of those directions relative to the dance floor. You cannot travel toward it or away from it using your empirical body. You cannot shoot an empirical arrow toward it. In vain would you try to find the warehouse with your empirical body and kick down its doors. It's not part of the same spatial manifold.

Let's call a spatial manifold *shared* if more than one person can participate in the same spatial environment, interacting with each other and experiencing themselves as acting upon the empirical objects around them in coherent, coordinated ways. For example, you and I might both be experiencing the same dance floor, from different points of view, as if we are facing each other. I might extend my empirical hand toward you, and you might see my hand coming and grasp it; and all of these experiences and empirical actions might be harmoniously coordinated, adjusting for our different viewpoints.

The *boundaries of a reality* (whether virtual or non-virtual) are the boundaries of that reality's spatial manifold. Importantly, this can include regions and empirical objects that are not currently being experienced by anyone, such as the treasure chest waiting behind the closed door in a virtual reality game. There is an intuitive sense in which that still-unseen treasure chest is part of the reality of the gameworld. If you and I occupy the shared virtual reality of that game, we might argue about what's behind the door. You say it's a dragon. I say it's a treasure chest. There's a sense in which I am right: There really is a treasure chest behind that door. Exactly how to make sense of unperceived empirical objects has troubled metaphysical idealists of all stripes. One approach is to say that they exist because, at least in principle, they would be perceived in the right conditions. The reason it's right to say that there really is a treasure chest behind that door in our shared virtual reality is that, in normal

circumstances, if we were to open that door we would experience that chest.[6]

There needn't be a single underlying computer object that neatly maps onto that unseen treasure chest. The computational structures beneath an experienced virtual reality might divide into ontological kinds very different from what could be discovered by even the most careful empirical exploration within that reality. The underlying structures might be disjunctive, distributed, half in the cloud under distant control, or a matter of just-in-time processes primed to activate only when the door is opened. They might be bizarrely programmed, redundant, kludgy, changeable, patchwork, luxuriously complex, dependent on intervention by outside operators, festooned with curlicues to delight an alien aesthetic—not at all what one would guess.

It is conceivable that intelligent, conscious beings like us could spend most or all of their lives in a shared virtual reality, acting upon empirical objects laid out in an immersive spatial environment, possibly not realizing that they have biological brains that aren't part of the same spatial manifold. One reason to think that this is conceivable is that central works of cyberpunk and related subgenres appear to depend for their narrative success and durable interest on ordinary people's ability to conceive of this possibility.

4 How to be a sim, fundamentality, and the noumenal

Nick Bostrom (2003) has famously argued that we might be *sims*. We might be artificial intelligences living in a shared virtual reality coordinated by a computer or network of computers. The crucial difference between this scenario and the virtual reality scenarios of Section 3 is that if you are a sim you don't have a biological brain. You yourself are instantiated computationally.

Many people think that we might someday create conscious artificial intelligences with robotic bodies and computer "brains"—like the android Data from *Star Trek* or Isaac Asimov's robots. For purposes of this essay, I'll assume that this is possible. So, then, imagine a conscious robot. Now imagine that it "jacks in" to cyberspace

[6] See, for example, Berkeley 1710/1965 and Mill 1867.

—that is, it creates a direct link between its computer brain and a computer-generated virtual reality, which it then empirically acts in. With a computer brain and a computer-generated virtual-reality environment, nothing biological would be required. Both the subject and its empirical reality would be wholly instantiated in computers. This would be one way to become a sim.

Alternatively, consider the popular computer game *The Sims*. In this game, artificial people stroll around and conduct their business in an artificial environment. You can watch and partly control them on your computer screen. The "people" are controlled by simple AI programs. However, we might imagine someday redesigning the game so that those AI programs are instead very sophisticated, with human-like perceptual experiences. These conscious sims would then interact with each other, and they would act on empirical objects in a spatial manifold that is distinct from our own.

Still another possibility is scanning and "uploading" a copy of your memories and cognitive patterns into a virtual reality, as imagined by some science fiction authors and futurists (e.g., Egan 1994, 1997; Kurzweil 2005; Chalmers 2010). In Greg Egan's version, biological humans scan their brains in detail, which destroys those brains, and then they live among many other "citizens" in virtual realities within highly protected supercomputers. Looking at these computers from the outside, a naive observer might see little of interest.

In a simulation, there's a *base level* of reality and a *simulated level* of reality. At the base level is a computer that implements the cognitive processing of the experiencing subjects and all of their transactions with their simulated environments. At the simulated level are the experiencing subjects and their empirical objects. At the base level there might be a gray hunk of computer in a small, dark room. At the simulated level, subjects might experience a huge, colorful world. At the same time, the base level computer might be part of a vast base-level spatial manifold far beyond the ken of the subjects within the simulation: computer plus computer operators plus the storage building, surrounding city, planet, galaxy.

The base level and the simulated level are *asymmetrically dependent*. The simulated level depends on what's going on at the base level but not vice versa. If the base-level computer is destroyed or loses power, the entire simulation will end. However, unless things have been

specially arranged in some way, no empirical activity within the simulation can have a world-destroying effect on base-level reality.

Similarly, the base level is more *fundamental* than the simulated level. Although fundamentality is a difficult concept to specify precisely, it seems clear that there's a reasonable sense of fundamentality on which this is so. Perhaps we can say that events in the computer "ground" events in the simulation, while events in the simulation do not similarly ground events in the computer; or that events in the simulation "reduce to" or are constituted by events in the computer, while events in the computer do not similarly reduce to, and are not constituted by, events in the simulation. Events in the simulation might "supervene" on events in the computer, but not vice versa. Maybe we can say that the treasure chest is "nothing but" computational processes in the base-level computer, while it's not equally accurate to say that the computational processes are nothing but the treasure chest.

Drawing again from Kant, we might distinguish *phenomena* from *noumena*. Phenomena are things considered as empirical objects of the senses. For the sims in our example, phenomena are things laid out in the spatial manifold of the simulation. The sims might understand that behind these phenomena some structure exists that undergirds their perceived reality, a noumenon (in Kant's "negative" sense) which is not for them a possible object of perception. (Contra Kant, it might be only contingently the case that the base-level reality is not a possible object of the sims' perception. Let's bracket that issue for now.)

As a stepping stone to transcendental idealism, we have so far imagined the base-level computer as an empirical object laid out in a spatial manifold (the same manifold as its operators at the base level of reality). Let's leave that stepping stone behind. We must attempt to conceive of this computer not as a spatially located, material object. Otherwise, we're still operating within a materialist picture.

5 Immaterial computation

Standard computational theory goes back to Alan Turing (1936). One of its most famous results is this: Any problem that can be solved purely algorithmically can in principle be solved by a very

simple system. Turing imagined a strip of tape, of unlimited length in at least one direction, with a read-write head that can move back and forth, reading alphanumeric characters written on that tape and then erasing them and writing new characters according to simple if-then rules. In principle, one could construct a computer along these lines—a "Turing machine"—that, given enough time, has the same ability to solve computational problems as the most powerful supercomputer we can imagine.

Hilary Putnam remarked that there is nothing about computation that requires it to be implemented in a material substance (1965: 43–4). We might, in theory, build a computer out of ectoplasm, out of immaterial soul-stuff. For concreteness, let's consider a broadly Cartesian concept of the soul (Descartes 1641/1984, 1647/1985). It is capable of thought and conscious experience. It exists in time, and it has causal powers. However, it does not have spatial properties such as extension or spatial position. To give it full power, let's assume that this soul has perfect memory. This need not be a *human* soul. Let's call it Angel.[7]

Such a soul might be impossible according to the laws of nature—at least the laws of *empirical* nature as we know it—but set that question aside for the moment. Coherent conceivability is sufficient for present purposes. In principle, could a Turing machine, or its computational equivalent, be built from an immaterial Cartesian Angel?

A proper Turing machine requires the following:

- a finite, non-empty set of possible *states* of the machine, including a specified starting state and one or more specified halting states;

[7] I will attribute moods, perceptual experiences, and imaginings to this soul, which Descartes believes arise from the interaction of soul and body. On my understanding of Descartes, these are possible in souls without bodies, but if necessary we could change to more purely intellectual examples, such mathematical thoughts. I am also bracketing Descartes' view that the soul is not a "machine", which appears to depend on commitment to a view of machines as necessarily material entities (1637/1985, part 5). If Angel is free not to implement the computational algorithm, that also introduces complications, if freedom requires the possibility of acting otherwise and if the computational description would have to incorporate that possibility.

- a finite, non-empty set of *symbols*, including a specified blank symbol;

- the capacity to move a *read/write head "right" and "left" along a tape* inscribed with those symbols, reading the symbol inscribed at whatever position the head occupies; and

- a finite *transition function* that specifies, given the machine's current state and the symbol currently beneath its read/write head, a new state to be entered and a replacement symbol to be written in that position, plus an instruction to then move the head either right or left.

A Cartesian soul ought to be capable of having multiple *states*. We might imagine that Angel has moods, such as bliss. Perhaps he can be in any one of several discrete moods along an interval from sad to happy. Angel's initial state might be the most extreme sadness and Angel might halt only at the most extreme happiness.

Although we normally think of an alphabet of *symbols* as an alphabet of written symbols, symbols might also be merely imagined. Angel might imagine a number of discrete pitches from the A three octaves below middle C to the A three octaves above middle C, with middle C as the blank symbol.

Instead of physical *tape*, Angel thinks of integer numbers. Instead of having a *read-write head that moves right and left* in space, Angel thinks of adding or subtracting one from a running total. We populate the "tape" with symbols using Angel's perfect memory: Angel associates 0 with one pitch, +1 with another pitch, +2 with another pitch, and so forth, for a finite number of specified associations. All unspecified associations are assumed to be middle C. Instead of a read-write head starting at a spatial location on a tape, Angel starts by thinking of 0, and recalling the pitch that 0 is associated with. Instead of the read-write head moving right to read the next spatially adjacent symbol on the tape, Angel adds one to his running total and thinks of the pitch that is associated with the updated running total. Instead of moving left, he subtracts one. Thus, Angel's "tape" is a set of memory associations like those in Figure 1, where at some point specific associations run out and middle C (C_4, i.e., C in the 4th

octave) is assumed on to infinity.

integer	associated pitch	
0	E_5	
+1	$D^{\#}_5$	
+2	E_5	
+3	$D^{\#}_5$	
+4	E_5	← current running total
+5	B_4	
+6	D_5	
+7	C_5	
+8	A_4	
…	etc.	

Figure 1: Immaterial Turing tape. An immaterial Angel remembers associations between integers and musical tones and keeps a mental running total representing a notional read-write head's current "position".

The *transition function* can be understood as a set of rules of this form: If Angel is in such and such a state (e.g., 23% happy) and is "reading" such and such a note (e.g., E_5), then Angel should "write" such-and-such a note (e.g, G_4), enter such-and-such a new state (e.g., 52% happy), and either add or subtract one from his running count. We rely on Angel's memory to implement the writing and reading: To "write" G_4 when his running count is +2 is to commit to memory the idea that next time the running count is +2 he will "read"—that is, actively recall—the symbol G_4 (instead of the E_5 he previously associated with +2).

As far as I can tell, Angel is a perfectly fine Turing machine equivalent. If standard computational theory is correct, he could execute any computational task that any ordinary material computer could execute. And he has no properties incompatible with being an immaterial Cartesian soul as such souls are standardly conceived.

I have chosen an immaterial Cartesian soul as my example in this section only because Cartesian souls are the most familiar example of a relatively non-controversially non-material type of conceivable (if not actually existing) entity. If there is something incoherent or otherwise objectionable about Cartesian souls, then imagine,

if possible, any entity or process (1) whose existence is disallowed by materialism and (2) which has sufficient structure to be Turing-machine equivalent. (If you think that no coherently conceivable entity is disallowed by materialism, then either your materialism lacks teeth or you have an unusually high bar for conceivability.)

6 The flexibility of computational implementation

Most of us don't care what our computers are made of, as long as they *work*. A computer can use vacuum tubes, transistors, integrated circuits on silicon chips, magnetic tape, laser discs, or pretty much any other technology that can be harnessed to implement computational tasks. Some technologies are faster or slower for various tasks. Some are more prone to breakdowns of various sorts under various conditions. But in principle all such computers are Turing-machine equivalent in the sense that, if they don't break down, then given enough time and memory space, they could all perform the same computational tasks. In principle, we could implement Neo's Matrix on a vast network of 1940s style ENIAC computers.

Now in theory it could matter, philosophically, whether a simulation is run using transistors and tape, versus integrated circuits and lasers, versus some futuristic technology like interference patterns in reflected light. Also, in theory, it could matter whether at the finest-grained functional level the machine uses binary symbols versus a hundred symbol types, and whether it uses a single read-write head or several that operate in parallel and at intervals integrate their results. Someone might argue that real spatial experience of an empirical manifold could arise in a simulation only if the simulation is built of integrated circuits and lasers rather than transistors and tape, even if the simulations are executing the same computational tasks at a more abstract level. Or someone might argue that real conscious experience requires parallel processing that is subsequently integrated rather than equivalently fast serial processing. Or someone might argue that speed is intrinsically important so that a slow enough computer simply couldn't host consciousness.

These are all coherent views—but they are more common among AI skeptics and simulation skeptics than among those who grant the possibility of AI consciousness and consciousness within simulated

realities. More orthodox, among AI and sim enthusiasts, is the view that the computational substrate doesn't matter. An AI or a simulation could be run on any substrate as long as it's functionally capable of executing the relevant computational tasks.[8]

For concreteness, let's imagine two genuinely conscious artificially intelligent subjects living in a shared virtual reality: Kate and Peer from Egan's *Permutation City*. Kate and Peer are lying on the soft dry grass of a meadow, in mild sunshine, gazing up at passing clouds. If we are flexible about implementation, then beneath this phenomenal reality might be a very fast 22nd-century computer operating by principles unfamiliar to us, an ordinary early 21st-century computer, or a 1940s ENIAC-style computer. Kate and Peer experience their languorous cloud-watching as lasting about ten minutes. On the 22nd-century computer it might take a split second for those events to transpire. On an early 21st-century computer maybe it takes a few hours or months (depending on how much computational power is required to instantiate human-grade consciousness in a virtual environment). On ENIAC it would take vastly longer and a perfect maintenance crew operating over many generations.

In principle, the whole thing could be instantiated on Turing tape. Beneath all of Kate's and Peer's rich phenomena, there might be only a read-write head following simple rules for erasing and writing 1s and 0s on a very long strip of paper. Viewed from outside—that is, from within the spatial manifold containing the strip of paper—one might find it hard to believe that two conscious minds and a shared phenomenal reality could arise from something so simple. But this is where a commitment to flexibility about implementation seems to lead us. The bizarreness of this idea is one reason to have some qualms about the assumptions and argumentative moves that brought us to it; but as I have argued elsewhere, *all* general theories of the conscious mind have some highly bizarre consequences, so bizarreness is not necessarily a defeater (Schwitzgebel 2014).

You know where this is headed. It is conceivable that our immaterial computer Angel, or some other entity disallowed by materialism, is the system implementing Kate's and Peer's phenomenal

[8] See for example Putnam 1965 on functionalism and probabilistic automata and Chalmers 1996 on the principle of organizational invariance.

reality. If Kate and Peer are conceivable, it is also conceivable that the computer implementing them is non-material.

7 From Kate and Peer to transcendental idealism

According to transcendental idealism as I have characterized it, space is not a feature of things as they are in themselves, although it is the necessary form of our perception of things. Beneath the phenomena of empirical objects that we experience is something more fundamental, something nonspatial and non-material—something beyond empirical inquiry. Part of the challenge in recognizing transcendental idealism as a viable competitor to materialism, I believe, is that it sounds so vague and mystical that it is difficult to conceive what it might amount to or how it could even possibly be true.

Here's what it might amount to: Beneath our perceptual experiences there might be an immaterial Cartesian soul implementing a virtual reality program in which we are embedded; this entity's fundamental structure might be unknowable to us, either by means of the empirical tools of physical science or by any other means such as a priori or transcendental reflection; and spatiality might be best understood as the way that our minds are constituted to track and manage interactions among ourselves and with other parts of that soul, somewhat analogously to the way that (on the view described in Section 2) our taste experiences help us navigate the edible world. (To be clear, I am suggesting that this last clause is a possible way of developing the Angel scenario; I have not argued that it is the only way.) If the world is like that, then transcendental idealism is correct and materialism is false.

Here's how it might possibly be true: We have no decisive evidence to rule this possibility out, unless we result to the Moorean move of just taking its falsity as a starting premise.[9] We might have excellent *empirical* evidence that everything that exists is material. We might even have excellent empirical evidence that consciousness can only occur in entities with brains composed of biological material. But all such evidence is consistent with things being very different at a more fundamental level. Artificial intelligences in a virtual

[9] Moore 1925; similarly Wittgenstein 1951/1969; Lycan 2001.

reality might have very similar empirical evidence.

I doubt that the most likely form of transcendental idealism is one in which we live within an Angel sadly imagining musical notes while keeping a running total of integers. But my hope is that once we vividly enough imagine this possibility, we begin to see how *in general* transcendental idealism might be true. If transcendental idealism is true, there's no good reason to suppose that things as they are in themselves are as easy to imagine as sadness and music.

I have articulated a possible transcendental idealism about space. But Kant himself was more radical. He argued that *time* also is transcendentally ideal, not a feature of things as they are in themselves independently of us. The nature of my example forces me to retain the transcendental reality of time: Computation appears to involve state transitions, which seem to require change over time.[10] I have also relied on our conception of an immaterial mind in a way that Kant would probably reject. Nonetheless, the possible transcendental ideality of space, coupled with the possible nonmateriality and undiscoverability of the fundamental features of things as they are in themselves, is already enough to constitute a transcendental idealist alternative to materialism. A more radical and more thoroughly Kantian transcendental idealism might dispose of all empirically-based concepts, including computation and time. Perhaps we can only negatively and abstractly imagine this possibility.

Is there any reason to regard transcendental idealism as anything other than an extremely remote possibility? I see four reasons.

First, materialism has problems as a philosophical position, including in the difficulty of articulating what it is to be "material" or "physical", the widespread opinion that it could never adequately explain phenomenal consciousness, and the fact that all well-worked out materialist approaches appear to commit to highly bizarre consequences of one sort or another.[11] Pressure against materialism is pressure in favor of an alternative position, and transcendental idealism is a historically important alternative position.

[10] Some have argued that computation can occur purely abstractly (Steinhart 2014, Tegmark 2014), but I see no need to employ that idea here.

[11] On the first point, see note 1 above; on the second see, e.g., Chalmers 1996; on the third, see Schwitzgebel 2014.

Second, if Bostrom (2003) is right, the possibility that we are living in a computer simulation deserves a non-trivial portion of our credence. If we grant that, I see no particular reason to assume that the base-level of reality is material, or spatially organized, or discoverable by inquirers living within the simulation.

Third, it is reasonable for us, I believe, to have considerable *cosmological skepticism*. Although the best current scientific cosmology is a Big Bang cosmology, cosmological theory has proven unstable over the decades, offers no consensus explanation of the cause (if any) of our universe, and is not even uniformly materialist. We have seen, perhaps, only a minuscule portion of the cosmos. We might be like fleas on the back of a dog, watching a hair grow and saying, "Ah, so *that's* how the universe works!"[12]

Fourth, whatever we know about external things, apart from what is knowable a priori or transcendentally, appears to depend on how those external things affect our senses. But things with very different underlying properties, call them A-type properties versus B-type properties, could conceivably affect our senses in identical ways. If so, we might have no good reason to suppose that they do have A-type properties rather than B-type properties.[13]

8 Structuralism about space

Here's a possible objection. We encounter a large, stable empirical world of objects laid out in space. We successfully navigate among those objects, and there are many empirical facts that correlate with their perceived spatial structure—how long it takes to walk somewhere, how much the magnet pulls one nail rather than another, where the stone we throw will be perceived to land. Whatever struc-

[12] On cosmological skepticism, see Schwitzgebel 2014, 2017a. Quote inspired by Hume 1779/1947.

[13] For versions of this fourth consideration in support of transcendental idealism, see Putnam 1981, Langton 1998. Maybe if A-type properties are much simpler than B-type properties, and if we have reason to suppose that the underlying reality is relatively simple, then we can infer A-type rather than B-type properties. Or maybe A-type properties are closer to common sense and we ought to stick with common sense unless there is compelling reason to reject it. Or....

ture undergirds all of these facts must have properties that explain those correlations. If it's a computer program, it must somehow be modeling spatial relationships like "next to", "within the gravity well of", and "across the river from". Maybe that's sufficient to justify saying that the computer has spatial properties, which our experiences of space are tracking. Spatiality might, in this sense, be a structural or functional or relational concept: All there is to spatiality is such stable patterns of relationships or the underlying structure (of whatever sort) that explains such patterns of relationships.[14] Angel might then be a spatial entity after all, and our experiences of space might be experiences of the real structural properties of Angel as he is in himself, independently of us, contra transcendental idealism.

I see at least three available replies for the transcendental idealist.

Reply 1. One possible structuralist view of space is straightforwardly friendly to transcendental idealism. On this view, among the essential functional, causal, or structural relata are *our experiences of space*. Spatiality is just whatever feature of the world is responsible for our patterns of spatial experience. To get the metaphysical oomph needed for this to constitute a form of idealism, the dependency must be modally strict. It can't be that we merely use our experiences of space to pick out whatever structures are actually responsible for those experiences, identify those structures as the spatial ones, and treat those structures as spatial in their own right, in a way that could hypothetically come entirely apart from our experiences of

[14] Leibniz's correspondence with Clarke is the locus classicus for relational views of space or spacetime (Alexander, ed., 1956). On structuralism about space or spacetime, see Chalmers 2010: ch. 7.5, 2017: 323–4; Greaves 2011; Lam 2017. Chalmers endorses a version of spatial structuralism or functionalism which understands space in terms of its causal role, including objects' "effects on each other and on our experiences" (2017b: 324; cf. 2010: 334–5). If the "on our experiences" aspect is ineliminable, then transcendental idealism as I have characterized it remains unthreatened. However, there are two distinct ways in which our experiences might be eliminable: First, the relevant form of structuralism might be "realizer functionalism" in the sense of Chalmers 2017a, on which spatial properties in fact play the causal role characteristic of space (including causing our spatial experiences) but might not have played that role. Second, even granting a modally stricter relationship between spatial properties and patterns of causal interaction, non-experiential object-object relationships might be sufficient for spatiality.

spatiality. On a transcendental idealist version of spatial structuralism, our experiences of space must be among the ineliminable relata constitutive of spatiality. Spatial properties must be spatial at least partly in virtue of their tendency to produce spatial experiences in us. If so, then things as they are in themselves are not spatial independently of our minds.

Consider sweetness again. If something is sweet just in virtue of its tendency to produce sweetness experiences in us, then sweetness essentially involves us. Spatiality might be conceptualized similarly. We might define spatial structures (nonrigidly) as whatever structures tend to produce these patterns of experience in us, while we retain Kantian agnosticism about the noumena beneath, which might be disjunctive, dependent on angelic minds, inconceivably weird, or a complicated network that maps poorly onto empirically discoverable ontological categories.

Reply 2. More threatening to transcendental idealism is a relationalism or structuralism about space that does not treat our experiences of space as ineliminably among the relata. In this case, to save the transcendental idealist possibility we might need to consider a smaller version of Angel without all of the structurally necessary object-to-object relationships fully realized independently of our experience, thereby defeating some of the presuppositions behind the structuralist objection. Given a sufficiently small Angel, for example, it might not be true that we are in fact embedded in a large, stable world of empirical objects that consistently enter the structurally required relationships with each other independently of us. This presupposition verges on being a skeptical possibility, depending on how much of Angel's structure is absent.

Reply 3. Deny the truth of structuralism about space. Structuralism is hardly the consensus opinion. Spatiality might be a property that Angel fundamentally lacks by virtue of being a Cartesian soul, regardless of structural relationships among his moods and thoughts.

A related concern puns on the mathematical concept of "space"; for of course Angel could be modeled geometrically in terms of a multi-dimensional state space. But this is not spatiality in the sense relevant to the truth of transcendental idealism. To see this, consider temporality for comparison. The fact that some set of relationships—for example, the heights of several people in a room—can

be arranged in a scalar "sequence" does not make that set of relationships temporal, despite being modeled by a mathematical structure similar to that by which we model temporal relationships.

9 Skepticism

Defenders of the possibility that we live in a simulated virtual reality, including Bostrom (2011), Chalmers (2012, 2017a), and Steinhart (2014), tend to emphasize that this needn't be construed as a skeptical possibility. Even if we are trapped in the Matrix by evil supercomputers, ordinary things like cups, desks, brains, and dance parties still exist. They are just metaphysically constituted differently than one might have supposed. Indeed, Kant and Chalmers both use arguments in this vicinity for *anti*-skeptical purposes. Roughly, the idea is that it doesn't greatly matter what specifically lies beneath the phenomenal world of appearances. Beneath it all, there might be a "deceiving" demon, or a network of supercomputers, or something else entirely incomprehensible to us. As long as phenomena are stable and durable, regular and predictable, then we know the ordinary things that we take ourselves to know: that the punch is sweet, that dawn will arrive soon, that the bass line is shaking the floor.

I am sympathetic with this move, but intended as a *general* rebuttal of radically skeptical scenarios, it is too optimistic (and Chalmers 2003/2010 and 2017a does explicitly limit the anti-skeptical force of his arguments to a very limited range of scenarios). For example, if we are living in a simulation, I see no compelling reason to believe that it must be a large, stable simulation. It might be a simulation set to run for only two subjective hours before shutting down. It might be a simulation that includes only you in your room, reading this essay. If *that's* what's going on beneath appearances, then much of what you probably ordinarily think you know—that the Sun will rise tomorrow, that Luxembourg exists—is false. If the fundamental nature of things might be radically different from the world as it appears to us, it might be radically different in a way that negates huge swaths of our supposed knowledge.

Consider these two simulation scenarios, both designed to push back against the durable stability assumption beneath the structuralist challenge to skepticism and to the transcendental ideality of space:

Toy simulation. Our simulated world might be purposely designed by creators. But our creators' purposes might not be grand ones that require us to live very long or in a large environment. Our creators might, like us, be limited beings with small purposes: scientific inquiry, mate attraction, entertainment. Huge and enduring simulations might be too expensive to construct. Most simulations might be small or short—only large or long enough to address their research questions, or to awe potential mates, or to enjoy as a fine little toy. If so, then we might be radically mistaken in our ordinary assumptions about the past, the future, or distant things.

Random simulation. The base level of reality might consist of an infinite number of randomly constituted computational systems, executing every possible program infinitely often. Only a tiny proportion of these computational systems might execute programs sophisticated enough to give rise to conscious subjects capable of considering questions about the fundamental nature of reality; but of course any subject who is considering questions about the fundamental nature of reality must be instantiated in one of those rare machines. If these rare machines are randomly constituted rather than designed for stability, it's possible that the overwhelming proportion of them host conscious subjects only briefly, soon lapsing into disorganization.[15]

If we are in a simulation, we might be in a Toy Simulation or a Random Simulation or some other epistemically unfortunate situation. It seems unwarranted to have a very high level of confidence that if we are in a simulation it is both large and stable. What could justify such confidence, except Moorean stipulation? If empirical facts about simulations are any guide, most simulations are small scale. If they are no guide, then we ought to feel even more at sea.[16]

Similar doubts ought to trouble transcendental idealists in general. If material stuff is fundamental, then the odds appear good that we are living in a large, stable, enduring universe with fixed laws that we can rely on. If, instead, fundamental reality is radically

[15] The cosmological literatures on Boltzmann brains and the anthropic principle are relevant here, e.g., Barrow and Tipler 1986, Bostrom 2002, Carroll 2010.

[16] I develop a related argument in Schwitzgebel 2017a. I explore a version of the Random scenario in Schwitzgebel 2017b.

incomprehensible to us, then empirical reality might be subject to whims and chances far beyond our ken. The Divine might stumble over the power cord at any moment, ending us all. The transcendental idealist ought to have some non-trivial doubts about our stability and future.[17]

Eric Schwitzgebel
Department of Philosophy
University of California at Riverside
Riverside, CA 92521-0201
USA
eschwitz at domain: ucr.edu

References

Alexander, H.G., ed. 1956. *The Leibniz-Clarke Correspondence.* Manchester, UK: Manchester University Press.

Allais, Lucy. 2015. *Manifest Reality.* Oxford: Oxford University Press.

Asimov, Isaac. 1982. *The Complete Robot.* Garden City, NY: Doubleday.

Barrow, John D.; and Frank J. Tipler. 1986. *The Anthropic Cosmological Principle.* Oxford: Oxford University Press.

Berkeley, George. 1710/1965. *A Treatise Concerning the Principles of Human Knowledge.* In *Principles, dialogues, and philosophical correspondence*, ed. C. M. Turbayne. New York: Macmillan.

Bostrom, Nick. 2002. *Anthropic Bias.* New York: Routledge.

Bostrom, Nick. 2003. Are we living in a computer simulation? *Philosophical Quarterly* 53: 243–55.

Bostrom, Nick. 2011. Personal communication publicly shared with permission as "Bostrom's response to my discussion of the simulation argument" at *The Splintered Mind* blog, Sep. 2, 2011. http://schwitzsplinters.blogspot.com/2011/09/bostroms-response-to-my-discussion-of.html.

Carroll, Sean. 2010. *From Eternity to Here.* New York: Penguin.

Chalmers, David J. 1996. *The Conscious Mind.* Oxford: Oxford University Press.

Chalmers, David J. 2003/2010. *The Matrix* as metaphysics. In *The Character of Consciousness.* Oxford: Oxford University Press.

Chalmers, David J. 2010. The singularity: a philosophical analysis. *Journal of*

[17] For helpful comments and discussion, thanks to Richard Brown, David Chalmers, Daniel Estrada, François Kammerer, Pierre Keller, Kris Rhodes, Jeremy Pober, Enrico Terrone, and Allen Wood; the audience at Institut Jean Nicod; and commenters on relevant posts on my blog and public Facebook page. (I should probably note that both Keller and Wood strenuously object to this approach to Kant.)

Consciousness Studies 17 (9–10): 7–65.
Chalmers, David J. 2012. *Constructing the World.* Oxford: Oxford University Press.
Chalmers, David J. 2017a. The structuralist response to skepticism. Unpublished manuscript. Downloaded from PhilPapers August 17, 2017.
Chalmers, David J. 2017b. The virtual and the real. *Disputatio* 9: 309–52.
Chen, Xiaoke; Gabitto, Mariano; Peng, Yueqing; Ryba, Nicholas J. P.; and Zuker, Charles S. 2011. A gustotopic map of taste qualities in the mammalian brain. *Science* 333: 1262–6.
Chomsky, Noam. 2009. The mysteries of nature: how deeply hidden? *Journal of Philosophy* 106: 167–200.
Descartes, René. 1637/1985. *Discourse on the Method.* In *The Philosophical Writings of Descartes, vol. I*, trans. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press.
Descartes, René. 1641/1984. *Meditations on First Philosophy.* In *The Philosophical Writings of Descartes, vol. II*, trans. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press.
Descartes, René. 1647/1985. *Principles of Philosophy.* In *The Philosophical Writings of Descartes, vol. I*, trans. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press.
Egan, Greg. 1994. *Permutation City.* London: Millennium.
Egan, Greg. 1997. *Diaspora*. London: Millennium.
Greaves, Hilary. 2011. In search of (spacetime) structuralism. *Philosophical Perspectives* 25: 189–204.
Hempel, Carl G. 1980. Comments on Goodman's *Ways of Worldmaking*. *Synthese* 45: 193–9.
Hume, David. 1779/1947. *Dialogues Concerning Natural Religion*, ed. N.K. Smith. Indianapolis: Bobbs-Merrill.
Kurzweil, Ray. 2005. *The Singularity is Near.* New York: Penguin.
Lam, Vincent. 2017. Structuralism in the philosophy of physics. *Philosophy Compass* 12: **e12421**.
Langton, Rae. 1998. *Kantian Humility.* Oxford: Oxford University Press.
Levin, J. 2000. Dispositional theories of color and the claims of common sense. *Philosophical Studies* 100: 151–74.
Locke, John. 1689/1975. *Essay Concerning Human Understanding*, ed. P.H. Nidditch. Oxford: Oxford University Press.
Lycan, William G. 2001. Moore against the new skeptics. *Philosophical Studies* 103: 35–53.
Mill, John Stuart. 1867. *An Examination of Sir William Hamilton's Philosophy.* London: Longmans, Green, Reader, and Dyer.
Montero, Barbara. 1999. The body problem. *Noûs* 33: 183–200.
Moore, G. E. 1925. A defence of common sense. In *Contemporary British philosophy,* ed. J.H Muirhead. London: George Allen and Unwin.
Peacocke, Christopher. 1984. Colour concepts and colour experience. *Synthese* 58: 365–81.
Putnam, Hilary. 1965. Psychological predicates. In *Art, Mind, and Religion*, ed.

W. H. Capitan and D. D. Merrill. Liverpool: University of Pittsburgh Press / C. Tinling.
Putnam, Hilary. 1981. *Reason, Truth, and History.* Cambridge: Cambridge University Press.
Schwitzgebel, Eric. 2014. The crazyist metaphysics of mind. *Australasian Journal of Philosophy* 92: 665–82.
Schwitzgebel, Eric. 2017a. 1% skepticism. *Noûs* 51: 271–90.
Schwitzgebel, Eric. 2017b. THE TURING MACHINES OF BABEL. *Apex* 98. URL: https://www.apex-magazine.com/the-turing-machines-of-babel/
Snodgrass, Melinda M.; and Scheerer, Robert. 1989. The measure of a man. *Star Trek: The Next Generation,* season 2, episode 9.
Stang, Nicholas F. 2016. Kant's transcendental idealism. In *Stanford Encyclopedia of Philosophy* (Spring 2016 ed.) https://plato.stanford.edu/archives/spr2016/entries/kant-transcendental-idealism.
Steinhart, Eric. 2014. *Your Digital Afterlives.* New York: Palgrave.
Stoljar, Daniel. 2010. *Physicalism.* Oxford: Routledge.
Strawson, Peter F. 1959. *Individuals.* London: Methuen.
Tegmark, Max. 2014. *Our Mathematical Universe.* New York: Random House.
Turing, A. M. 1936. On computable numbers, with an application to the *Entscheidungsproblem*. *Proceedings of the London Mathematical Society, Series 2* 42: 230–65.
Van Cleve, James. 1995. Putnam, Kant, and secondary qualities. *Philosophical Papers* 24: 83–109.
Wittgenstein, Ludwig. 1951/1969. *On Certainty*, ed. G.E.M. Anscombe and G.H. von Wright. New York: Harper.