

The Weirdness of the World

Eric Schwitzgebel

Contents

1. In Praise of Weirdness	3
Part One: Bizarreness and Dubiety	
2. If Materialism Is True, the United States Is Probably Conscious.....	11
• Chapter Two Appendix: Six Objections.....	41
3. Universal Bizarreness and Universal Dubiety	58
4. 1% Skepticism	99
5. Kant Meets Cyberpunk	135
Part Two: The Size of the Cosmos	
6. Experimental Evidence for the Existence of an External World	163
7. Almost Everything You Do Causes Almost Everything (Under Certain Not Wholly Implausible Assumptions); or Infinite Puppetry.....	192
Part Three: More Perplexities of Consciousness	
8. An Innocent and Wonderful Definition of Consciousness	217
9. The Loose Friendship of Visual Experience and Reality	235
10. Is There Something It's Like to Be a Garden Snail? Or: How Sparse or Abundant Is Consciousness in the Universe?.....	250
11. The Moral Status of Future Artificial Intelligence: Doubts and a Dilemma	283
12. Weirdness and Wonder	313
Acknowledgements.....	320
References.....	321

The Weirdness of the World

Chapter One In Praise of Weirdness

The weird sisters, hand in hand,
Posters of the sea and land,
Thus do go about, about:
Thrice to thine and thrice to mine
And thrice again, to make up nine.
Peace! the charm's wound up
(Macbeth, Act I, scene iii)

Weird often saveth
The undoomed hero if doughty his valor!
(Beowulf, X.14-15, trans. L. Hall)

The word “weird” reaches deep back into old English, originally as a noun for fate or magic, later as an adjective for the uncanny or peculiar. By the 1980s, it had fruited as the choicest middle-school insult against unstylish kids like me who spent their free time playing with figurines of wizards and listening to obscure science fiction radio shows. If the “normal” is the conventional, ordinary, predictable, and readily understood, the weird is what defies that.

The world is weird. It wears mismatched thrift-shop clothes, births wizards and monsters, and all of the old science fiction radio shows are true. Our changeable, culturally specific sense of normality is no rigorous index of reality.

One of the weirdest things about Earth is that certain complex bags of mostly water can pause to reflect on the most fundamental questions there are. We can philosophize to the limits of our comprehension and peer into the fog beyond those limits. We can think about the foundations of the foundations of the foundations, even with no clear method and no great hope of an answer. In this respect, we vastly out-geek bluebirds and kangaroos.

1. What I Will Argue in This Book.

Consider three huge questions: What is the fundamental structure of the cosmos? How does human consciousness fit into it? What should we value? What I will argue in this book – with emphasis on the first two questions, but also sometimes drawing implications for the third – is (1.) the answers are currently beyond our capacity to know, and (2.) we do nonetheless know at least this: Whatever the truth is, it’s weird. Careful reflection will reveal all of the viable theories on these grand topics to be both bizarre and dubious. In Chapter 3 (“Universal Bizarreness and Universal Dubiety”), I will call this the Universal Bizarreness thesis and the Universal Dubiety thesis. Something that seems almost too crazy to believe must be true, but we can’t resolve which of the various crazy-seeming options is ultimately correct. If you’ve ever wondered why every wide-ranging, foundations-minded philosopher in the history of Earth has held bizarre metaphysical or cosmological views (each philosopher holding, seemingly, a different set of bizarre views), Chapter 3 offers an explanation.

I will argue that given our weak epistemic position, our best big-picture cosmology and our best theories of consciousness are strange, tentative, and modish.

Strange: As I will argue, every approach to cosmology and consciousness has bizarre implications that run strikingly contrary to mainstream “common sense”.

Tentative: As I will also argue, epistemic caution is warranted, partly *because* theories on these topics run so strikingly contrary to common sense and also partly because they test the limits of scientific inquiry. Indeed, the nature and value of scientific inquiry itself rests upon dubious assumptions about the fundamental structure of mind and world, as I discuss in Chapters 4 (“1% Skepticism”), 5 (“Kant Meets Cyberpunk”), and 6 (“Experimental Evidence for the Existence of an External World”).

Modish: On a philosopher's time scale – where a few decades ago is “recent” and a few decades hence is “soon” – we live in a time of change, with cosmological theories and theories of consciousness rising and receding based mainly on broad promise and what captures researchers' imaginations. We ought not trust that the current range of mainstream academic theories will closely resemble the range in a hundred years, much less the actual truth.

2. Varieties of Cosmological Weirdness.

To establish that the world is cosmologically bizarre, maybe all that is needed is relativity theory and quantum mechanics.

According to relativity theory, if your twin accelerates away from you at nearly light speed then returns, much less time will have passed for the traveler than for you who stayed here on Earth – the so-called Twin Paradox. According to quantum mechanics, if you observe the decay of a uranium atom, there's also an equally real, equally existing version of you in another “world” who shares your past but who observed the atom not to have decayed. Or maybe your act of observation caused the decay, or maybe some other strange thing is true, depending on your favored interpretation of quantum mechanics. Oddly enough, the many-worlds hypothesis appears to be the most straightforward interpretation of quantum mechanics.¹ If we accept that view, then the cosmos contains a myriad of slightly different, equally real worlds each containing different versions of you and your friends and everything you know, each splitting off from a common history.

¹ Greene 2011; Wallace 2012; Carroll 2019. For a review of the leading interpretations, see Maudlin 2019.

I won't dwell on those particular cosmological weirdnesses, since they are familiar to academic readers and well-handled elsewhere (for example, in recent books by Sean Carroll and Brian Greene).² However, some equally fundamental cosmological issues are typically addressed by philosophers rather than scientific cosmologists.

One is the possibility that the cosmos is nowhere near as large as we ordinarily assume – perhaps just you and your immediate environment (Chapter 4) or perhaps even just your own mind and nothing else (Chapter 6). Although these possibilities might not be likely, they are worth considering seriously, to assess how confident we ought to be in their falsity and on what grounds. I will argue that it's reasonable not to *entirely* dismiss such skeptical possibilities. Alternatively, and more in line with mainstream physical theory, the cosmos might be infinite, which brings its own train of bizarre consequences (Chapter 7, “Almost Everything You Do Causes Almost Everything (Under Certain Not Wholly Implausible Assumptions); or Infinite Puppetry”).

Another possibility is that we live inside a simulated reality or a pocket universe, embedded in a much larger structure about which we know virtually nothing (Chapters 4 and 5). Still another is that our experience of three-dimensional spatiality is a product of our own minds that doesn't reflect the underlying structure of reality (Chapter 5) or maps only loosely onto it (Chapter 9, “The Loose Friendship of Visual Experience and Reality”).

Still another set of questions concerns the relationship of mind to cosmos. Is conscious experience abundant in the universe, or does it require the delicate coordination of rare events (Chapter 10, “Is There Something It's Like to Be a Garden Snail? Or: How Sparse or Abundant Is Consciousness in the Universe?”)? Is consciousness purely a matter of having the right

² Carroll 2010, 2019; Greene 2011, 2020.

physical structure, or might it require something nonphysical (Chapter 3)? Under what conditions might a group of organisms give rise to group-level consciousness (Chapter 2, “If Materialism Is True, the United States Is Probably Conscious”)? What would it take to build a conscious machine, if that is possible at all – and what ought we to do if we don’t know whether we have succeeded (Chapter 11, “The Moral Status of Future Artificial Intelligence: Doubts and a Dilemma”)?

In each of our heads are about as many neurons as stars in the galaxy, and each neuron is arguably more structurally complex than any star system that does not contain life. There is as much complexity and mystery inside as out.

I will argue that in the most fundamental matters of consciousness and cosmology, neither common sense, nor early 21st-century empirical science, nor armchair philosophical theorizing is entirely trustworthy. The rational response is to distribute your credence across a wide range of bizarre options.

3. Philosophy That Closes Versus Philosophy That Opens.

You are reading a philosophy book – voluntarily, let’s suppose. Why? What do you like about philosophy? Some people like philosophy because they believe it reveals profound, fundamental truths about the one way the world is and the one right manner to live. Others like the beauty of grand philosophical systems. Still others like the clever back-and-forth of philosophical combat. What I like most is none of these. I love philosophy best when it opens my mind – when it reveals ways the world could be, possible approaches to life, lenses through which I might see and value things around me, which I might not otherwise have considered.

Philosophy can aim to open or to close. Suppose you enter Philosophical Topic X imagining three viable possibilities, A, B, and C. The philosophy of closing aims to reduce the three to one. It aims to convince you that possibility A is correct and the others wrong. If it succeeds, you know the truth about Topic X: A is the answer! In contrast, the philosophy of opening aims to add new possibilities to the mix – possibilities that you hadn't considered before or had considered but too quickly dismissed. Instead of reducing three to one, three grows to maybe five, with new possibilities D and E. We can learn by addition as well as subtraction. We can learn that the range of viable possibilities is broader than we had assumed.

For me, the greatest philosophical thrill is realizing that something I'd long taken for granted might not be true, that some "obvious" apparent truth is in fact doubtable – not just abstractly and hypothetically doubtable, but really, seriously, in-my-gut doubtable. The ground shifts beneath me. Where I'd thought there would be floor, there is instead open space I hadn't previously seen. My mind spins in new, unfamiliar directions. I wonder, and wondrousness seems to coat the world itself. The world expands, bigger with possibility, more complex, more unfathomable. I feel small and confused, but in a good way.

Let's test the boundaries of the best current work in science and philosophy. Let's launch ourselves at questions monstrously large and formidable. Let's contemplate these questions carefully, with serious scholarly rigor, pushing against the edge of human knowledge. That is an intrinsically worthwhile activity, worth some of our time in a society generous enough to permit us such time, even if the answers elude us.

4. To Non-Specialists: An Invitation and Apology.

I will try to write plainly and accessibly enough that most readers who have come this far can follow me. I think it is both possible and important for academic philosophy to be comprehensible to non-specialists. But you should know also that I am writing primarily for my peers – fellow experts in epistemology, philosophy of mind, and philosophy of cosmology. There will be slow and difficult patches, where the details matter. Most of the chapters are based on articles published in technical philosophy journals – articles revised, updated, and integrated into what I hope is an intriguing overall vision. These articles have been lengthened and deepened, not shortened and simplified. The chapters are designed mostly to stand on their own, with cross-references to each other. If you find yourself slogging, please feel free to skip ahead. Trust your sense of what’s interesting to you.

My middle-school self who used dice and thrift-shop costumes to imagine astronauts and wizards is now a fifty-four-year old who uses 21st century science and philosophy to imagine the shape of the cosmos and the magic of consciousness. Join me! If doughty our valor, the weird may saveth us.

[Illustration 1 (no caption): a nerdy middle-schooler in a loose-fitting wizard’s robe, floating on a magic carpet, against a background of stars and nebulas]

The Weirdness of the World

Part One: Bizarreness and Dubiety

The Weirdness of the World

Chapter Two If Materialism Is True, the United States Is Probably Conscious

I begin with the question of group consciousness. I start with this issue because, if I'm right, it's a quicksand of weirdness. Every angle you pursue, whether pro or con, has strange implications. The more you wriggle and thrash, the deeper you sink, with no firm, unweird bottom. In Chapter 3, with this example in mind, I'll lay out the broader framework.

For simplicity, I will assume that you favor *materialism* as a cosmological position. According to materialism, everything in the universe¹ is composed of, or reducible to, or most fundamentally, material stuff, where “material stuff” means things like elements of the periodic table and the various particles or waves or fields that interact with or combine to form such elements, whatever those particles, waves, or fields might be, as long as they are not intrinsically mental.² Later in the book, I'll discuss alternatives to materialism.

If materialism is true, the reason you have a stream of conscious experience – the reason there's something it's like to be you while there's nothing it's like (presumably) to be a bowl of chicken soup, the reason you possess what Anglophone philosophers call *consciousness* or

¹ For present purposes, let's consider our “universe” to include all and only what is spatiotemporally related to us. If there are other universes (Carroll 2010; Green 2011; Tegmark 2014) or “possible worlds” (Lewis 1986) that are not spatiotemporally related to us, materialism need make no commitment about what is going on there.

² The materialist position is difficult to characterize precisely: Hempel 1980; Crane and Mellor 1990; Montero 1999; Chomsky 2009; Stoljar 2010. I hope the current characterization will suffice. See Chapters 3 and 5 for further discussion. For etymological reasons, I prefer “material” (which suggests the intrinsic mindlessness of the fundamental stuff of which we are composed) over “physical” (which suggests deference to physical science). Arguably, the distinction matters if physicists decide that mentality is among the fundamental structures of the world.

phenomenology or *phenomenal consciousness* (I use the terms equivalently³) – is that your basic constituent elements are organized the right way. Conscious experience arises, somehow, from the interplay of tiny, mindless bits of matter. Most early 21st century Anglophone philosophers are materialists in this sense.⁴ You might find materialism attractive if you reject the thought that people are animated by immaterial spirits or possess immaterial properties.

Here's another thought that you will probably reject: The United States is literally, like you, phenomenally conscious. That is, the United States literally possesses a stream of experiences over and above the experiences of its members considered individually. This view stands sharply in tension both with ordinary common sense opinion in our culture and with the opinion of the vast majority of philosophers and scientists who have published on the topic.⁵

³ In Chapter 8, I will define “consciousness” more rigorously.

⁴ Bourget and Chalmers 2014, interpreting “materialism” and “physicalism” equivalently.

⁵ Admittedly, the empirical literature on ordinary people's opinions about group consciousness is more equivocal than I would have thought: Knobe and Prinz 2008; Sytsma and Machery 2010; Arico 2010; Huebner, Bruno, and Sarkissian 2010; Phelan, Arico, and Nichols 2013; Huebner 2016. Group mentality without commitment to (or normally even discussion of) group consciousness, is commonly discussed in the scholarly literature. In social philosophy, see Gilbert 1989; Clark 1994; Bratman 1999; Rupert 2005; Tuomela 2007; Searle 2010; List and Pettit 2011; Huebner 2014; Epstein 2018; Overgaard and Salice 2019. In group psychology, see Le Bon 1895/1995; Bosanquet 1899/1923; McDougal 1920; Canetti 1960/1962; Tarrow 1994/2011; Wilson 2004. On sharing individual tokens (and not just types) of emotion, see Krueger 2013; Schmid 2014; León, Szanto, and Zahavi 2019.

Until recently, the few defenders of literal group-level consciousness of social groups have mostly been non-materialists: Espinas 1877/1924; Schäffle 1896; and especially Teilhard de Chardin 1955/1965; maybe Wundt 1897/1897; maybe Strawson 1959; and deeper in history maybe Averroes 12th c./2009 on the “active intellect”. Some recent materialists or near-materialists have defended group consciousness for specially structured hypothetical entities but not actually existing nations: Lycan 1981; Brooks 1986; Huebner 2014; Fekete, van Leeuwen, and Edelman 2016. A few others have made passing favorable remarks: Edelman 2008; Lewis and Viharo 2011; Koch 2012; Malone 2018. Pettit 2018 defends group level “coawareness”, which appears to imply group-level consciousness of the sort defended in this chapter, though that implication is open to interpretation. Lerner 2020 endorses nation-level group consciousness in the context of social science research on international relations. Works of science fiction with relatively detailed discussions of physically plausible group consciousness include Vinge 1992, 2011; Leckie 2013. Group consciousness might also be defensible in a

In this chapter, I will argue that accepting the materialist idea that you probably like (if you're a typical 21st century Anglophone philosopher) should lead you to accept some ideas about group consciousness that you probably don't like (if you're a typical 21st century Anglophone philosopher), unless you choose instead to accept some other ideas that you probably ought to like even less.

The argument in brief is this. If you're a materialist, you probably think that rabbits have conscious experiences. And you ought to think that. After all, rabbits are a lot like us, biologically and neurophysiologically. If you're a materialist, you probably also think that conscious experience would be present in a wide range of naturally evolved alien beings behaviorally very similar to us even if they are physiologically very different. And you ought to think that. After all, it would be insupportable Earthly chauvinism to deny consciousness to alien species behaviorally very similar to us, even if they are physiologically different. But, I will argue, a materialist who accepts consciousness in hypothetical weirdly formed aliens ought also to accept consciousness in spatially distributed group entities. If you then also accept rabbit consciousness, you ought also accept the possibility of consciousness in rather dumb group entities. Finally, the United States is a rather dumb group entity of the relevant sort (or maybe even it's rather smart, but that's more than I need for my argument). If we set aside our prejudices against spatially distributed group entities, we can see that the United States has all the types of properties that materialists normally regard as indicative of consciousness.

panpsychist context, according to which consciousness is ubiquitous in the universe, including possibly but perhaps only derivatively in groups: Roelofs 2019.

My claim is conditional and gappy. *If materialism is true, probably* the United States is conscious. Alternatively, if materialism is true, *the most natural thing to conclude* is that the United States is conscious.

1. Sirian Supersquids, Antarean Antheads, and Your Own Horrible Contiguism.

We are deeply prejudiced beings. Whites are prejudiced against Blacks, Gentiles against Jews, overestimators against underestimators.⁶ Even when we intellectually reject such prejudices, they can color our behavior and implicit assumptions.⁷ If we ever meet interplanetary travelers similar to us in overall intelligence and moral character, we will likely be prejudiced against them too, especially if they look strange.

It's hard to imagine a prejudice more deeply ingrained than our prejudice against entities that are visibly spatially discontinuous – a prejudice built, perhaps, even into the basic functioning of our visual system.⁸ Analogizing to racism, sexism, and speciesism, let's call such prejudice *contiguism*.

You might think that so-called contiguism is always justified and thus undeserving of a pejorative label. You might think, for example, that spatial contiguity is a necessary condition of objecthood or entityhood, so that it makes no more sense to speak of a spatially discontinuous entity than it makes sense – barring a very liberal ontology⁹ – to speak of an entity composed of

⁶ On the last, see Bettencourt et al. 1992.

⁷ Brownstein 2015/2019; Gawronski and Brannon 2017; Kurdi et al. 2019.

⁸ Especially if the entity's parts move on diverse trajectories. See Campbell 1958; Spelke et al. 1992; Scholl 2007; Carey 2009; Luria and Vogel 2014. See Barnett 2008 and Madden 2015 for philosophical arguments that we do not intuitively attribute consciousness to scattered objects and Chomanski 2019 for a discussion of how these issues relate to psychological work on commonly inducible illusions of bodily discontinuity.

⁹ See discussions in Elder 2011; Korman 2011/2020; Biro 2017.

your left shoe, the Eiffel Tower, and the rings of Saturn. If you'll excuse me for saying so, such an attitude is foolish provincialism! The contiguous creatures of Earth are not the only kinds of creatures there might be. Let me introduce you to two of my favorite possible alien species.

1.1. Sirian supersquids.

In the oceans of a planet orbiting Sirius lives a naturally-evolved creature with a central head and a thousand tentacles. It's a very smart creature – as smart, as linguistic, as artistic and creative as human beings are, though the superficial forms of its language and art differ from ours. Let's call these creatures "supersquids".

The supersquid's brain is not centrally located like our own. Rather, the supersquid brain is distributed mostly among nodes in its thousand tentacles, while its head houses digestive and reproductive organs and the like.¹⁰ However, despite the spatial distribution of its cognitive processes across its body, the supersquid's cognition is fully integrated, and supersquids report having a single, unified stream of conscious experience. Part of what enables their cognitive and experiential integration is this: Instead of relatively slow electrochemical nerves, supersquid nerves are reflective capillaries carrying light signals, something like Earthly fiber optics. The speed of these signals ensures the tight temporal synchrony of the cognitive activity shooting among their tentacular nodes.

Supersquids show all external signs of consciousness. They have covertly visited Earth, and one is a linguist who has mastered English well enough to be indistinguishable from an

¹⁰ Earthly cephalopods and other molluscs already tend somewhat in this direction. For a fascinating treatment, see Godfrey-Smith 2016b. In Chapter 10 I will explore the physiology, cognition, and possible low-grade consciousness of another type of mollusk, the common garden snail, most of whose neurons are in its posterior tentacles.

ordinary English speaker in verbal tests, including in discussions of consciousness.¹¹ Like us, the supersquids have communities of philosophers and psychologists who write eloquently about the metaphysics of experience, about emotional phenomenology, about their imagery and dreams. Any unbiased alien observer looking at Earth and looking at the supersquid home planet would see no good grounds for ascribing consciousness to us but not them. Some supersquid philosophers doubt that Earthly beings are genuinely phenomenally conscious, given our radically different physiological structure. (“What? *Chemical* nerves? How protozoan!”¹²) However, I’m glad to report that only a small minority holds that view.

[Illustration 2 (Caption: A Sirian supersquid philosopher): The supersquid is a nerdy, intellectual squid with a thousand tentacles (mostly attached but some detached). Picture it in an underwater library, reading a (waterproof) book titled “Are Humans Conscious?”. It looks puzzled. Several of its detachable tentacles are floating around doing various tasks: fetching a book from the shelves, holding a reading light, writing a letter, carrying a kelp snack, or similar.]

Here’s another interesting feature of supersquids: They can detach their limbs. To be detachable, a supersquid limb must be able to maintain homeostasis briefly on its own and suitable light-signal transceivers must occupy both the surface of the limb and the surface to which the limb is usually attached. Once the squids began down this evolutionary path, selective advantages nudged them farther along, revolutionizing their hunting and foraging. Two major subsequent adaptations were these: First, the nerve signals between the head and limb-surface

¹¹ In other words, the linguist passes the traditional Turing Test (Turing 1950). For discussion of the Turing Test and consciousness see Harnad 2003. For a recent adaptation addressing some methodological and metaphysical concerns, see Schneider 2019 (critiqued in Udell and Schwitzgebel 2021).

¹² For fictional portrayals of this regrettable attitude, see Bisson 1991/2008 and Frankish 2018.

transceivers shifted to wavelengths less readily degraded by water and obstacles. Second, the limb-surface transceivers evolved the ability to communicate directly among themselves without needing to pass signals through the head. Since the speed of light is negligible, supersquids can now detach arbitrarily many limbs and send them roving widely across the sea with hardly any disruption of their cognitive processing. The energetic costs are high, but they supplement their diet and use technological aids.

In this limb-roving condition, supersquid limbs are not wandering independently under local limb-only control, then reporting back. Limb-roving squids remain as cognitively integrated as do contiguous squids and as intimately in control of their entire spatially-distributed selves. Despite all the spatial intermixing of their limbs with those of other supersquids, each individual's cognitive processes remain private because each squid's transceivers employ a distinctive signature wave pattern. If a limb is lost, a new limb can be artificially grown and fitted, though losing too many limbs at once substantially degrades memory and cognitive function. The supersquids have begun to experiment with limb exchange and cross-compatible transceiver signals. This has led them toward what human beings might regard as peculiarly overlap-tolerant views of personal identity, and they have begun to re-envision the possibilities of marriage, team sports, and scientific collaboration.¹³

I hope you'll agree with me, and with the opinion universal among supersquids, that supersquids are coherent entities. Despite their spatial discontinuity, they aren't mere collections. They are integrated systems that can be treated as beings of the sort that might house consciousness. And if they might, they do. Or so you should probably say if you're a mainline

¹³ These last thoughts are inspired by Churchland 1981; Parfit 1984; Egan 1992; and Leckie 2013.

philosophical materialist. After all, supersquids are naturally evolved beings who act and speak and write and philosophize just like we do.

Does it matter that this is only science fiction? I hope you'll agree that supersquids, or entities relevantly similar, are at least *physically possible*. And if such entities are physically possible, and if the universe is as large as most cosmologists currently think it is – maybe even infinite, maybe even one among infinitely many infinite universes¹⁴ – then it might not be a bad bet that some such spatially distributed intelligences actually exist *somewhere*. Biology can be provincial, maybe, but not fundamental metaphysics – not any theory that aims, as ambitious general theories of consciousness do, to cover the full range of possible cases. You need room for supersquids in your universal theory of what's so.

1.2. *Antarean antheads*.

Among the green hills and fields of a planet near Antares dwells a species that looks like woolly mammoths but acts much like human beings. Gazing into my crystal ball, here's what I see: Tomorrow, they visit Earth. They watch our television shows, learn our language, and politely ask to tour our lands. They are sanitary, friendly, excellent conversationalists, and well supplied with rare metals for trade, so they are welcomed across the globe. They are quirky in a few ways, however. For example, they think at about one-tenth our speed. This has no overall effect on their intelligence, but it does test the patience of people unaccustomed to the Antareans' slow pace. They also find some tasks easy that we find difficult and vice versa. They are baffled and amused by our trouble with simple logic problems like the Wason Selection Task¹⁵ and

¹⁴ For example, Carroll 2010; Green 2011; Tegmark 2014.

¹⁵ Wason 1968, then a large subsequent literature. You see four cards. On two, you see letters: A on one, C on the other. On the other two you see numbers: 4 on one, 7 on the other.

tensor calculus, but they are impressed by our skill in integrating auditory and visual information.

Over time, some Antareans migrate permanently down from their orbiting ship. Patchy accommodations are made for their size and speed, and they start to attend our schools and join our corporations. Some achieve political office and display approximately the normal human range of virtue and vice. Although Antareans don't reproduce by coitus, they find some forms of physical contact arousing and have broadly human attitudes toward love-bonding. Marriage equality is achieved. What a model of interplanetary harmony! Ordinary non-philosophers all agree, of course, that Antareans are conscious.

Here's why I call them "antheads": Their heads and humps contain not neurons but rather ten million squirming insects, each a fraction of a millimeter across. Each insect has a complete set of tiny sensory organs and a nervous system of its own, and the antheads' behavior arises from complex patterns of interaction among these individually dumb insects. These mammoth creatures are much-evolved descendants of Antarean ant colonies that evolved in symbiosis with a brainless, living hive. The interior insects' interactions are so informationally efficient that neighboring insects can respond differentially to the behavioral or chemical effects of other insects' individual outgoing efferent nerve impulses. The individual ants vary in size, structure, sensoria, and mobility. Specialist ants have various affinities, antagonisms, and predilections,

You also know that each card has a letter on one side and a number on the other side. Thus, the lettered cards have hidden numbers on the back and the numbered cards have hidden letters on the back. What card or cards do you need to flip to test the following rule: If there's a consonant on one side, there's an even number on the other side. Most people get it wrong. In my experience, most people get it wrong even after being told to be careful because most people get it wrong. I find the test fascinating because it's so logically simple and yet we are so bad at it!

but no ant individually approaches human intelligence. No individual ant, for example, has an inkling of Shakespeare despite the Antareans' great fondness for Shakespeare's work.

There seems to be no reason in principle that such an entity couldn't execute any computational function that a human brain could execute or satisfy any high-level functional description that the human organism could satisfy, according to standard theories of computation and functional architecture.¹⁶ Every computable input-output relation and every medium-to-coarse-grained functionally describable relationship that human beings execute via patterns of neural excitation should be executable by such an anthead. Nothing about being an anthead should prevent Antareans from being as clever, creative, and strange as Earth's best scientists, comics, and artists, on standard materialist approaches to cognition.

Maybe there are little spatial gaps between the ants. Does it matter? Maybe, in the privacy of their homes, the ants sometimes disperse from the body, exiting and entering through the mouth. Does it matter? Maybe if the exterior body is badly damaged, the ants recruit a new body from nutrient tanks – and when they march off to do this, they retain some cognitive coordination, able to remember and later report thoughts they had mid-transfer. “Oh it's such a free and airy feeling to be without a body! And yet it's a fearful thing too. It's good to feel again the power of limbs and mouth. May this new body last long and well. Shall we dance, then, love?”

We humans are not *so* different perhaps. In one perspective, we ourselves are but symbiotic aggregates of simpler entities that invested in cooperation.¹⁷

¹⁶ For more on the theory of computation, see Chapter 5.

¹⁷ John Maynard Smith vividly and influentially defends such a perspective in Maynard Smith and Szathmáry 1995. See also Gilbert, Sapp, and Tauber 2012; Arnellos 2018.

The Sirian and Antarean examples establish the following claim well enough, I hope, that most materialists should accept it: At least some physically possible spatially scattered entities could reasonably be judged to be coherent entities with a unified stream of conscious experience.

2. Dumbing Down and Smarting Up.

You probably think that rabbits are conscious – that there’s “something it’s like” to be a rabbit, that rabbits feel pain, have visual experiences, and maybe have feelings like fear. Some philosophers would deny rabbit consciousness; more on that later. For now, I’ll assume you’re on board.

If you accept rabbit consciousness, you probably ought to accept consciousness in the Sirian and Antarean equivalents of rabbits. Consider, for example, the Sirian squidbits, a squidlike species with approximately the intelligence of rabbits. When chased by predators, squidbits will sometimes eject their limbs and hide their central heads. Most Sirians regard squidbits as conscious entities. Whatever reasoning justifies attributing consciousness to Earthly rabbits, parallel reasoning justifies attributing consciousness to Sirian squidbits. If humans are justified in attributing consciousness to rabbits due to rabbits’ moderate cognitive and behavioral sophistication, squidbits have the same types of moderate cognitive and behavioral sophistication. If, instead, humans are justified in attributing consciousness to rabbits due to rabbits’ physiological similarity to us, then supersquids are justified in attributing consciousness to squidbits due to squidbits’ physiological similarity to them. Antares, similarly, hosts antrabbits. If we accept this, we accept that consciousness can be present in spatially distributed entities that lack humanlike intelligence, a sophisticated understanding of their own minds, or linguistic reports of consciousness.

Let me knit together Sirius, Antares, and Earth: As the squidbit continues to evolve, its central body shrinks – thus easier to hide – and the limbs gain more independence, until the primary function of the central body is just reproduction of the limbs. Earthly entomologists come to refer to these heads as “queens”. Still later, squidbits enter into a symbiotic relationship with brainless but mobile hives, and the thousand bits learn to hide within for safety. These mobile hives look something like woolly mammoths. Individual fades into group, then group into individual, with no sharp, principled dividing line. On Earth, too, there is often no sharp, principled line between individuals and groups, though this is more obvious if we shed our obsession with vertebrates. Corals, aspen forests connected at the root, sea sponges, and networks of lichen and fungi often defy easy answers concerning the number of individuals. Opposition to group consciousness is more difficult to sustain if “group” itself is a somewhat arbitrary classification.¹⁸

We can also, if we wish, increase the size of the Antareans and the intelligence of the ants. Maybe Antareans are the size of shopping malls and filled with naked mole rats. This wouldn't seem to affect the argument. Maybe the ants or mole rats could even have human-level intelligence, while the Antareans' behavior still emerges in roughly the same way from the system as a whole. Again, this wouldn't seem to affect the argument (though see the Dretske/Kammerer objection in Chapter 2 Appendix, section 5).¹⁹

¹⁸ See especially Dupré 2012 on promiscuous individualism – the view that there is often more than one way to divide living systems into biological individuals, depending on one's time scale and practical interests.

¹⁹ We might also imagine a truly huge and slow entity, which shows all the signs and functions of consciousness if viewed on a scale of thousands of kilometers and a pace of months per action, and which is composed of people who through their communications knowingly enact, without fully comprehending, this large entity's cognitive processing and behavioral choices. I explore some possible cases in Schwitzgebel 2012a and Schwitzgebel 2017b.

The present view might seem to conflict with biological (or “type materialist”) views that equate human consciousness with specific biological processes. I don’t think it needs to conflict, however. Most such views allow that strange alien species might in principle be conscious, even if constructed rather differently from us. The experience of pain, for example, might be constituted by one biological process in us and a different biological process in a different species. Alternatively, the experience or “phenomenal character” of *pain*, in the specific manner it’s felt by us, might require Earthly neurons, while Antareans have conscious experiences of *schmain*, which feels very different but plays a similar functional role. Still another possibility is that whatever biological properties ground consciousness, those properties are sufficiently coarse or abstract that species with very different low-level structures (neurons vs. light signals vs. squirming bugs) can all equally count as possessing the required biological properties.²⁰

3. A Telescopic View of the United States.

A planet-sized alien who squints might see the United States as a single, diffuse entity consuming bananas and automobiles, wiring up communication systems, touching the Moon, and regulating its smoggy exhalations – an entity that can be evaluated for the presence or absence of consciousness.

You might object: Even if a Sirian supersquid or Antarean anthead is a coherent entity evaluable for the presence or absence of consciousness, the United States is not such an entity.

²⁰ See McLaughlin 2017 for a review of type materialism. For details of how some of the options described in this paragraph might play out, see Lewis 1980; Bechtel and Mundale 1999; Hill 2009; Polger and Shapiro 2016. Block 2002/2007 illustrates the skeptical consequences of embracing type-identity materialism without committing to a possibility of broadly this sort.

For example, it is not a biological organism. It lacks a life cycle. It doesn't reproduce. It's not an integrated system of biological materials maintaining homeostasis.

To this concern I have two replies.

First, it's not clear why being conscious should require any of those things. Properly-designed androids, brains in vats, gods – they aren't organisms in the standard biological sense, yet they are sometimes thought to be potential loci of consciousness. (I'm assuming materialism, but some materialists believe in actual or possible gods.) Having a distinctive mode of reproduction is often thought to be a central, defining feature of organisms, but it's not clear why reproduction should matter to consciousness. Human beings might vastly extend their lives and cease reproduction, or they might conceivably transform themselves technologically so that any specific condition on having a biological life cycle is dispensed with, while our brains and behavior remain largely the same. Would we no longer be conscious? Being composed of cells and organs that share genetic material might also be characteristic of biological organisms, but as with reproduction it's unclear why that should be necessary for consciousness.

Second, it's not clear that nations aren't biological organisms. The United States *is*, after all, composed of cells and organs that share genetic material, to the extent it is composed of people who are composed of cells and organs and who share genetic material. The United States maintains homeostasis. Farmers grow crops to feed non-farmers, and these nutritional resources are distributed via truckers on a network of roads. Groups of people organized as import companies draw in resources from the outside environment. Medical specialists help maintain the health of their compatriots. Soldiers defend against potential threats. Teachers educate future generations. Home builders, textile manufacturers, telephone companies, mail carriers, rubbish haulers, bankers, judges, all contribute to the stable well-being of the organism.

Politicians and bureaucrats work top-down to ensure the coordination of certain actions, while other types of coordination emerge spontaneously from the bottom up, just as in ordinary animals. Viewed telescopically, the United States is arguably a pretty awesome biological organism.²¹ Now, some parts of the United States are also individually sophisticated and awesome, but that subtracts nothing from the awesomeness of the U.S. as a whole – no more than we should be less awed by human biology as we discover increasing evidence of our dependence of microscopic symbionts.

Nations also reproduce – not sexually but by fission. The United States and several other countries are fission products of Great Britain. In the 1860s, the United States almost fissioned again. And fissioning nations retain traits of the parent that enhance the fitness of future fission products – intergenerationally stable developmental resources, if you will. As in cellular fission, there's a process by which subparts align on different sides and then separate physically and functionally.

Even if you don't accept that the United States is literally a biological organism, you still probably ought to accept that it has sufficient organization and coherence to qualify as a concrete though scattered entity of the sort that can be *evaluated* for the presence or absence of consciousness. On Earth, at all levels, from the molecular to the neural to the social, there's a vast array of competitive and cooperative pressures; at all levels, there's a wide range of actual and possible modes of reproduction, direct and indirect; and all levels show manifold forms of mutualism, parasitism, partial integration, agonism, and antagonism.²² There isn't as radical a

²¹ See Stock 1993 for a similar perspective in lively detail. On Godfrey-Smith's 2013 three-dimensional taxonomy of "Darwinian individuals", the United States would appear to be an intermediate case, comparable to a sponge.

²² Shapiro 2007; Dupré 2012; Gilbert, Sapp, and Tauber 2012; Trewavas 2014; Figdor 2018; Nicholson and Dupré 2018.

difference in kind as people are inclined to think between our favorite level of organization and higher and lower levels.

I'm asking you to think of the United States as a planet-sized alien might: as a concrete entity composed of people (and maybe some other things), with boundaries, inputs, outputs, and behaviors, internally organized and regulated.

We can now ask our main question: Is this entity *conscious*? More specifically, does it meet the criteria that mainstream scientific and philosophical materialists ordinarily regard as indicative of consciousness?

If those criteria are applied fairly, without prejudice, it does appear to meet them, as I will now endeavor to show.

4. *What Is So Special about Brains?*

According to mainstream philosophical and scientific materialist approaches to consciousness, what's really special about us is our brains. Brains are what make us conscious. Maybe brains have this power on their own, so that even a bare brain in an otherwise empty universe would have conscious experiences if it was structured in the right way; or maybe consciousness arises not strictly from the brain itself but rather from a thoroughly entangled mix of brain, body, and environment.²³ But all materialists agree: Brains are central to the story.

Now what is so special about brains, on this view? Why do brains give rise to conscious experience while a similar mix of chemicals in chicken soup does not? It must be something about how the materials are organized. Two general features of brain organization stand out:

²³ For example, Hurley 1998; Noë 2004; Wilson 2004; Rockwell 2005; Kirchhoff and Kiverstein 2019.

their complex high order / low entropy information processing, and their role in coordinating sophisticated responsiveness to environmental stimuli. These two features are of course related. Brains also arise from an evolutionary and developmental history, within an environmental context, which might play a constitutive role in determining function and cognitive content.²⁴ According to a broad class of materialist views, any system with sophisticated enough information processing and environmental responsiveness, and perhaps the right kind of historical and environmental embedding, should have conscious experiences. The central claim of this chapter is: The United States seems to have what it takes, if standard materialist criteria are straightforwardly applied without post-hoc noodling. It is mainly unjustified morphological prejudice that prevents us from seeing this.

Consider, first, the sheer quantity of information transfer among members of the United States. The human brain contains about a hundred billion neurons exchanging information through an average of about a thousand connections per neuron, firing at peak rates of several hundred times a second. The United States, in comparison, has only about three hundred million people. But those people exchange a lot of information. How much? We might begin by considering how much information flows from one person to another by stimulation of the retina. The human eye contains about a hundred million photoreceptor cells. Most people in the United States spend most of their time in visual environments that are largely created by the actions of people (including their past selves). If we count even one three-hundredth of this visual neuronal stimulation as the relevant sort of person-to-person information exchange, then the quantity of visual connectedness among people is similar to the neural connectedness of the brain (a hundred

²⁴ As on “externalist” views of cognitive content such as Putnam 1975; Burge 1979; Millikan 1984, 2017; Dretske 1988, 1995; Neander 2017.

trillion connections). Very little of this exchanged information makes it past attentional filters for further processing, but analogous considerations apply to information exchange among neurons. Or here's another angle: If at any time one three-hundredth of the U.S. population is viewing internet video at one megabit per second, that's a transfer rate among people of a trillion bits per second in this one minor activity alone.²⁵ Furthermore, it seems unlikely that conscious experience requires achieving the degree of informational connectedness of the entire neuronal structure of the human brain. If mice are conscious, they manage it with under a hundred million neurons.

A more likely source of concern, it seems to me, is that the information exchange among the U.S. population isn't of the right *type* to engender a genuine stream of conscious experience. A simple computer download, even if it somehow managed to involve a hundred trillion bits per second, presumably wouldn't by itself suffice. For consciousness, presumably there must be some organization of the information in service of coordinated, goal-directed responsiveness; and maybe, too, there needs to be some sort of sophisticated self-monitoring.

But the United States has these properties too. The population's information exchange is not in the form of a simply-structured internet download. The United States is a goal-directed entity, flexibly self-protecting and self-preserving. The United States responds, intelligently or semi-intelligently, to opportunities and threats – not less intelligently than a small mammal. The United States expanded west as its population grew, developing mines and farms in traditionally Native American territory. When Al Qaeda struck New York, the United States responded in a

²⁵ See also Moravec 1997; Kurzweil 2005; Hilbert and López 2011. It is probably too simplistic to conceptualize the connectivity of the brain as though all that mattered were neuron-to-neuron connections; but those who favor complex models of the internal activity of the brain should probably also for similar reasons favor complex models of interactivity among citizens and residents of the United States.

variety of ways, formally and informally, in many branches of government and in the population as a whole. Saddam Hussein shook his sword and the United States invaded Iraq. The U.S. acts in part through its army, and the army's movements involve perceptual or quasi-perceptual responses to inputs: The army moves around the mountain, doesn't crash into it. Similarly the spy networks of the CIA detected the location of Osama bin Laden, whom the U.S. then killed. The United States monitors space for asteroids that might threaten Earth. Is there less information, less coordination, less intelligence than in a hamster? The Pentagon monitors the actions of the Army, and its own actions. The Census Bureau counts residents. The State Department announces the U.S. position on foreign affairs. The Congress passes a resolution declaring that Americans hate tyranny and love apple pie. This is self-representation. Isn't it? The United States is also a social entity, communicating with other entities of its type. It wars against Germany, then reconciles, then wars again. It threatens and monitors North Korea. It cooperates with other nations in threatening and monitoring North Korea. As in other linguistic entities, some of its internal states are well known and straightforwardly reportable to others (who just won the Presidential election, the approximate unemployment rate) while others are not (how many foreign spies have infiltrated the CIA, why the population consumes more music by Elvis Presley than Ella Fitzgerald).

One might think that for an entity to have real, intrinsic representations and meaningful utterances, it must be richly historically embedded in the right kind of environment. Lightning strikes a swamp and "Swampman" congeals randomly by freak quantum chance. Swampman might utter sounds that we would be disposed to interpret as meaning "Wow, this swamp is humid!", but if he has no learning history or evolutionary history, some have argued, this utterance would have no more meaning than a freak occurrence of the same sounds by a random

perturbation of air.²⁶ But I see no grounds for objection here. The United States is no Swampman. The United States has long been embedded in a natural and social environment, richly causally connected to the world beyond – connected in a way that would seem to give meaning to its representations and functions to its parts.²⁷

I am asking you to think of the United States as a planet-sized alien might, that is, to evaluate the behaviors and capacities of the United States as a concrete, spatially distributed entity with people as some or all of its parts – an entity in which people play roles somewhat analogous to the roles that individual cells play in your body. If you are willing to jettison contiguism and other morphological prejudices, this is not, I think, an intolerably strange perspective. As a house for consciousness, a rabbit brain is not clearly more sophisticated. I leave it open whether we include objects like roads and computers as part of the body of the U.S. or instead as part of its environment.

The representations and actions of the United States all presumably depend on what's going on among the people of the United States. In some sense, arguably, its representations and actions reduce to, and are nothing but, patterns of activity among its people and other parts (if any). Yes, right, and granted! But if materialism is true, something similar can be said of you. All of *your* representations and actions depend on, reduce to, are analyzable in terms of, and are nothing but, what's going on among your parts, for example, your cells. This doesn't make you non-conscious. As long as these lower-level events hang together in the right way to contribute to the whole, you're conscious. Materialism as standardly construed just is the view on which consciousness arises at a higher level of organization (e.g., the person) when lower-level parts

²⁶ See Davidson 1987; Dretske 1995; Millikan 2010; Neander 2016.

²⁷ In this respect, the case of the United States is importantly different from the more artificial cases discussed in Block 1978/2007; Lycan 1981; and Brooks 1986.

(e.g., brain cells) interact in the right way. Maybe everything ultimately comes down to and can in principle be fully understood as nothing but the activity of a few dozen fundamental particles in massively complex interactions.²⁸ The reducibility of X to Y does not imply the non-consciousness of X.²⁹ On standard materialist views, as long as there are the right functional, behavioral, causal, informational, etc., patterns and relationships in X, it detracts not a whit that it can all be explained in principle by the buzzing about of the smaller-scale stuff Y that composes X.

I'm not arguing that the United States has any exotic consciousness juice, or that its behavior is in principle inexplicable in terms of the behavior of people, or anything fancy or metaphysically complicated like that. My argument is really much simpler: There's something awesomely special about brains such that they give rise to consciousness, and if we examine the standard candidate explanations of what makes brains special, the United States seems to be special in just the same sorts of ways.

What is it about brains, as hunks of matter, that makes them so amazing? Consider what materialist scientists and philosophers tend to say in answer: sophisticated information processing, flexible goal-directed environmental responsiveness, representation, self-representation, multiply-ordered layers of self-monitoring and information-seeking self-

²⁸ Here I bracket non-reductionist materialists such as Davidson 1970/2001. Most materialists are at least "in principle" reductionists, or "theological reductionists" as Dupré 1983 terms them, in the sense that even if they think the world is too complex ever to permit reductive explanation by actual human scientists, in principle an omniscient deity could understand in detail how psychological states arise from the activity of fundamental particles. See also Kim 1989.

²⁹ This allows me to, I hope, dodge the debate in social philosophy about whether group behavior or mental states can be reduced to, or explained entirely in terms of, the behavior and mental states of individual members of the group. This chapter in no way turns on embracing an anti-reductionist answer to that question. See the social philosophy references in note 4.

regulation, rich functional roles, a content-giving historical embeddedness. The United States has all those same features. In fact, it seems to have them to a greater degree than do some entities, like rabbits, that we ordinarily regard as conscious.

5. *Three Ways Out.*

Let me briefly consider three more conservative views about the distribution of consciousness in the universe, to see if they can provide a suitable exit from the conclusion that the United States literally has conscious experiences.

5.1. Consciousness isn't real.

Maybe the United States isn't conscious because *nobody* is conscious – not you, not me, not rabbits, not aliens. Maybe “consciousness” is a corrupt, broken concept, embedded in a radically false worldview, and we should discard it entirely, as we discarded the concepts of demonic possession, the luminiferous ether, and the fates.³⁰

In this chapter, I've tried to use the concept of consciousness in a plain way, unburdened with dubious commitments like irreducibility, immateriality, or infallible self-knowledge. In Chapter 8, I will further clarify what I mean by consciousness in this hopefully innocent sense. But let's allow that I might have failed. Permit me, then, to rephrase: Whatever it is in virtue of which human beings and rabbits have quasi-consciousness or consciousness* (appropriately unburdened with dubious commitments), the United States has that same thing.

³⁰ Some philosophers have argued that objectionable commitments are built into the very notion of “consciousness” or “phenomenal consciousness” and thus that no such thing exists: Feyerabend 1963; P.S. Churchland 1983; Frankish 2016; Kammerer 2021; for reviews, see Irvine and Sprevak 2020 and Niikawa 2021. In Schwitzgebel 2020a, I argue that eliminativists only make their arguments plausible by defining “consciousness” in an inflated way.

The most visible philosophical eliminativists about terms from everyday “folk psychology” still seem to have room in their theories for consciousness, suitably stripped of objectionable epistemic or metaphysical commitments.³¹ So if you take this path, you’re going farther than they. Denying that consciousness exists at all seems at least as bizarre as believing that the United States is conscious.

5.2. Extreme sparseness.

Here’s another way out: Argue that consciousness is rare, so that really only very specific types of systems possess it, and then argue that the United States doesn’t meet the restrictive criteria. If the criteria are specifically neural, this position is *neurochauvinism*, which I’ll discuss in Section 5.3. Setting aside neurochauvinism, the most commonly endorsed extreme sparseness view is one in which language is required for consciousness. Thus, dogs, wild apes, and human infants aren’t conscious. There’s nothing it’s like to be such beings, any more than there’s something it’s like (most people think) to be chicken soup or a fleck of dust. To a dog, all is dark inside, or rather, not even dark. This view is no less seemingly bizarre than accepting that the U.S. is conscious. Like the consciousness-isn’t-real response, it trades one bizarreness for another. It is no escape from the quicksand of weirdness. It is also, I suspect, a serious overestimation of the gulf between us and our nearest relatives.³²

³¹ P.M. Churchland 1984/1988; P.S. Churchland 2002; Stich 2009; Frankish 2016 (in his reply to Schwitzgebel 2016b). P.M. Churchland even says several things that seem, jointly, to commit him to the idea that cities or countries would be conscious (though he doesn’t to my knowledge explicitly draw the conclusion): See his characterizations of life and consciousness on pages 173 and 178 of his 1984/1988.

³² For discussion see Allen and Trestman 1995/2016; Andrews 2015.

Moreover, it's not clear that requiring language for consciousness actually delivers the desired result. The United States does seemingly speak as a collective entity, as I've mentioned. It linguistically threatens and self-represents, and these threats and self-representations influence the linguistic and non-linguistic behavior of other nations.

5.3. Neurochauvinism.

A third way out is to assume that consciousness requires *neurons* – neurons bundled together in the right way, communicating by ion channels and all that, rather than by voice and gesture. All the entities that we have actually met and that we normally regard as conscious do have their neurons bundled in that way, and the 3×10^{19} neurons of the United States are not as a whole bundled that way.

Examples from Ned Block and John Searle lend intuitive support to this view.³³ Suppose we arranged the people of China into a giant communicative network resembling the functional network instantiated by the human brain. It would be absurd, Block says, to regard such an entity as conscious.³⁴ Similarly, Searle asserts that no arrangement of beer cans, wire, and windmills, however cleverly structured, could ever host a genuine stream of conscious experience.³⁵ According to Block and Searle, what these entities are lacking isn't a matter of large-scale functional structure of the sort that is revealed by input-output relationships, responsiveness to an external environment, or coarse-grained functional state transitions. Consciousness requires not that, or not only that. Consciousness requires human biology.

³³ Block 1978/2007; Searle 1980, 1984.

³⁴ Though see Lycan 1981 for a reply.

³⁵ Though see Cuda 1985 for a reply.

Or rather, consciousness on this view requires something *like* human biology. In what way like? Here Block and Searle aren't very helpful. According to Searle, "any system capable of causing consciousness must be capable of duplicating the causal powers of the brain".³⁶ In principle, Searle suggests, this could be achieved by "altogether different" physical mechanisms. But what mechanisms could do this and what mechanisms could not, Searle makes no attempt to adjudicate, other than by excluding certain systems, like beer-can systems, as plainly the wrong sort of thing. Instead, Searle gestures hopefully toward future science.

The reason for not insisting strictly on neurons, I suspect, is this: If we're playing the common sense game – that is, if bizarreness by the standards of current common sense is our reason for excluding beer-can systems and organized groups of people – then we're going to have to allow the possibility, at least in principle, of conscious beings from other planets who operate other than by neural systems like our own. By whatever commonsense or intuitive standards we judge beer-can systems nonconscious, by those very same standards, it seems, we would judge at least some hypothetical Martians, with different internal biology but intelligent-seeming outward behavior, to be conscious.

From a cosmological perspective it would be strange to suppose that of all the possible beings in the universe that are capable of sophisticated, self-preserving, goal-directed environmental responsiveness, beings that could presumably be (and in a vast enough universe presumably actually are) constructed in myriad strange and diverse ways, somehow only we with

³⁶ Searle 1992, p. 92.

our neurons have genuine conscious experience, and all the rest are mere empty shells, so to speak.³⁷

If they're to avoid un-Copernican³⁸ neuro-fetishism, the question must become, for Block and Searle, *what* feature of neurons, possibly also possessed by non-neural systems, gives rise to consciousness? In other words, we're back with the question of Section 4: What is so special about brains? And the only well-developed answers on the near horizon seem to involve appeals to features that the United States has, like massively complex informational integration, functional self-monitoring, and a long-standing history of sophisticated environmental responsiveness.

6. Conclusion.

In sum, the argument is this. There is no principled reason to deny psychological properties to spatially distributed beings if they are sufficiently integrated in other ways. By this criterion, the United States is at least a *candidate* for the literal possession of real psychological states, including consciousness. If we're willing to entertain this perspective, the question then becomes whether the U.S. meets plausible criteria for consciousness, according to the usual standards of mainstream philosophical and scientific materialism. My suggestion is that if those

³⁷ However, see Lee 2019 for a perspective according to which alien entities might lack consciousness but instead have quasi-consciousness, similar enough to consciousness for moral purposes.

³⁸ According to the “Copernican Principle” of mainstream cosmology, we should assume that we are not in any particularly special region in the universe, such as its exact center. In Schwitzgebel 2020b, I suggest that application of this principle extends to the assumption that there is no unexplained lucky relationship between our cognitive sophistication and our consciousness. Among all the actual or hypothetical species capable of sophisticated cognitive and linguistic behavior, it would be un-Copernican to suppose that we are among the special small portion who also have conscious experiences.

criteria are liberal enough to include both small mammals and highly intelligent aliens, then the United States probably does meet those criteria.

You still have an objection? I'm unsurprised. Consult the appendix of this chapter where I address six. This chapter argues for a thesis that most people are inclined to resist, inspiring the creative construction of counterarguments.

I would have liked to strengthen the arguments of this chapter by applying particular, detailed materialist metaphysical theories to the question at hand, showing how each does (or does not) imply that the United States is literally conscious. Unfortunately, any attempt to do so presents four obstacles, in combination nearly insurmountable. First: Few materialist theoreticians explicitly discuss the plausibility of literal group consciousness.³⁹ Thus, it's a matter of speculation how to properly apply their theory to a case that might have been overlooked in the theory's design and presentation. Second: Many theories, especially those by neuroscientists and psychologists, implicitly or explicitly limit themselves to *human* consciousness or at most consciousness in entities with neural structures like ours, and thus are silent about how consciousness might work in other types of entities.⁴⁰ Third: Further limiting

³⁹ Notable exceptions include Lycan 1981; Brooks 1986; Wilson 2004; Huebner 2014; Tononi and Koch 2015; Pacherie 2017; List 2018; Roelofs 2019. Tononi and Koch and on similar grounds List deny the group consciousness of nations (see note XXX below). Lycan, Brooks, and Huebner endorse hypothetical group consciousness under certain counterfactual conditions (e.g., Brooks's "Brain City" in which people mimic the full neuronal structure of the brain), while refraining from stating that their arguments extend to any group entities that actually exist. Wilson I am inclined to read as rejecting group consciousness on the grounds that it had been advocated only sparsely and confusedly, with no advocate meeting a reasonable argumentative burden of proof. Pacherie allows that groups might be conscious if certain general theories of consciousness are correct, but we do not yet know if such theories are correct. Roelofs allows that within a panpsychist framework groups may have consciousness, but Roelofs adds that to have anything more than faint, hazy, or blurred experience the group would need more structure than currently existing groups tend to have.

⁴⁰ For example, Baars 1988; Crick 1994; Prinz 2012; Dehaene 2014.

the pool of relevant theories is the fact that few thinkers really engage the question from top to bottom, including all of the details that would be relevant to assessing whether the U.S. would literally be conscious according to their theories.⁴¹ Fourth: When first working through my thoughts on this topic I arrived at what I thought would be a representative sample of four prominent, metaphysically ambitious, top-to-bottom theories of consciousness, it proved rather complex to assess how each view applied to the case of the U.S. – too complex to tack to the end of an already-long chapter.⁴²

Thus, I think further progress on this issue will require having some specific proposals to evaluate, that is, some ambitious, general materialist theories of consciousness that address the question of group consciousness in a serious and careful way, with enough detail that we can assess the theory's implications for specific, real groups like the United States. No theorist has yet, to my knowledge, risen to the occasion.

Large things are hard to see properly when you're in their midst. Too vivid an appreciation of the local mechanisms overwhelms your view. In the 18th century, Leibniz imagined entering into an enlarged brain and looking around as if in a mill. You wouldn't be

⁴¹ For example, most theoreticians advocating “higher order” models of consciousness (in which a mental state is conscious if the organism simultaneously has the right kind of “higher order” representation of that mental state) don't provide sufficient detail on the nature of “lower order” mental states for me to evaluate whether the United States would qualify as having such lower-order mental states – though if it does, it would probably have the higher-order states too. For a review of higher-order theories see Carruthers and Gennaro 2001/2020.

⁴² The theories I chose were Humphrey's, Dennett's, and Tononi's pre-2012 view. You can see some of my preliminary efforts in several blog posts: Schwitzgebel 2012a,d,e,f,g,h. See also Koch's sympathetic 2012 treatment of pre-2012 Tononi (which might stand in tension with their later joint work in Tononi and Koch 2015). On the most natural interpretations of these four test-case views, I thought that readers sympathetic with any of these authors' general perspectives ought to accept that the United States is conscious. And I confess I still do think that, despite protests from Humphrey, Dennett, Dretske, and Tononi themselves in personal communication. See the comments section of Schwitzgebel 2012f for Humphrey's reaction, and Chapter 2 Appendix, Section 1, 3, and 5 for Tononi, Dennett, and Dretske.

inclined, he says, to attribute consciousness to the mechanisms.⁴³ Leibniz intended this as an argument against materialism, but the materialist could respond that the scale is misleading. Our intuitions about the mill shouldn't be trusted. Parallel reasoning might explain some of our reluctance to attribute consciousness to the United States. The space between us is an airy synapse.

If the United States is conscious, is Google? Is an aircraft carrier?⁴⁴ And if such entities are conscious, do they have rights? I don't know, but if we continue down this argumentative path, I expect this will prove quite a mess.

Then get off the path, you might say! As I see it, we have four choices.

First: We could accept that the United States is probably conscious. Maybe this is where our best philosophical and scientific theories most naturally lead us – and if it seems bizarre or contrary to common sense, so much the worse for most ordinary people's intuitive sense of what is plausible about consciousness.

Second: We could reject materialism. We could grant that mainstream materialist theories generate either the result that the United States is literally conscious or some other implausible seeming result (one of the various seemingly unattractive ways of escaping that conclusion, discussed above), and we could treat this as a reason to doubt that whole class of theories. In Chapters 3 and 5, I'll discuss and partially support some alternatives to materialism. However, as I'll argue there, if you're drawn this direction in hopes of escaping bizarreness, that's futile. You're doomed.

⁴³ Leibniz 1714/1989.

⁴⁴ Hutchins 1995 vividly portrays distributed cognition in a military vessel. However, I don't know whether he would extend his conclusions to phenomenal consciousness.

Third: We could treat this argument as a challenge to which materialists might rise. The materialist might accept one of the escapes I mention, such as denying rabbit consciousness (see also Chapter 10), accepting neurochauvinism, accepting an anti-nesting principle (see section 1 of the appendix to this chapter), or denying that the U.S. is a concrete entity, and then work to defend that view. Alternatively, the materialist might devise a new, attractive theory that avoids all of the difficulties I've outlined here. Although this is not at all an unreasonable strategy, it is unlikely for reasons I'll explain in Chapter 3 that such a view will be entirely free of bizarre-seeming implications of one sort or another. The metaphysics of consciousness will have weird consequences whichever way you turn.

Fourth: We could go *quietist*. We could say so much the worse for theory. Of course the United States is not conscious, and of course humans and rabbits are conscious. The rest is speculation, and if that speculation turns us in circles, well, as Ludwig Wittgenstein suggested, speculative philosophical theorizing might be more like an illness requiring cure than an enterprise we should expect to deliver truths.⁴⁵

Oh, how horrible that fourth reaction is! Take any of the other three, please. Or take, as I prefer, some indecisive superposition of those three. But do not tell me that speculating as best we can about consciousness and the basic structure of the universe is a disease.

⁴⁵ Wittgenstein interpretation is notoriously fraught. For discussion of the scope and nature of Wittgenstein's quietism, see Cray and Read 2000; McDowell 2009.

Chapter Two Appendix: Six Objections.

1. *Objection 1: Anti-Nesting Principles.*

Here's one way out of the potentially unappealing conclusion of Chapter 2: Evoke a general principle according to which a conscious organism cannot have conscious parts – what I will call an *anti-nesting principle*. Anti-nesting says that if consciousness arises at one level of organization in a system, it cannot also arise at any smaller or larger levels of organization.⁴⁸ We know that we are conscious. It would then follow from anti-nesting that no larger entity that contains us as parts, like the United States, would also be conscious.

Anti-nesting principles have rarely been discussed or evaluated in detail.⁴⁹ I am aware of two influential articulations of general anti-nesting principles in the philosophical and scientific literature. Both articulations are only thinly defended, almost stipulative, and both carry consequences approximately as weird as the group-level consciousness of the United States.

The first is due to the philosopher Hilary Putnam. In an influential series of articles in the 1960s, Putnam described and defended a functionalist theory of the mind, according to which having a mental state is just a matter of having the right kinds of relationships among states that are definable functionally, that is, causally or informationally.⁵⁰ Crucially, on Putnam's

⁴⁸ If we apply anti-nesting strictly, we will have to commit to denying that you as an organism can be conscious at the same time your brain (a subpart of you as an organism) is conscious. Anti-nesting implies that, strictly speaking, human consciousness transpires at one particular level of organization (for example, the whole brain) and attributing it to larger entities (the entire human organism) or smaller entities (the visual cortex) is sloppy. See Unger 1999; Merricks 2003; Sider 2003; Noë 2009; Sutton 2014; Fekete, van Leeuwen, and Edelman 2016; Simon 2017; Mørch 2019; Roelofs 2019; Blackmon 2021.

⁴⁹ Kammerer 2015 frames his objection to U.S. consciousness as depending on a “Sophisticated Anti-Nesting Principle” but I think it is better conceptualized as similar to Dretske's objection than as a general opposition to nesting. Therefore, I treat it in section 5.

⁵⁰ Papers collected in Putnam 1975; but see Putnam 1988 for some of Putnam's own later concerns about this view.

functionalism, it doesn't matter what sorts of material structures implement the functional relationships. Consciousness arises whenever the right sorts of causal or informational relationships are present, whether in human brains, in very differently structured octopus brains, in computers made of vacuum tubes or integrated circuits, in hypothetical aliens, or even in ectoplasmic soul stuff. Roughly speaking, any entity that acts and processes information in a sophisticated enough way is conscious. Putnam's pluralism on this issue was hugely influential, and functionalism of one stripe or other is probably the standard view about consciousness in academic philosophy. It's an attractive view for those of us who suspect that complex intelligence is likely to arise more than once in the (probably vast or infinite) universe and that what matters to consciousness is not its implementation by neurons but rather the presence of appropriately sophisticated behavior and informational processing.

Putnam's functionalist approach to consciousness is simple and elegant: *Any* system that has the right sort of functional structure is conscious, no matter what it's made of. Or rather – Putnam surprisingly adds – any system that has the right sort of functional structure is conscious, no matter what it's made of *unless it is made of other conscious systems*. This is the sole exception to Putnam's otherwise liberal attitude about the composition of conscious entities. A striking qualification! However, Putnam offers no argument for this qualification apart from the fact that he wants to rule out “swarms of bees as single pain-feelers”.⁵¹ Putnam never explains why single-pain-feeling is impossible for actual swarms of bees, much less why no possible future evolutionary development of a swarm of conscious bees could ever also be a single pain-feeler. Putnam embraces a general anti-nesting principle, but he offers only a brief, off-the-cuff defense of that principle.

⁵¹ Putnam 1967, p. 43.

The other prominent advocate of a general anti-nesting principle is more recent: the neuroscientist Giulio Tononi (and his collaborators). In a series of articles, Tononi advocates the view that consciousness arises from the integration of information, with the amount of consciousness in a system being a complex function of the system's informational structure. Tononi's theory, Integrated Information Theory (IIT), is influential – though also subject to a range of, in my judgment, rather serious objections.⁵² One aspect of Tononi's theory, introduced starting in 2012, is an “Exclusion Postulate” according to which whenever one informationally integrated system contains another, consciousness occurs only at the level of organization that integrates the most information. Tononi is thus committed to a general anti-nesting principle about consciousness.

Tononi's defense of the Exclusion Postulate is not much more substantive than Putnam's defense of his anti-nesting principle. Perhaps this is excusable: It's a “postulate” and postulates are often offered without much defense, in hopes that they will be retrospectively justified by the long-term fruits of the larger system of axioms and postulates to which they belong (compare Euclid's geometry). While we await the long-term fruits, though, we can still ask whether the postulate is independently plausible – and Tononi does have a few brief things to say. First, he says that the Exclusion Postulate is defensible by Occam's Razor, the famous principle forbidding us to “multiply entities beyond necessity”, that is, not to admit more than is strictly required into our ontology or catalog of what exists. Second, Tononi suggests that it's unintuitive to suppose that group consciousness would arise from two people talking.⁵³

⁵² Aaronson 2014; Schwitzgebel 2014c; Bayne 2018; Doerig, Schurger, Hess, and Herzog 2019; Barrett and Mediano 2019; Hanson 2020; Michel 2021.

⁵³ Both arguments appear in Tononi 2012, p. 304. In earlier work (Tononi 2010, note 9) discusses an anti-nesting principle without endorsing it, saying that such a principle is “in line with the intuitions that each of us has a single, sharply demarcated consciousness”. In his more

However, no advocate of IIT should rely in this way on ordinary intuitions about whether small amounts of consciousness would be present when small amounts of information are integrated: IIT attributes small amounts of consciousness even to simple logic gates and photodiodes if they integrate small amounts of information.⁵⁴ Why not some low-level consciousness from the group too? And Occam's razor is a tricky implement. Although admitting unnecessary entities into your ontology seems like a bad idea, what's an "entity" and what's "unnecessary" is often unclear, especially in part-whole cases. Is a hydrogen atom necessary once you admit the proton and electron into your ontology? What makes it necessary, or not, to admit the existence of consciousness in the first place? It's obscure why the necessity of attributing consciousness to Antarean antheads or the United States should depend on whether it's also necessary to attribute consciousness to the individual ants or people.

Furthermore, anti-nesting principles like Putnam's and Tononi's, though seemingly designed to avoid the bizarreness of group consciousness, bring different bizarre consequences in their train. As Ned Block argues against Putnam, such principles appear to have the unintuitive consequence that if ultra-tiny conscious organisms were somehow to become incorporated into your brain – perhaps, for reasons unknown to you, playing the role of one neuron or one part of one neuron – you would be rendered nonconscious, even if all of your behavior and all of your cognitive functioning, including your reports about your experiences, remained the same.⁵⁵ IIT and presumably any other information-quantity-based anti-nesting principle would also have the

recent articles, Tononi does not repeat that defense of the Exclusion Postulate – though he does mention that idea in connection with the quite different and in this context confusingly named Exclusion *Axiom*. For one possible exception to anti-nesting within the framework of IIT, see Tononi and Koch 2015, note 7, which allows nesting “as long as there is no causal overlap at the relevant scales”.

⁵⁴ Tononi 2004, 2008; Oizumi, Albantakis, and Tononi 2014.

⁵⁵ Block 1978/2007.

further bizarre consequence that we could all lose consciousness by having the wrong sort of election or social networking application. Imagine a large election, with many ballot measures.

Organized in the right way, the amount of integrated information could be arbitrarily large.⁵⁶

Tononi's Exclusion Postulate would then imply that all of the individual voters would lose

⁵⁶ The amount of information Φ is computationally intractable for most systems with more than a dozen nodes, but we could imagine that the citizens form a sufficiently large expander graph (Aaronson 2014) or maybe we could imagine a network in which some informational states encode millions or billions of bits of information in a single state, instead of the 1 or 0 used in most simple models of Φ (e.g., citizen 1 voted A, B, C, D; citizen 2 voted -A, B, -C, D; ...), potentially amplifying the total Φ values. Tononi and Koch (2015) – perhaps reacting to my arguments in Schwitzgebel 2012f and 2015c – assert that the United States would necessarily have a lower Φ than its citizens, but it's unclear how they can justify this claim. As long as there is *any* realistic way of structuring and describing the U.S. as a system, at any spatial or temporal grain, that generates a higher Φ than the highest Φ person in the system, my argument would carry. (Given the high sensitivity of Φ to differences in structural detail, no inference can be justified from their simple and idealized figure 5a to the case of the United States.) Hoel, Albantakis, and Tononi (2013) argue that macro-level systems can have higher Φ than lower level systems, especially when they contain “heavily interconnected groups of elements with spontaneous activity and the ability to distinguish between intragroup and intergroup connections” (p. 19795), which seems to describe the U.S. case rather well. List (2018; see also Pacherie 2017) carries the Tononi and Koch objection a bit farther but still relies on undefended claims about relative independence and interdependence among people and its likely relation to Φ – claims which furthermore risk ruling out unitary consciousness at the human level if human cognitive subsystems are relatively informationally independent, feeding their outputs into a Global Workspace (see discussion in section 4).

Note also this odd result (from Schwitzgebel 2014c): Compose a system containing people as subparts where the highest Φ person (Max) has a Φ of X, the system as a whole has a Φ of X-1, and every other person has a Φ of at most X-2. Assume also that Φ decreases during dreamless sleep (as Tononi claims). By Exclusion, the system is not conscious, since it has a subsystem (Max) with higher Φ . Now imagine that Max goes to sleep and his Φ falls to X-2. Suddenly, the highest Φ belongs to the system as a whole. By Exclusion every single person in the system now loses consciousness – even though they might continue to function and self-report normally and have no way of knowing whether Max is asleep or awake. If Max then wakes, the situation changes again. The root problem here is that there is no principled reason to think that minor changes in Φ values need have any major effects on the functioning of the various subsystems, but given the threshold nature of Exclusion, a minor fluctuation in Φ can suddenly and completely obliterate consciousness in subsystems or supersystems. This creates the possibility of “flickering qualia” cases of the sort Chalmers (1996) discusses in arguing for a tight connection generally between consciousness and functional structure. See Favela 2019 for an attempt to reconcile IIT with the possibility of nested levels of consciousness.

consciousness. Furthermore, since “greater than” is a dichotomous property, there ought on Tononi’s view be an exact point at which polity-level integration crosses the relevant threshold, causing all human-level consciousness suddenly to vanish.⁵⁷ At this point, the addition of a single vote would cause every individual voter instantly to lose consciousness – even with no detectable behavioral or self-report effects, or any loss of integration, at the level of those individual voters. It’s odd to suppose that so much, and simultaneously so little, could depend on the discovery of a single mail-in ballot.

There’s a fundamental reason that it’s easy to generate unintuitive consequences from anti-nesting principles, either looking up and out or looking down and in or both. According to anti-nesting principles that look up and out, the consciousness of a system depends not exclusively on what’s going on internally within the system itself but also on facts about larger structures containing the system, structures potentially so large as to be beyond the smaller system’s knowledge. Changes in these larger structures could then potentially add or remove consciousness from the systems within them even with no internal changes whatsoever to the systems themselves. (This issue will arise again when I discuss nonlocality in Chapter 3.) According to anti-nesting principles that look down and in, the consciousness of a system can depend on structures within it that are potentially so small that they have no measurable impact on the system’s functioning and remain below its level of awareness. Such anti-nesting principles imply the possibility of unintuitive dissociations between what we would normally think of as organism-level indicators or constituents of consciousness (such as organism-level cognitive function, brain states, introspective reports, and behavior) and that organism’s actual

⁵⁷ See especially Tononi 2010, note 9.

consciousness, if those indicators or constituents happen to embed or be embedded in the wrong sort of much larger or much smaller things.

I conclude that neither the existing theoretical literature nor intuitive armchair reflection support commitment to a general anti-nesting principle.

2. Objection 2: The U.S. Has Insufficiently Fast and/or Synchronous Cognitive Processing.

Andy Clark has prominently argued that consciousness requires high-bandwidth neural synchrony – a type of synchrony that is not currently possible between the external environment and structures interior to the human brain. Thus, he says, consciousness stays in the head.⁵⁸ Now in the human case, and generally for Earthly animals with central nervous systems, maybe Clark is right – and maybe such Earthly animals are all he really has in view. However, we can consider elevating this principle to a necessity. The informational integration of the brain is arguably qualitatively different in its synchrony from the informational integration of the United States. If consciousness, in general, requires massive, swift, parallel information processing, then maybe mammals are conscious but the United States is not.

However, this move has a steep price if we are concerned, as the ambitious general theorist should be, about hypothetical cases. Suppose we were to discover that some people, though outwardly very similar to us, or some alien species, operated with swift but asynchronous serial processing rather than synchronous parallel processing. A fast serial system might be very difficult to distinguish from a slower parallel system, and models of parallel processes are often implemented in serial systems or systems with far fewer parallel streams.⁵⁹ Would we really be

⁵⁸ Clark 2009; with clarification and critique in Vold 2015.

⁵⁹ For example, massively “connectionist” architectures are often modeled on standard desktop computers. In philosophy of mind, the “Church-Turing Thesis” is sometimes interpreted

justified in thinking that such entities had no conscious experiences? Or what if we were to discover a species of long-lived, planet-sized aliens whose cognitive processes, though operating in parallel, proceeded much more slowly than ours, on the order of hours rather than milliseconds? If we adopt the liberal spirit that admits consciousness in Sirian supersquids and Antarean antheads – the most natural development of the materialist view, I’m inclined to think – it seems we can’t insist that high-bandwidth neural synchrony is necessary for consciousness. To justify a more conservative view on which consciousness requires a particular type of architecture, we need some principled motivation for denying the consciousness of any hypothetical entity that lacks that architecture, regardless of how similar that entity might be in its outward behavior. No such motivation suggests itself here.

Analogous considerations will likely trouble most other attempts to deny U.S. consciousness on broad architectural grounds of this sort.

3. Objection 3: The U.S. Is So Radically Structurally Different That Our Terms Are Infelicitous.

Daniel Dennett is arguably the most prominent living theorist of consciousness.⁶⁰ When I was initially drafting the arguments of this chapter, I emailed him, and he offered a pragmatic objection: To the extent that the United States is radically unlike human beings, it’s unhelpful to ascribe consciousness to it.⁶¹ Its behavior is impoverished compared to ours and its functional

as showing that any computable function could in principle be implemented by a sufficiently swift serial computer, but that characterization might not be accurate: See Copeland 1997/2020. See also Townsend and Wegner 2004 for discussion the difficulty of empirically distinguishing serial and parallel processes even in the human case.

⁶⁰ Dennett 1991, 1996, 2017.

⁶¹ The objections by Dennett, Chalmers, and Dretske are all shared with their permission, and each philosopher explicitly approved my formulation of their concerns before I published the article on which this chapter is based.

architecture is radically unlike our own. Ascribing consciousness to the United States is not as much straightforwardly false as it is misleading. It invites the reader to too closely assimilate human architecture and group architecture.

To this objection I respond, first, that the United States is not behaviorally impoverished. It does many things, as described in Sections 3 and 4 above – probably more than any individual human does. (In this way it differs from the aggregate of the U.S., Pakistan, and South Africa, and *maybe* also from the aggregate of all humanity.⁶²) Second, to hang the existence of consciousness too sensitively on details of architecture runs counter to the spirit that admits Sirians and Antareans to the realm of entities who would (hypothetically) be conscious. Thus, this objection faces concerns similar to those I raise for Objection 2. And third, we can presumably dodge such practical worries about over-assimilating groups and individuals by being restrained in our inferences. We can refrain from assuming, for example, that when the U.S. is angry its anger is felt in a way that is experientially similar to human anger. We can even insist that “anger” is not a great word and simply the best we can do with existing language. The U.S. can’t feel blood rush to its head; it can’t feel tension in its arms; it can’t “see red”. It can muster its armies, denounce the offender via spokespeople in Security Council meetings, and enforce an embargo. What it feels like, if anything, to enforce an embargo, defenders of U.S. consciousness can wisely refrain from claiming to know.⁶³

⁶² The entity composed of all of humanity, for example, does not appear to linguistically communicate with other entities of its type or enter into social interactions, and top-down, “self-conscious” organized control of behavior is more common at a national level than at the level of all of humanity. However, see Teilhard de Chardin 1955/1965 for an engaging perspective on species-level consciousness.

⁶³ One intriguing possibility, however, would be to draw on James C. Scott’s reflections in *Seeing Like a State* (1998), treating the “seeing” more literally than Scott probably intends. Scott argues that state-level management and control requires simplifying facts about the world

4. Objection 4: Most of the Cognitive Processing of the U.S. is Within Its Subsystems Rather Than Between Them.

I also tried these ideas on David Chalmers, perhaps the world's most prominent currently active scholarly critic of materialism. In response, Chalmers proposed (but did not endorse) the following objection: The United States might lack consciousness because the complex cognitive capacities of the United States arise largely in virtue of the complex cognitive capacities of the people composing it and only to a small extent in virtue of the functional relationships between the people composing it.⁶⁴ To feel the pull of Chalmers's suggestion, consider an extreme example – a two-seater homunculus, such as an Antarean controlled not by ten million insects but instead by two homunculi living in the mammoth's hump, in constant verbal communication.⁶⁵ Assuming that such a system's cognitive capacities arise almost entirely in virtue of the capacities of the two individual homunculi, while the interaction between the homunculi serves a secondary, coordinating role, one might plausibly deny consciousness to the system as a whole even while granting consciousness to systems whose processing is more distributed, such as Earthly rabbits and ten-million-insect antheads. Maybe the United States is like a two-seater homunculus?

Chalmers's objection seems to depend on something like the following principle: The complex cognitive capacities of a conscious system (or at least the cognitive capacities in virtue

and its citizens: The state might "see" a forest simply as a source of so much annual lumber or a citizen as simply a taxpayer with such-and-such income and property.

⁶⁴ Although Chalmers is not a materialist, for the issues at hand his view invites similar treatment. See Chalmers 1996 and Dahan 2017. List 2018 develops an objection similar to Chalmers' and relates it to Integrated Information Theory (see note XXX).

⁶⁵ Schwitzgebel 2019c, ch. 38, contains a more detailed example of a two-seater homunculus.

of which it is conscious) must arise largely in virtue of the functional relationships *between* the subsystems composing it rather than in virtue of the capacities *of* its subsystems. If such a principle is to defeat U.S. consciousness, it must be the case that both (a.) the United States has no such complex capacities that arise largely in virtue of the functional relationships between people, and (b.) no conscious system could be conscious largely in virtue of the capacities of its subsystems. Part (a) is difficult to assess, but given the boldness of the claim – there are exactly zero such capacities – it seems a risky bet unless we can find solid empirical grounds for it.

Part (b) is even bolder. Consider a rabbit’s ability to visually detect a snake. This complex cognitive capacity, presumably an important contributor to rabbit visual consciousness, might exist largely in virtue of the functional organization of the rabbit’s visual subsystems, with the results of processing then communicated to the organism as a whole, precipitating further reactions. Indeed, turning (b) almost on its head, some models of human consciousness treat subsystem-driven processing as the normal case: The bulk of our cognitive work is done by subsystems that cooperate by feeding their results into a “global workspace” or that compete for “fame” or control of the organism.⁶⁶ So grant (a) for the sake of argument: The relevant cognitive work of the United States is done largely within individual subsystems (people or groups of people) who then communicate their results across the U.S. as a whole, competing for fame and control via complex patterns of looping feedback. At the very abstract level of description relevant to Chalmers’s objection, such an organization might not be so different from the actual organization of the human mind. And it is of course much bolder to commit to the further view, per part (b), that no conscious system could *possibly* be organized in such a subsystem-driven way. It’s hard to see what could justify such a claim.

⁶⁶ For example, Baars 1988; Dennett 2005; Dehaene 2014.

The two-seater homunculus is strikingly different from the rabbit or ten-million-insect anthead because the communication is between only two subentities, at a low information rate. But the U.S. is composed of three hundred million subentities whose informational exchange is massive. The cases aren't sufficiently similar to justify transferring intuitions from the one to the other.

5. Objection 5: The Representations of the U.S. Depend on Others' Representations.

A fourth objection arose in email correspondence with philosopher Fred Dretske, whose 1995 book *Naturalizing the Mind* is an influential defense of the view that consciousness arises when a system has a certain type of sophisticated representational capacity, and also in a published exchange with François Kammerer, who was at the time a philosophy graduate student in Paris.⁶⁷ Dretske suggested that The United States cannot be conscious because its representational states depend on the conscious states of others. In his lingo, that renders its representations “conventional” rather than “natural”.⁶⁸ Kammerer argued, relatedly, that if the reason a system acts as if it is conscious is that it contains smaller entities within it who have conscious representations of that larger entity, then that larger entity is not in fact conscious.⁶⁹ Both Dretske and Kammerer resist attributing consciousness to the United States as a whole on the grounds that the group-level representations of the United States depend, in a particular

⁶⁷ See Kammerer 2015 and Schwitzgebel 2016b.

⁶⁸ In his 1995 book, Dretske says that a representation is natural if it is not “derived from the intentions and purposes of its designers, builders, and users” (p. 7) rather than the more general criterion of independency from “others”.

⁶⁹ The full specification of Kammerer's principle is more complex than this summary statement, but not I believe in a way that makes a difference to the present argument. In Schwitzgebel 2016a, I reply in more detail than is possible here.

consciousness-excluding way, on the individual level representations of the people who compose the U.S.

To see the appeal of Dretske's idea, consider a system that has some representational functions, but only extrinsically, because outside users impose representational functions upon it. That doesn't seem like a good candidate for a conscious entity. We don't make a mercury column conscious by calling it a thermometer, nor do we make a machine conscious by calling it a robot and interpreting its output as speech acts. The machine either is or is not conscious, it seems, independently of our intentions and labels. A wide range of theorists, I suspect, will and should accept that an entity cannot be conscious if all of its representational functions depend in this way on external agents.

But citizens and residents of the United States are parts of the U.S. rather than external agents, and it's not clear that dependency on the intentions and purposes of *internal* agents threatens consciousness in the same way, if the internal agents' behavior is properly integrated with the whole. The internal and external cases, at least, are sufficiently dissimilar that before accepting Dretske's principle in its general form we should at least consider some potential internal agent cases. The Antarean antheads are just such a case, and I've suggested that the most natural materialist position is to allow that they are conscious.

Kammerer explicitly focuses on internal agents, thus dodging my reply to Dretske. Also, Kammerer's principle denies consciousness only if the internal agents represent the larger entity as a whole, thus maybe permitting Antarean anthead consciousness while denying U.S. consciousness, if the ants are sufficiently ignorant. However, it's not clear whether Kammerer's approach can actually achieve that seemingly desirable result concerning the antheads. On a weak sense of "representing the whole", some ants plausibly could or would represent the whole

(e.g., representing the anthead's position in egocentric space, as some human neural subsystems might also do⁷⁰); while on more demanding senses of “representing the whole” (e.g., representing the U.S. as a concrete entity with group-level conscious states), individual members of the U.S. might not represent the whole. Furthermore, and more fundamentally, Kammerer's criterion appears largely to be motivated post-hoc to deliver the desired result, rather than having any independent theoretical appeal. *Why*, exactly, should a part's representing the whole cause the whole to lose consciousness? What would be the mechanism or metaphysics behind this?

The fundamental problem with both Dretske's and Kammerer's principles is similar to the fundamental problem with anti-nesting principles. Although such principles might exclude counterintuitive cases of group consciousness, they engender different counterintuitive consequences of their own, such as that you would lose consciousness upon inhaling a Planck-sized person with the right kind of knowledge or motivations. These counterintuitive consequences follow from the disconnection such principles introduce between function and behavior at the larger and smaller levels. On both Dretske's and Kammerer's principles, and I suspect on any similar principle, entities that behave similarly on a large scale and have superficially similar evolutionary and developmental histories might either have or lack consciousness depending on micro-level differences that are seemingly unreportable, unintrospectible, unrelated to what they say about The Beatles, and thus, it seems natural to suppose, irrelevant.

Dretske conceives of his criterion as dividing “natural” representations from “conventional” or artificial ones. Maybe it is reasonable to insist that a conscious entity have natural representations. But from a telescopic perspective national groups and their

⁷⁰ Holmes and Spence 2004; Ekstrom and Isham 2017.

representational activities are entirely natural – as natural as the structures and activities of groups of cells clustered into spatially contiguous individual organisms. What should matter on a broadly Dretskean approach, I'm inclined to think, is that the representational functions emerge naturally from within rather than being artificially imposed from outside, and that they are properly ascribed to the whole entity rather than only to a subpart. Both Antarean opinions about Shakespeare and the official U.S. position on North Korea's nuclear program appear to meet those criteria.

6. Objection 6: Methodological Concerns.

Riffling through existing theories of consciousness, we could try to find, or we could invent, some necessary condition for consciousness that human beings meet, that the United States fails to meet, and that sweeps in the most plausibly conscious non-human entities. As mentioned in the conclusion to Chapter 2, I wouldn't object to treating the argument as a challenge to which materialists might rise: Let's find, if we can, an independently plausible criterion that delivers this appealing conclusion! The suggestions above, if they can be adequately developed, might be a start. But it's not clear what if anything would justify taking the non-consciousness of the United States as a fixed point in such discussions. The confident rejection of otherwise plausible theories simply to avoid implications of U.S. consciousness would only be justified if we had excellent independent grounds for denying U.S. consciousness. It's hard to see what those grounds might be.

We have basically two choices of grounds for holding firm against U.S. consciousness: One possible ground would be good *theoretical* reason to deny consciousness to the U.S. However, I've argued that we have no such good theoretical reason and that in fact the bulk of

mainstream theories, interpreted at face value, instead seem to *support* the idea of U.S. consciousness.

The other possible reason to hold firm is just that it seems intuitively implausible, bizarre, and contrary to good old-fashioned common sense to think that the United States could literally be conscious. It's not unreasonable to resist philosophical conclusions on commonsense grounds. Chapter 3 will be all about this. However, resistance on these grounds should be accompanied by an acknowledgement that what initially seems bizarre can sometimes ultimately prove to be true. Scientific cosmology is pretty bizarre! Microphysics is bizarre. Higher mathematics is bizarre. The more we discover about the fundamentals of the world, the weirder things seem to become. Our sense of strangeness is often ill-tuned to reality. We ought at least remain open to the possibility of group consciousness if on continuing investigation our best theories point in that direction.

Some readers – perhaps especially empirically-oriented readers – might suggest that the argument does little more than display the bankruptcy of speculation about bizarre cases. How could we hope to build any serious theory on science-fictional intuitions? Perhaps we should abandon any aspiration for a truly universal theory that would cover the whole range of hypothetical entities. The project seems so ungrounded, so detached from our best sources of evidence about the world!

Part of me sympathizes with that reaction. There's something far-fetched and wildly speculative about this enterprise. Still, let me remind you: The United States is not a hypothetical entity. We can skip the Sirians and Antareans if we want and go straight to the question of whether the actual United States possesses the actual properties that actual consciousness scientists treat as indicative of consciousness. The hypothetical cases mainly

serve to clarify the implications and to block certain speculative countermoves. They can be skipped if you're allergic. Furthermore – and what really drives me – it's a fundamental assumption of this book that we should permit ourselves to wonder about and try to think through, as rigorously as we can and with an appropriately liberal helping of doubt, speculative big picture questions like whether group consciousness is possible, even if the science comes up short. It would just be too sad if the world had no space for speculation about wild hypotheticals.

The Weirdness of the World

Chapter Three Universal Bizarreness and Universal Dubiety

“The universe is not only queerer than we suppose, but queerer than we *can* suppose”
(Haldane 1927, p. 286).

Here are two things I figure you know:

1. You’re having some conscious experiences.
2. Something exists in addition to your conscious experiences (for example, some objects external to you).

Both facts are difficult to deny, especially if we don’t intend anything fancy – as I don’t – by “conscious experiences” or “objects”.⁷¹ You have experiences, and oh hey, there’s some stuff. It requires some creative, skeptical energy to doubt either of these headsmackingly obvious facts. As it happens, I myself have lots of creative, skeptical energy! Later, I’ll try doubting versions of these propositions. Let’s assume them for now.

But how are (1) and (2) related? How is mentality housed in the material or seemingly-material world? If you take mindless, material stuff and swirl it around, under what conditions will consciousness arise? Or, alternatively, could consciousness never arise merely from mindless, material stuff? Or is there, maybe, not even such a thing as mindless, material stuff?

⁷¹ See Chapter 8 for my un-fancy definition of “conscious experience”. See Chapters 5 and 6 for skeptical scenarios that are consistent with the existence of *something* besides oneself – even if it’s just an immaterial computational Angel or a chess-playing intelligence composed of who-knows-what.

About this, philosophers and scientists have accumulated centuries-worth of radically divergent theories.

In this chapter I will argue that every possible theory of the relationship of (1) and (2) is both bizarre and dubious. Something radically strange and counterintuitive must be true about the relationship between mind and world. And for the foreseeable future, we cannot know which among the various bizarre and dubious alternatives is in fact correct.

Part One: Universal Bizarreness

1. Why Is Metaphysics Always Bizarre?

Start your exploration of metaphysics as far away as you please from the “metaphysics” aisle of the bookstore. Inquire into the fundamental nature of things with as much sobriety and bland common sense as you can muster. You will always, if you arrive anywhere, arrive at weirdness. The world defies bland sobriety. Bizarreness pervades the cosmos.

Philosophers who explore foundational metaphysical questions typically begin with some highly plausible initial commitments or commonsense intuitions, some solid-seeming starting points – that there is a prime number between two and five, that they could have had eggs for breakfast, that squeezing a clay statue would destroy the statue but not the lump of clay. They think long and hard about what these claims imply. In the end, they find themselves positing a nonmaterial realm of abstract Platonic entities (where the numbers exist), or the real existence of an infinite number of possible worlds (in at least one of which a counterpart of them ate eggs for breakfast), or a huge population of spatiotemporally coincident objects on their mantelpiece

(some but not others of which would be destroyed by squeezing).⁷² In thirty-plus years of reading philosophy, I have yet to encounter a single broad-ranging exploration of the fundamental nature of things that doesn't ultimately entangle its author in seeming absurdities. Rejection of these seeming absurdities then becomes the commonsense starting point of a new round of metaphysics by other philosophers, generating a complementary bestiary of metaphysical strangeness. Thus are philosophers happily employed.

I see three possible explanations of why metaphysics, the study of the fundamental nature of things, is always bizarre.

First possible explanation. Without some bizarreness, a metaphysics won't sell. It would seem too obvious, maybe. Or it would lack theoretical panache. Or it would conflict too sharply with recent advances in empirical science.

The problem with this explanation is that there should be at least a small market for a thoroughly commonsensical metaphysics with no bizarre implications, even if that metaphysics is gauche, boring, and scientifically stale. Common sense might not be quite as fun as Nietzsche's eternal recurrence or Leibniz's windowless monads.⁷³ It won't be as scientifically current as the latest incarnation of multiverse theory.⁷⁴ But a commonsensical metaphysics ought to be attractive to a certain portion of philosophers. It ought at least be a curiosity worth study. It shouldn't be so downmarket as to be entirely invisible.

⁷² For example, see Frege 1884/1968 and 1918/1956 on Platonic entities (for interpretative caveats see Reck 2005); Lewis 1986 on possible worlds; and Yablo 1987 on the essential properties of statues and lumps of clay.

⁷³ On windowless monads see Leibniz 1714/1989. On eternal recurrence, see Nietzsche 1882/1974, §341, and 1901/1968. I favor the old-fashioned view that Nietzsche intended eternal recurrence literally as a cosmological hypothesis, as argued in Loeb 2013. For some related ideas, see Schwitzgebel 2019c, ch. 44.

⁷⁴ On multiverse theory: Carroll 2010; Green 2011; Tegmark 2014.

Second possible explanation. Metaphysics is difficult. A thoroughly commonsensical metaphysics with no bizarre implications is out there to be discovered. We simply haven't found it yet. If all goes well, someday someone will piece it together, top to bottom, with no serious violence to common sense anywhere in the system.

This is wishful thinking against the evidence. Philosophers have worked at it for centuries, failing again and again. Indeed, often the most thorough metaphysicians – Leibniz, David Lewis – are the ones whose systems seems the strangest.⁷⁵ It's not as though we've been making slow progress toward ever less bizarre views and only a few more pieces need to fall into place. There is no historical basis for hope that a well-developed commonsense metaphysics will eventually arrive.

Third possible explanation. Common sense is incoherent in matters of metaphysics. Contradictions thus inevitably flow from commonsense metaphysical reflections, and no coherent metaphysical system can adhere to every aspect. Although common sense serves us well in practical maneuvers through the social and physical world, common sense has proven an unreliable guide in theoretical physics, probability theory, neuroscience, macroeconomics, evolutionary biology, astronomy, medicine, topology, chemical engineering.... If, as it seems to, metaphysics more closely resembles those endeavors than it resembles reaching practical judgments about picking berries and planning parties, we might reasonably doubt the

⁷⁵ For a brief review of Leibniz on windowless monads and pre-established harmony, see Kulstad and Carlin 2020. I've already mentioned Lewis's realism about the existence of possible worlds, where counterparts of you exist whose behavior determines what counterfactual statements about you are true or false. (It's true that I could have had eggs for breakfast because a counterpart Eric Schwitzgebel in another possible world did have eggs for breakfast.) Lewis famously acknowledges that such modal realism typically draws, in response, an "incredulous stare" (1986, p. 133). Lewis's views about consciousness are also bizarre, including in ways that aren't widely appreciated among scholars of his work, as I will discuss in Section 3.3 below on nonlocality, and which I explore in a more fanciful manner in Schwitzgebel 2015e.

dependability of common sense as a guide to metaphysics.⁷⁶ Undependability doesn't imply incoherence, of course. But it seems a natural next step, and it would neatly explain the historical facts at hand.

On the first explanation, we could easily enough invent a thoroughly commonsensical metaphysics if we wanted one, but we don't want one. On the second explanation, we do want one, or enough of us do, we just haven't finished constructing it yet. On the third explanation, we can't have one. The third explanation best fits the historical evidence and best matches the pattern in related disciplines.

Let me clarify the scope of my claim. It concerns only broad-ranging explorations of fundamental metaphysical issues, especially issues around which seeming absurdities have historically congregated: mind and body, causation, identity, the catalog of entities that really exist. On these sorts of questions, thorough-minded philosophers have inevitably found themselves in thickets of competing bizarreness, forced to weigh violations of common sense against each other and against other types of philosophical consideration. It's not the aim of this section to defend this claim in detail, but I expect most seasoned metaphysicians will agree with this description of the state of the discipline (if not about the cause of that state); and the remainder of the chapter (more broadly, the book as a whole) will explore in more detail the bizarre metaphysics of the mind-body relation.

⁷⁶ Critiques of the role of common sense or philosophical intuition as a guide to metaphysics and philosophy of mind can be found in, for example, Churchland 1981; Stich 1983; Kornblith 1998; Dennett 2005; Ladyman and Ross 2007; Weinberg et al. 2010; Segal 2020; Frances 2021. I also count Hume 1740/1978 and Kant 1781/1787/1998 as allies on this issue. Even metaphysical views that treat metaphysics as largely a matter of building a rigorous structure out of commonsense judgments often envision conflicts with common sense so that the entirety of common sense can't be preserved: Ayer 1967; Kriegel 2011.

Common sense is culturally variable. So whose common sense do I mean? It doesn't matter! All well-elaborated metaphysical systems in the philosophical canon, Eastern and Western, ancient and modern, conflict both with the common sense of their milieu and with current Western Anglophone common sense. Kant's noumena, Plato's forms, Lewisian possible worlds, Buddhist systems of no-self and dependent origination – these were never part of any society's common sense.⁷⁷

Common sense changes. Heliocentrism used to defy common sense, but it no longer does. Maybe if we finally get our metaphysics straight and then teach it patiently to generations of students, sowing it deep into our culture, eventually people will say, “Ah yes, of course, modal realism and the transcendental aesthetic, what could be more plain and obvious to the common fellow!” One can always dream.

Some readers will doubt the existence of the phenomenon I aim to explain. They will think or suspect that there is a thoroughly commonsensical metaphysics already on the market. To them, I issue a challenge: Who might count as a thoroughly commonsensical metaphysician? Find me one. As far as I see, there are simply none to be found.

I've sometimes heard it suggested that Aristotle was a common sense metaphysician. Or Scottish “common sense” philosopher Thomas Reid. Or G. E. Moore, famous for his “Defence of Common Sense”. Or “ordinary language” philosopher P. F. Strawson. Or the later Wittgenstein. But Aristotle didn't envision himself as defending a commonsense view. In the introduction to the *Metaphysics* Aristotle says that the conclusions of sophisticated inquiries such as his own will often seem “wonderful” to the untutored and contrary to their opinions; and he

⁷⁷ On Kant's noumena, see discussion in Chapter 5. On Plato's forms, see Silverman 2003/2014. On Lewisian possible worlds, see note 5 above. On Buddhist metaphysics and philosophy of mind, see Siderits 2007 and Garfield 2015.

sees himself as aiming to distinguish the true from the false in common opinion.⁷⁸ Moore, though fierce in wielding common sense against his philosophical foes, is unable to preserve all commonsense opinions when he develops his positive views in detail, for example in his waffling about the metaphysics of “sense data” (entities you supposedly see both when hallucinating and in ordinary perception).⁷⁹ Strawson, despite his preference for ordinary ways of talking and thinking insofar as possible, struggles similarly, especially in his 1985 book, in which he can find no satisfactory account of mental causation.⁸⁰ Wittgenstein does not commit to a detailed metaphysical system.⁸¹ Reid I will discuss in Section 4 below.

Since I can’t be expert in the entire history of global philosophy, maybe there’s someone I’ve overlooked – a thorough and broad-ranging metaphysician who nowhere violates the common sense at least of their own culture. I welcome the careful examination of possible counterexamples.

My argument is an empirical explanatory or “abductive” one. The empirical fact to be explained is that across the entire history of philosophy, all well-developed metaphysical systems – all broad-ranging attempts to articulate the general structure of reality, including especially attempts to explain how our mental lives relate to the existence of objects around us – defy common sense. Every one of them is in some respect jaw-droppingly bizarre. An attractive possible explanation of this striking empirical fact is that people’s commonsensical metaphysical intuitions form an incoherent set.

⁷⁸ Aristotle 4th c. BCE/1928, 983a; θαυμάσιον: “wonderful” in the sense of tending to cause wonder, or amazing.

⁷⁹ See Moore 1925 on common sense and Moore 1922, 1953, and 1957 on sense data.

⁸⁰ Strawson 1985. Also noteworthy, given Chapter 2 above: In his influential 1959 book, Strawson appears to endorse or at least take seriously the possibility of group consciousness.

⁸¹ This fits with Wittgenstein’s quietism, which I briefly criticized in Chapter 2 above.

2. *Bizarre, Dubious, and Wild.*

Let's call a position *bizarre* if it's contrary to common sense. And let's say that a position is contrary to common sense if most people without specialized training on the issue confidently, but perhaps implicitly, believe it to be false. This usage traces back to Cicero, who calls principles that ordinary people assume to be obvious *sensus communis*. Cicero describes violation of common sense as the orator's "worst possible fault" – presumably because the orator's views will be regarded as ridiculous by his intended audience.⁸² Common sense is what it seems ridiculous to deny.

To call a position *bizarre* is not necessarily to repudiate it. The seemingly ridiculous is sometimes true. Einsteinian relativity theory is *bizarre* (e.g., the Twin Paradox). Various *bizarre* things are true about the infinite (e.g., the one-to-one correspondence between the points in a line segment and the points in the whole of a space that contains that line segment as a part, and thus their equal cardinality or size). Common sense errs, and we can be justified in thinking so.

However, we are not ordinarily justified in believing *bizarre* things without compelling evidence. In the matters it traverses, common sense serves as an epistemic starting point that we reject only with sufficient warrant. To believe something *bizarre* on thin grounds – for example, to think that the world was created five minutes ago or that you are constantly cycling through different immaterial souls – seems weird or wacky or wild. I stipulate, then, the following technical definition of a *wild* position: A position is *wild* if it's *bizarre* and we are not epistemically compelled to believe it.⁸³ Theoretical wildness is then type of weirdness in the

⁸² Cicero 1st c. BCE/2001: *De Oratore* I.ii.12.

⁸³ In Schwitzgebel 2014b, I used the term "crazy" instead of "wild" for this idea, and "crazyism" for the conjunction of Universal Bizarreness and University Dubiety. However,

sense of Chapter 1. It's weird – strikingly contrary to the normal or ordinary – to hold a wild theory.⁸⁴

Not all wild positions are as far out as the two just mentioned. Many philosophers and some scientists favor views contrary to common sense and for which the evidence is less than compelling. In fact, to convert the wild to the merely bizarre might be the highest form of academic success. Einstein, Darwin, and Copernicus (or maybe Kepler) all managed the conversion – and at least in the case of Copernicus common sense eventually relented. Intellectual risk-takers nurture wild views and see what marvels bloom. The culture of contemporary Anglophone academia, perhaps especially philosophy, overproduces wildness like a plant produces seeds.

Let's call a topic or general phenomenon a *theoretical wilderness* if something wild must be among the core truths about that topic or general phenomenon. Sometimes we can be justified in believing that one among several bizarre views must be true about the matter at hand, while the balance of evidence leaves no individual view decisively supported over all others. We might find ourselves rationally compelled to believe that either Theory 1, Theory 2, Theory 3, or Theory 4 must be true, where each of these theories is bizarre and none rationally compels belief.

Consider interpretations of quantum mechanics. The “many worlds” interpretation, for example, according to which reality consists of vastly many worlds constantly splitting off from each other, seems to conflict sharply with common sense.⁸⁵ It also seems that the balance of

given the perhaps too negative connotations of “crazy” and its ableist history, I downplay the term in this chapter.

⁸⁴ Okay, one exception: If for some reason it becomes normal or ordinary to hold a bizarre and dubious theory, then holding that the theory is no longer weird. Maybe the Catholic doctrine of the trinity is an example.

⁸⁵ Theoretical physicist Bryce DeWitt, for example, writes:

evidence does not decisively favor this view over all competitors.⁸⁶ Thus, the many worlds interpretation is “wild” in the sense defined. If the same holds for all viable interpretations of quantum mechanics, then quantum mechanics is a theoretical wilderness.

The thesis of this chapter is that the metaphysics of mind is a wilderness. *Something* bizarre must be true, but we can’t know – at least not yet, not with our current tools – which among the various bizarre possibilities is correct. I will divide approaches to the mind-world relation into four broad types: materialist (according to which everything is material), dualist (according to which both material and immaterial substances exist), idealist (according to which only immaterial substances exist), and a grab-bag of views that reject all three alternatives or attempt to reconcile or compromise among them. Part One defends Universal Bizarreness: Each of these four approaches is bizarre, at least when details are developed and implications pursued. Part Two defends Universal Dubiety: None of the four approaches compels belief.

Furthermore, as the example of idealism especially illustrates, bizarreness and dubiety concerning the metaphysics of mind entails bizarreness and dubiety about the very structure of the cosmos. If my arguments in this chapter are correct, we live in a bizarre cosmos that defies our understanding.

3. *Materialist Bizarreness.*

Recall that in Chapter 2 I characterized materialism as the view that everything in the universe is composed of, or reducible to, or most fundamentally, material stuff, where “material

I still recall vividly the shock I experienced on first encountering this multiworld concept. The idea of 10^{100+} slightly imperfect copies of oneself all constantly splitting into further copies, which ultimately become unrecognizable, is not easy to reconcile with common sense (1970, p. 33).

⁸⁶ See Maudlin 2019 for a review of the main competing interpretations.

stuff” means things like elements of the periodic table and the various particles or waves or fields that interact with or combine to form such elements, whatever those particles, waves, or fields might be, as long as they are not intrinsically mental. The two historically most important competitor positions are idealism and substance dualism, both of which assert the existence of immaterial souls.

Materialism *per se* might be bizarre. People have a widespread, and maybe developmentally and cross-culturally deep, tendency to believe that they are more than just material stuff.⁸⁷ Not all unpopular views violate common sense, however. It depends on how strongly the popular view is held, and that’s difficult to assess in this case.

My claim is not that materialism *per se* is bizarre. Rather, my claim is that *any well-developed materialist view* – any materialist view developed in sufficient detail to commit on specific metaphysical issues about, for example, the nature of mental causation and the distribution of consciousness among actual and hypothetical systems – will inevitably be bizarre.

⁸⁷ In the 2018 *General Social Survey* of U.S. residents (<http://www3.norc.org/GSS+Website>), 76% of respondents reported believing in life after death and 21% reported disbelieving; also, 75% reported believing in God, 13% reported belief in an impersonal “higher power”, and only 5% reported atheism. We can probably safely assume that most of theists and afterlife believers are not materialists (though some are materialists, e.g., Baker 1995; Murphy 2006). Some other industrialized Western nations are more secular than the U.S., but even in those societies religiosity is widespread (Zuckerman 2007; *World Values Survey*, 2017-2020, <https://www.worldvaluessurvey.org/wvs.jsp>), and religiosity or belief in entities not tolerated by materialism might even be something like a cultural universal (Brown 1991; McCauley 2000; Boyer 2001; De Cruz and De Smedt 2017). Paul Bloom (2004) has argued on developmental and cross-cultural grounds that it is innately natural to human beings to think of mental life as the product of an immaterial soul, even if some of us reject dualism on an “airy intellectual level” (for discussion, see Bering 2006; Richert and Harris 2006; Hodge 2008; Slingerland and Chudek 2011; Lane 2021). In David Bourget’s and David Chalmers’ 2009 PhilPapers survey of faculty in leading Anglophone philosophy departments, 73% of “target faculty” respondents reported accepting or leaning toward atheism. Yet even in this remarkably secular group, only 57% reported accepting or leaning toward physicalism (Bourget and Chalmers 2014).

I offer four considerations in support of this claim.

3.1. Antecedent plausibility.

As I noted above, traditional commonsense folk opinion has proven radically wrong about central issues in physics, biology, and cosmology. On broad inductive grounds, scientifically inspired materialists ought to be entirely unsurprised if the best materialist metaphysics of mind sharply conflicts with ordinary commonsense opinion. Folk opinion about mentality has likely been shaped by evolutionary and social pressures ill-tuned to the metaphysical truth, especially concerning cases beyond the run of ordinary daily life, such as pathological, science fictional, and phylogenetically remote cases. Materialists have reason to suspect serious shortcomings in non-specialists' ideas about the nature of mind.

3.2. Leibniz's mill and Crick's oscillations.

As I mentioned near the end of Chapter 2, in 1714 Leibniz asked his readers to imagine entering into a thinking machine as though one were entering into a mill. Looking around, he said, you will see only parts pushing each other – nothing to explain perceptual experience.⁸⁸ In the 20th century, Frank Jackson's "Mary" and the "zombies" of Robert Kirk and David Chalmers invite a similar thought. Jackson's Mary is a genius super-scientist who knows every physical fact about color but who lives in an entirely black and white room and has never actually seen any colors. It seems that she might know every physical fact about color (wavelength, neurophysiological reactions to light, linguistic and behavioral facts about how people describe and react to colors) while remaining ignorant of some further fact that is not among those

⁸⁸ Leibniz 1714/1898, §7, p. 215.

physical facts: what it's like to see red.⁸⁹ Kirk's and Chalmers' zombies are hypothetical entities physically identical to normal human beings but entirely lacking conscious experience. If we can coherently conceive of such creatures, Kirk and Chalmers argue, then there must be some property we have that zombies if they existed would lack, which by stipulation couldn't be a physical property.⁹⁰ I won't attempt to evaluate the soundness of such arguments here, but most of us can feel at least the initial pull. Leibniz, Jackson, Kirk, and Chalmers all appeal to something in us that rebels against the materialist idea that particular motions and configurations of material stuff could ever give rise to, fully explain, or constitute full-color conscious experience without the addition of something extra.

The more specific the materialist's commitments, the stiffer the resistance seems to be. Francis Crick, for example, equates human consciousness with 40-hertz oscillations in the subset of neurons corresponding to an attended object.⁹¹ Nicholas Humphrey equates consciousness with re-entrant feedback loops in an animal's sensory system.⁹² Both Humphrey and Crick report that popular audiences vigorously resist their views. Common sense fights them hard. Their views are more than just slightly unintuitive. When a theorist commits to a particular materialist account – *this* material process is what consciousness is! – the bizarreness of materialist theorizing about the mind shines through.

⁸⁹ Jackson 1986 is the original source. See Nida-Rümelin and Conaill 2002/2019 for a review of subsequent discussion and Alter and Walter 2007 for a collection of essays on the topic. Jackson himself later rejects the Mary argument in Jackson 1998.

⁹⁰ Kirk 1974 is the original source, with discussion exploding after Chalmers 1996 gave the thought experiment a prominent place in its influential critique of materialism. Kirk 2005 reverses position, rejecting the conceivability of zombies.

⁹¹ Crick 1994.

⁹² Humphrey 1992, 2011.

3.3. Mad pain, Martian pain, and nonlocality.

On several issues, materialist theories must, I think, commit to one among several alternatives each one of which violates common sense. One such issue concerns the implications of materialist theories of consciousness for cases that are functionally and physiologically remote from the human case. What is it, for example, to be in pain? Materialists generally take one of three approaches: a brain-based approach, a causal/functional approach, or a hybrid approach. Each approach has bizarre consequences.

This section will read more easily if you are familiar with David Lewis' classic – and brief! and fun! – “Mad Pain and Martian Pain”.⁹³ If you don't already know that essay, my feelings won't be hurt if you just go read it right now. Alternatively, muddle through my condensed presentation. Yet more alternatively, skip to §3.4.

You might think that to be in pain is just to be in a certain brain state. The simplest version of this view is associated with J. J. C. Smart.⁹⁴ To be in pain is just to have such-and-such a neural configuration discoverable by neuroscience, for example, X-type activity in brain region Y. Unfortunately, the simplest form of this view has the following bizarre consequence: No creature without a brain very much like ours could ever feel pain. For this reason, simple brain-state theories of consciousness have few adherents among philosophers interested in general accounts of the metaphysics of consciousness applicable across species. Hilary Putnam vividly expresses the standard objection:

Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that any organism (not just a mammal) is

⁹³ Lewis 1980.

⁹⁴ Smart 1959. In later work, Smart treats Lewis's more complex view as a natural development of his 1959 view: Smart 2000/2017.

in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusca, and certainly feel pain), etc. At the same time, it must not be a possible (physically possible) state of the brain of any physically possible creature that cannot feel pain. Even if such a state can be found, it must be nomologically certain that it will also be a state of the brain of any extra-terrestrial life that may be found that will be capable of feeling pain before we can even entertain the supposition that it may be pain.⁹⁵

To avoid this mess, one can shift to a causal or functional view: To experience pain is to be in a state that plays the right type of causal or functional role.⁹⁶ According to *simple functionalism*, to experience pain is to be in whatever state tends to be caused by tissue damage or tissue stress and in turn tends to cause withdrawal, avoidance, and protection (the exact details are tweakable). Any entity in such a state feels pain, regardless of whether its underlying architecture is a human brain, an octopus brain, a hypothetical Martian's cognitive hydraulics, or the silicon chips of a futuristic robot. (I briefly discussed Putnam's version of simple functionalism in the appendix to Chapter 2.)

However, bizarre consequences also follow from simple functionalism. First, as emphasized especially by Ned Block and John Searle, if mentality is all about occupying states that play these types of causal roles and the underlying architecture doesn't matter at all, you

⁹⁵ Putnam 1967, p. 44. But see Bechtel and Mundale 1999 and Shapiro and Polger 2016 for critique of this argument.

⁹⁶ Putnam 1967, 1975.

could get genuine pain experiences in some truly strange functional architectures – such as a large group of people communicating by radio or a huge network of beer cans and wire powered by windmills and controlling a marionette.⁹⁷ Such systems could presumably, at least in principle, be arranged into arbitrarily complex functional patterns, generating internal state transitions and outward behavior indistinguishable (except in speed) from that of commonsensically conscious entities like dogs, babies, or even adult humans. If you arranged a vast network of beer cans just right, it could presumably enter states caused by structural damage and stress and that in turn cause withdrawal, protection, and avoidance.

Second, if pain is all about being in a state that plays a causal role of this sort, then no one could feel pain without being in a state that is playing that causal role. But it seems to be simple common sense that one person might enter an experiential state under one set of conditions and tend to react to it in one way while someone else might enter that same experiential state under very different conditions and react to it very differently. In his critique of simple functionalism, Lewis imagines a “madman” whose pain experiences are caused in unusual ways (by moderate exercise on an empty stomach) and in turn cause unusual reactions (finger snapping and concentration on mathematics). Madness, pathology, disability, drugs, brain stimulation, personality quirks, and maybe radical existential freedom can amply scramble, it seems, the actual and counterfactual causes and effects of any experiential state. So while the brain-oriented view seems to be neural chauvinism that wrongly excludes creatures different from us in their underlying structure, insisting that pain is all about causal role appears to chauvinistically exclude people and creatures with atypical causes of or reactions to pain.

⁹⁷ Block 1978/2007; Searle 1980, 1984. See also Chapter 2, section 5.2, on neurochauvinism.

Lewis aimed to thread the needle by hybridizing. According to Lewis, entities who experience pain are entities in states with any one of a variety of possible biological or physical configurations – just those configurations that *normally* play the causal role of pain, even if things sometimes aren't quite normal. The madman feels pain, according to Lewis, because he's in Brain State X, which is the brain state normally caused by tissue damage in humans and which normally causes withdrawal and avoidance, even though in his particular case the causes and effects are atypical.

Normality can be understood in various ways. Maybe what matters is that you're in the biological or physical configuration that plays the causal role of pain in your species, even if it isn't playing that causal role for you right now.⁹⁸ Or maybe what matters is that the configuration played that role, or was selected to play that role, in your developmental or evolutionary history.⁹⁹ Either way, you're in pain whenever you're in whatever state normally plays the causal role of pain for you or your group.

What's central to all such appeals to normality or biological history is that pains no longer “supervene” locally: Whether you're in pain now depends on how your current biophysical configuration is seated in the broader universe. It depends on who else is in your group or on events in the past. In other words, on such views there's a *nonlocality* in the metaphysical grounds of pain experience. If normality depends upon the past, then you and I might be molecule-for-molecule identical with each other right now, screaming and writhing equally, equally cursing our tormentor, and utterly indistinguishable by any measure of our current brain structure or brain activity no matter how precise – and yet because of differences in

⁹⁸ Lewis 1980.

⁹⁹ Dretske 1995; Tye 1995, 2009.

our personal or evolutionary history, you're in pain and I'm not.¹⁰⁰ This would seem to be action at a historical distance. If pain depends, instead, on what is currently normal for your species or group, it has an even more bizarre implication, rarely appreciated (and not explicitly discussed by Lewis): Your pain state could change with selective genocide or with a speciation event beyond your ken.¹⁰¹ Strange forms of anaesthesia! On such a current normality view, to eliminate his pain, a tyrant in painful Brain State X could kill everyone for whom Brain State X plays the causal role of pain, while selectively preserving a minority for whom Brain State X plays some other causal role, such as the role characteristic of feeling a tickle. With no interior change, the tyrant's pain will transform into a tickle. While nonlocality is commonsensical for relational properties – whether I have a niece, whether I'm remembering coffee cup A or qualitatively identical coffee cup B – it is bizarre for pain.

Furthermore, any view of pain which doesn't confine the "supervenience base" to the organism's past threatens to permit faster-than-light communication (an implication that Lewis and most other materialists would probably prefer to avoid). To see this, first consider the relational property of having a niece. If my wife's sister emigrates to a distant star system, I become an uncle the instant her baby is born, even though it may take me years to hear the news (though of course when exactly that "instant" occurs will depend on the frame of reference being used). No possibility of faster-than-light communication arises, since I cannot detect, in that same instant, my sudden unclehood. Similarly, if whether I'm in pain depends on what's currently normal for my species, and 90% of my species is on a faraway planet, then changes in

¹⁰⁰ The standard thought experiment here is the "Swampman" case (alternatively "Swampcow") of an entity molecule-for-molecule identical to a human being but who congealed by freak quantum chance. Davidson 1987; Dretske 1995; Millikan 2010; Neander 2016.

¹⁰¹ In Schwitzgebel 2015e, I explore such a hypothetical scenario in detail with reference to David Lewis' view in particular.

that distant population (such as the sudden death of enough people that Y is now what's normal in that population rather than X) would seem also, instantly, to change whether I'm in pain. But on the assumption that I can detect whether I am in pain without waiting to hear the news, then I could know faster than light that the population has undergone a change. In this way, denying that pain depends exclusively on the organism's past implies either the possibility faster-than-light communication, both bizarre in itself and contrary to orthodox relativity theory, or the possibility of radical self-ignorance about whether one is in pain.

The issue appears to present a trilemma for the materialist. Either accept neural chauvinism (no Martian pain and probably no octopus pain), accept simple functionalism (beer can pain and no mad pain), or deny locality (action at a distance and/or anaesthesia by genocide). Maybe some materialist view can evade all three horns. They don't seem logically exhaustive. But if so, I don't see the view out there yet.

The argument of this subsection doesn't require the inescapability of the trilemma. My only claim is this: The issues raised here are such that any relatively specific materialist theory that attempts to address them will crash against common sense somewhere. It will have to "bite the bullet" somewhere, accepting some bizarre implications. Exactly which bizarre implications will depend on the details of the theory. All of the views on which materialists have so far alighted have bizarre implications. It is, I think, a reasonable guess that no plausible, well-developed materialist view can simultaneously respect all of our commonsense judgments about this cluster of issues.¹⁰²

¹⁰² See also Adams and Dietrich 2004. See Hill 2009 for a rather different argument that the folk understanding of pain is incoherent.

3.4. Group consciousness.

In Chapter 2, I argued that if materialism is true, the United States is probably conscious, that is, literally possessed of a stream of conscious experience over and above the experiences of its citizens and residents. If we look in broad strokes at the types of properties that materialists tend to regard as indicative of the presence of conscious experience – complex information processing, rich functional roles in a historically embedded system, sophisticated environmental responsiveness, complex layers of self-monitoring – the United States, conceived of as a concrete, spatially distributed entity with people as parts, appears to meet the criteria for consciousness. I expect you'll agree that it's contrary to common sense to suppose that the United States is literally conscious in this way.

In Chapter 2, I described several possible ways the materialist could dodge this bizarre conclusion. However, all the dodges that weren't mere hopeful handwaving (“future science will straighten everything out”) violated common sense in other respects. This is evidence that any serious attempt to negotiate the issue will likely violate our commonsense picture somewhere.

The literal group consciousness of the United States, anaesthesia by genocide, beer-can pain – these are the types of striking bizarreness that I have in mind. Mainstream materialism might not seem bizarre at a first pass – but that is because it's often presented in a sketchy way that masks the bizarre implications of the theoretical choices that are swiftly forced upon the thoughtful materialist.

Even if I have failed in my specific examples, I hope that the general point is plausible. The more we learn about cosmology, microphysics, mathematics, and other such foundational matters, the grosser the violations of common sense seem to become. The materialist should expect no less strangeness from the metaphysics of mind.

4. *Dualist Bizarreness.*

One alternative to materialism is dualism, the view that people have both material bodies and immaterial souls. (By “dualism”, unqualified, I mean what philosophers usually call “substance dualism”. “Property dualism” I will discuss briefly in Section 6.¹⁰³) Common sense might be broadly dualist.¹⁰⁴ However, from the 17th century to the present, the greatest philosophers of the Western world have universally found themselves forced into bizarreness when attempting to articulate the metaphysics of immateriality. This history is significant empirical evidence that a well-developed metaphysics of substance dualism will unavoidably be bizarre.

Attempts at commonsense dualism founder on two broad issues: the causal powers of the immaterial mind and the class of beings with immaterial minds.

The causal powers issue can be posed as a dilemma: Does the immaterial soul have the causal power to affect material entities like the brain? If yes, then material entities like neurons must be regularly and systematically influenced by immaterial events. A neuron must be caused to fire not just because of the chemical, electrical, and other influences on it but also because of immaterial happenings in spiritual substances. That forces a subsidiary choice. Maybe events in the immaterial realm transfer some physical or quasi-physical push that makes neurons behave other than they would without that push. But that runs contrary to both ordinary ideas and mainstream scientific ideas about the sorts of events that can alter the behavior of small,

¹⁰³ I intend “immaterial soul” in a fairly broad but traditional sense. Some metaphysical systems that call themselves substance dualist, notably Lowe’s (2008), are probably closer to compromise/rejection views in my taxonomy.

¹⁰⁴ See discussion in note 16 above.

material, mechanistic-seeming things like the subparts of neurons. Alternatively, maybe the immaterial is somehow causally efficacious with no push and no capacity to make a material difference. The immaterial has causal powers even though material events transpire exactly as they would have done without those causal powers. That seems at least as strange a view. Consider, then, the other horn of the dilemma: The immaterial soul has no causal influence on material events. If immaterial souls do anything, they engage in rational reflection. On a no-influence view, such rational reflection could not influence the movements of the body. You can't make a rational decision that has any effect on the physical world. I've rolled quickly here over some complex issues, but I think that informed readers of the history of dualism will agree that dualists have perennially struggled to accommodate the full range of commonsense opinion on mental-physical causation, for approximately the reasons I've just outlined.¹⁰⁵

The scope of mentality issue can be expressed as a quadrilemma. Horn 1: Only human beings have immaterial souls. Only we have afterlives. Only we have religious salvation. (Substance dualists needn't be theistic, but many are.) There's a cleanliness to this idea. But if the soul is the locus of conscious experience, as is standardly assumed, then this view implies that dogs are mere machines with no consciousness. No dog ever feels pain. No dog ever has any sensory experiences. There's nothing it's like to be a dog, just as there's nothing it's like (most people assume) to be a stone or a toy robot. That seems bizarre. Horn 2: Everybody and everything is in: humans, dogs, frogs, worms, viruses, carbon chains, lone hydrogen atoms sailing past the edge of the galaxy – we're all conscious! That view seems bizarre too. Horn 3: There's a line in the sand. There's a sharp demarcation somewhere between entities with souls and conscious experiences and those without. But that's also a bizarre view. Across the range of

¹⁰⁵ Heil 1998/2020; Patterson 2005.

animals, how could there be a sharp line between the ensouled and the unensouled creatures? What, toads in, frogs out? Grasshoppers in, crickets out? If the immaterial soul is the locus of conscious experience, it ought to do some work. There ought to be big psychological differences between animals with and without souls. But the only remotely plausible place to draw a sharp line is between human beings and all the rest – and that puts us back on Horn 1. Horn 4: Maybe we don't have to draw a sharp line. Maybe having a soul is not an on-or-off thing. Maybe there's a smooth gradation of ensoulment so that some animals – snails? – kind-of-have immaterial souls? (Or have kind-of-immaterial souls?) But that's bizarre too. What would it mean to kind of have or halfway have an immaterial soul? Immateriality doesn't seem like a vague property. It's seems quite different from being red or bald or tall, of which there are gradations and in-between cases.¹⁰⁶

I don't intend the causal dilemma and scope-of-mentality quadrilemma as *a priori* metaphysical arguments against dualism. Rather, I propose them as a diagnosis of an empirically observed phenomenon: the failure of Descartes, Malebranche, Leibniz, Bayle, Berkeley, Reid, Kant, Hegel, Schopenhauer, etc., up to and including 21st century nonmaterialists like David Chalmers, Philip Goff, Howard Robinson, and William Robinson, to develop non-bizarre views of the metaphysics of immateriality.¹⁰⁷ Some of these philosophers are better described as idealists or what I will call “compromise/rejection” theorists than substance dualists, but the quagmire they faced was the same and my explanation of their bizarre metaphysics is the same:

¹⁰⁶ The same quadrilemma arises if immateriality is regarded as essential to life, as on the types of vitalist theories that were discarded in the early 20th century and on immaterialist views of the “vegetative soul”. Nor would successful resolution of the vitalist quadrilemma resolve the core question about mentality, as emphasized by Pierre Bayle (1697/1702/1965, “Rorarius”; see also Des Chene 2006).

¹⁰⁷ Chalmers 1996; H. Robinson 2016; Goff 2017; W. Robinson 2019.

Anyone developing a metaphysics of immateriality unavoidably faces theoretical choices about mental causation and the scope of mentality. Immaterial souls either have causal powers or they do not. Either only humans have souls, or everything has a soul, or there's a bright line between the ensouled and unensouled animals, or there's a hazy line. Good old common sense reasoning can recognize that these are the options. But then, incoherently, common sense also rejects each option considered individually. Consequently, no coherent commonsense metaphysics of immateriality exists to be articulated.¹⁰⁸

You might suspect that some philosopher somewhere has developed a substance dualist metaphysics that violates common sense in no important respect. Of course I can't treat every philosopher on a case-by-case basis, but let me briefly discuss two: Thomas Reid, who enjoys a reputation as a "common sense philosopher" and René Descartes, whose interactionist substance dualism has perhaps the best initial intuitive appeal.

Reid's explicit and philosophically motivated commitment to common sense often leads him to refrain from advancing detailed metaphysical views – which is of course no harm to the Universal Bizarreness thesis. However, in keeping with that thesis, on those occasions where Reid does develop views on the metaphysics of dualism, he drops his commitment to common sense. On the scope of mentality, Reid is either silent or embraces a radically abundant view: He attributes immaterial souls to vegetables, but it's unclear whether he thinks possession of an immaterial soul is sufficient for consciousness. If it is, then grasses and cucumber plants have

¹⁰⁸ Actually, materialists face similar theoretical choices concerning mental causation and the scope of consciousness. However, without the seemingly bright metaphysical line between materiality and immateriality, they have more avenues of escape. On mental causation, this is sometimes known as the "exclusion problem". How can high-level mental events and low-level neural events both cause you to raise your arm without some causal overdetermination? (See Kim 1998.) On the scope of consciousness, see my discussion of the "slippery slope" argument for the abundance of consciousness in Chapter 10, §3.1, and Schwitzgebel 2022.

conscious experiences. If not, then Reid did not develop a criterion of non-human consciousness and so his theory is not “well developed” in the relevant sense.¹⁰⁹ On causal powers, Reid regards material events as causally inert or epiphenomenal. Only immaterial entities have genuine causal power. Material objects cannot produce motion or change, or even cohere into stable shapes, without the regular intervention of immaterial entities.¹¹⁰ Reid recognizes that this view conflicts with the commonsense opinions of ordinary people – though he says that this mistake of “the vulgar” does them no harm. Despite his general commitment to common sense, Reid explicitly acknowledges that on some issues human understanding is weak and common sense errs.¹¹¹

Descartes advocates an interactionist approach to the causal powers of the soul, according to which activities of the soul can exert a causal influence the brain, changing what it would otherwise do. Although this view is probably somewhat less jarring to common sense than other options, it does suggest an odd and seemingly unscientific view of the behavior of neurons, and it requires contortions to explain how the rational, non-embodied processes of the immaterial soul can be distorted by drugs, alcohol, and sleepiness.¹¹² On Descartes’ view, non-human animals, despite their similarity to human beings in physiology and much of their behavior, have no more thought or sensory experience than a cleverly made automaton.¹¹³ Some of Descartes’ later opponents imagined Descartes flinging a cat from a second-story window while asserting that animals are mere machines – testament to the sharp division between Descartes’ and the common person’s view about the consciousness of cats. The alleged defenestration was, or was

¹⁰⁹ Reid 1774-1778/1995, 3.X.

¹¹⁰ Reid 1774-1778/1995, 1788/2010.

¹¹¹ Reid 1788/2010, IV.3, 1785/2002, I.1.

¹¹² La Mettrie 1748/2002 is especially vivid on this point, contra Descartes 1649/1985.

¹¹³ Descartes 1649/1991.

intended to be, the very picture of the bizarreness of Cartesian metaphysics.¹¹⁴ Descartes' interactionist dualism is no monument of common sense.

I conclude that we have good grounds to believe that any well-developed dualist metaphysics of mind will somewhere conflict sharply with common sense.

5. Idealist Bizarreness.

A third historically important position is idealism, the view that there is no material world at all but instead only a world of minds or spirits in interaction with each other or with God, or wholly solipsistic. In the Western tradition, Berkeley, Fichte, Schelling, and maybe Hegel are important advocates of this view, and in the non-Western tradition the Indian Advaita Vedānta and Yogācāra traditions, or some strands of them, may be idealist in this sense.¹¹⁵ As Berkeley acknowledges, idealism is not the ordinary view of non-philosophers: "It is indeed an opinion strangely prevailing amongst men that houses, mountains, rivers, and, in a word, all sensible objects have an existence, natural or real, distinct from their being perceived by the understanding."¹¹⁶ No one, it seems, is born an idealist. They are convinced, against common sense, by metaphysical arguments or by an unusual meditative or religious experience.

Idealism also inherits the bizarre choices about causation and scope of mentality that trouble dualism. If a tree falls, is this somehow one idea causing another, in however few or many minds happen to observe it? Do non-human animals exist only as ideas in our minds or do

¹¹⁴ Grayling 2005, p. 135, offers a recent account of this apocryphal event.

¹¹⁵ On idealism in the Indian tradition, see Lusthaus 2002; Trivedi 2005; Siderits 2007; Gold 2011/2021; Garfield 2015; Albahari 2019; Grego 2020.

¹¹⁶ Berkeley 1710-1713/1965, PHK 4.

they have minds or their own? And if the latter, how do we avoid the slippery slope to electron consciousness?

The bizarreness of materialism and dualism might not be immediately evident, manifesting only when details are developed and implications pursued. Idealism, in contrast, is bizarre on its face.

6. The Bizarreness of Compromise/Rejection Views.

There might be an alternative to the classic trio of materialism, substance dualism, and idealism; or there might be a compromise position. Maybe Kant's transcendental idealism is such an alternative or compromise.¹¹⁷ Or maybe some Russellian or Chalmersian neutral monism or property dualism is.¹¹⁸ I won't enter into these views in detail, but I hope it's fair to say that Kant, Russell, and Chalmers do not articulate commonsense views of the metaphysics of mind. For example, Chalmers' property dualism (which allows both irreducibly material and irreducibly nonmaterial properties) offers no good commonsense answer to the problem of immaterial causation or the scope of mentality, tentatively favoring epiphenomenalism and panpsychism: All information processing systems, even thermostats, have conscious experiences or at least "proto-consciousness", but such immaterial properties play no causal role in their physical behavior. In Chapter 5, I will discuss Kant's transcendental idealism at length – the view that "empirical objects" are in some sense constructions of our minds upon an unknowable noumenal reality. The attractions of Kant, Russell, and Chalmers lie, if anywhere, in their elegance and rigor rather than their commonsensicality.

¹¹⁷ Kant 1781/1787/1998.

¹¹⁸ Russell 1921, 1927; Chalmers 1996.

Alternatively, maybe there's no metaphysical fact of the matter. Maybe the issue is so ill-conceived that debate about it is hopelessly misbegotten.¹¹⁹ Or maybe metaphysical questions of this sort run too far beyond the proper bounds of language use to be meaningful.¹²⁰ This type of view is also bizarre. The whole famous mind-body dispute is over nothing real or nothing it makes sense to try to talk about? There is no fact of the matter about whether something in you goes beyond the merely physical or material? We can't legitimately ask whether some immaterial part of you might transcend the grave? It's one thing to allow that facts about transcendent existence might be unknowable – an agnosticism probably within the bounds of commonsense options – and it's one thing to express the view, as some materialists do, that dualists speak gibberish when they invoke the soul; but it's quite another thing, a much more radical and unintuitive thing, to say that there's no legitimate sensible interpretation of the dualist-materialist(-idealist) debate, not even sense enough in it to allow materialists to coherently express their rejection of the dualist's transcendent hopes.

7. Universal Dubiety and Universal Bizarreness.

I am making an empirical claim about the history of philosophy and offering a psychological explanation of this putative empirical fact. The empirical claim is that all well-developed accounts of the metaphysics of the mind are bizarre. Across the entire history of written philosophy, every theory of the relationship between the stream of conscious experience and the apparently material objects around and within us defies common sense. The psychological explanation is that common sense is incoherent in the metaphysics of mind.

¹¹⁹ As suggested in Carnap 1928/1967, Appendix B.

¹²⁰ This might seem a broadly Wittgensteinian position, but it's probably not Wittgenstein's own position; see esp. 1945-1949/1958, p. 178, and 1947/1980, vol. 1, §265.

Common sense, and indeed simple logic, requires that one of four options be true: materialism, dualism, idealism, or a compromise/rejection view. And yet common sense rejects each option, either on its face or implicitly as revealed when theoretical choices are made and implications pursued. If common sense is indeed incoherent, then it will not be possible to develop a non-bizarre metaphysics of mind with specific commitments on tricky issues like mind-body causation and the scope of mentality. This is the Universal Bizarreness thesis.

I aim to conjoin the Universal Bizarreness thesis with a second thesis, Universal Dubiety, to which I now turn. Universal Dubiety is the thesis that none of the various bizarre options compels belief. Even on a fairly coarse slicing of the alternatives – materialism vs. dualism vs. idealism vs. compromise/rejection – no one position probably deserves credence much over 50%. And probably no moderately specific variant, such as materialist functionalism or interactionist substance dualism, merits credence even approaching 50%. If Universal Bizarreness and Universal Dubiety are both true, then every possible approach will be both bizarre and unconvincing and therefore the metaphysics of mind is, in the technical sense of Section 2, a landscape of wild views, a wilderness. Something that seems almost too crazy to believe must be the case.

Part Two: Universal Dubiety

8. An Argument from Disagreement.

Usually, when experts disagree, doubt is the most reasonable response. You might have an opinion about whether the Chinese stock market will rise next year. You might have an opinion about the best explanation of the fall of Rome. But unless you have some privileged

information others lack, you should feel some doubt if there's no consensus among the world's leading experts. You should probably acknowledge, at least, that the evidence doesn't decisively support your preferred view over all others. You might still prefer your view. You might champion it, defend it, argue for it, see the counterarguments as flawed, think those who disagree are failing to appreciate the overall balance of considerations. But appropriate epistemic humility and recognition of your history of sometimes misplaced confidence ought, probably, to inspire substantial uncertainty in moments of judicious assessment. This is true, of course, when you are a novice, but it is also often true when you yourself are among the experts.¹²¹

The world's leading experts disagree about the metaphysics of mind. If we confine ourselves strictly to the best-known, currently active Anglophone philosophers of mind, some – for example, Daniel Dennett and Ned Block – are materialists, of rather different stripes (Dennett focusing more on broad functional structure and patterns of behavior, Block focusing more on specific interior mechanisms).¹²² Others – for example, David Chalmers – are dualists or compromise/rejection theorists, who think that standard-issue materialism omits something important.¹²³ If we cast our net more widely, across a broader range of experts, or across different cultures, or across a broader time period, we find quite a range of materialists, dualists,

¹²¹ For discussion of this issue, including possible exceptions, see Kelly 2005; Christensen 2007; Christensen and Lackey 2013; Frances 2014; Frances and Matheson 2018/2019.

¹²² Block 1978/2007, 2002/2007; Dennett 1991, 2017. Other highly-cited recent materialists or near-materialists include David Lewis and Hilary Putnam (discussed above), David M. Armstrong, Donald Davidson, Fred Dretske, Jerry Fodor, Jaegwon Kim, Ruth Millikan, and John Searle – treating high citation rates in the *Stanford Encyclopedia of Philosophy* as the marker of eminence in recent mainstream Anglophone philosophy (see Schwitzgebel 2019b).

¹²³ Chalmers 1996. Other highly-cited recent critics of materialism include Saul Kripke, early Frank Jackson, and Thomas Nagel.

idealists, and compromise/rejection theorists, of many varieties. The appropriate reaction to this state of affairs should be doubt concerning the metaphysics of mind.

There are two conditions under which expert disagreement can reasonably be disregarded. One condition is when you have good reason to suppose that the experts on all but one side are epistemically deficient in some way, for example, disproportionately biased or ignorant. Consider conspiracy theorists who hold that Hillary Clinton sold children into sex slavery from the back of a pizza restaurant. They might cite many apparently supportive facts and sport a kind of expertise on the issue, but we can reasonably discount their expertise given the universal rejection of this view by sources we have good reason to regard as more careful and even-handed. However, I see no grounds for such dismissals in the metaphysics of mind. Disagreeing experts in the metaphysics of mind appear to have approximately equivalent levels of knowledge and epistemic virtue.

A second condition in which you might (arguably) reasonably disregard expert opinion is when you have thoroughly examined the arguments and evidence of the disagreeing experts, and you remain unconvinced.¹²⁴ If you're already well acquainted with the basis of their opinion, you have already taken into account, as best you can, whatever force their evidence and arguments have. They can appeal to no new considerations you aren't already aware of. Instead, you might treat their disagreement as evidence that perhaps they are not as expert, careful, and sensible as they might otherwise have seemed, or maybe that you have some special insight that they lack.

I take no stand here on the merits of disregarding expert disagreement in ideal conditions in which you are thoroughly familiar with all of the arguments and evidence. In the metaphysics

¹²⁴ See Kelly 2005; though see Christensen 2007; Kelly 2010; Lackey 2010.

of mind, it is simply not possible to examine the grounds of every opposing expert opinion. The field is too large. Instead, expertise is divided. Some philosophers are better versed in *a priori* theoretical armchair arguments, others in arguments from the history of philosophy, others in the empirical literature – and these broad literatures divide into sub-literatures and sub-sub-literatures with which philosophers are differently acquainted. The details of these sub-sub-literatures are sometimes highly relevant to philosophers’ big-picture metaphysical views. One philosopher’s view might turn crucially on the soundness of a favorite response to the Lewis-Nemirow objection to Jackson’s Mary argument.¹²⁵ Another’s might turn on empirical evidence of a tight correlation in invertebrates between the capacity for trace conditioning and having allocentric maplike representations of the world.¹²⁶ Every year, hundreds of directly relevant books and articles are published, plus thousands of indirectly relevant books and articles. No one person could keep up. Even the most expert among us lack relevant evidence that other well-informed, non-epistemically-vicious disagreeing experts have. We must divide assessment among ourselves, relying partly on others’ judgments.

Furthermore, philosophers differ in the profile of skills they possess. Some philosophers are more careful readers of opponents’ views. Some are more facile with complicated formal arguments. Some are more imaginative in constructing hypothetical scenarios. And so on. The evidence and arguments in this area are sufficiently difficult that they challenge the upper boundaries of human capacity in several of these skills. World-class intellectual ability in any of these respects could substantially improve the quality of one’s assessment of arguments in the metaphysics of mind; and no one is so excellent in every relevant intellectual skill that there isn’t

¹²⁵ For example, Raymont 1999 in response to Nemirow 1980, 1990; Lewis 1988/1990.

¹²⁶ See footnote XXX in Chapter 10.

some metaphysician somewhere with a different opinion who isn't importantly better in at least one skill. Even if we all could assess exactly the same evidence and arguments, we might reach different conclusions depending our skills.

Every philosopher's preferred metaphysical position is rejected by a substantial portion of philosophers who are overall approximately as well informed and intellectually skilled and who are also in some respects *better* informed and *more* intellectually skilled. Under these conditions, a high degree of confidence is unwarranted. It's perhaps not unreasonable to retain some trust in your own assessment of the metaphysical situation, preferring it over the assessments of others. It's perhaps not unreasonable to think that you have some modest degree of special insight, good judgment, or philosophical *je ne sais quoi*. This argument doesn't require simply replacing your favorite perspective with some weighted compromise of expert opinion.¹²⁷ All that's required is the realistic acknowledgement of substantial doubt, given the complexity of the issues.

Consider this analogy. Maybe in all of Santa Anita, you're among the four best at picking racehorses. Maybe you're the unmatched very best in some respects – the best, say, at evaluating the relationship between current track conditions and horses' past performance in varying track conditions. And maybe you have some information that none of the other experts have. You chatted privately with two of the jockeys last night. However, the other three expert pickers exceed you in other respects (reading the mood of the horses, assessing fit between jockey and horse) and have some information you lack (some aspects of training history, some

¹²⁷ Thus, this argument does not require adopting “conciliationism” regarding disagreement (Christensen 2007; Frances and Matheson 2018/2019). It is, I think, compatible with moderate versions of steadfastness, as I'm inclined to read Van Inwagen 1996; Foley 2001; Henderson, Horgan, Potrč, and Tierney 2017.

biometric data). If you pick Night Trampler then learn that some of the other experts favor instead Daybreak or Afternoon Valor, you ought to worry.

Or try this thought experiment. You are shut in a seminar room, required to defend your favorite metaphysics of mind for six hours – or six days, if you prefer – against the objections of the world’s leading philosophers of mind. For concreteness, imagine that those philosophers are Ned Block, David Chalmers, Daniel Dennett, Saul Kripke, and Ruth Millikan.¹²⁸ Just in case we aren’t now living in the golden age of metaphysics of mind, let’s invite Aristotle, Dignāga, Hume, Husserl, Kant, Jaegwon Kim, Leibniz, David Lewis, and Zhu Xi too. (First, we’ll catch them up on recent developments.) If you don’t imagine yourself emerging triumphant, then you might want to acknowledge that your grounds for your favorite position might not be compelling.

Consider everyone’s favorite philosophy student. She vigorously champions her opinions while at the same time being intellectually open and acknowledging the substantial doubt that appropriately flows from knowing that her views differ from the views of others who are in many respects more capable and better informed. Concerning the metaphysics of mind, even the best professional philosophers still are such students, or should aspire to be, only in a larger classroom.¹²⁹

¹²⁸ Before debating them, I recommend that you review Block 2007; Chalmers 1996, 2012; Dennett 1991, 2017; Kripke 1980; and Millikan 1984, 2017.

¹²⁹ For arguments resembling those in this section, though not on the metaphysics of mind in particular, see Goldberg 2009; Kornblith 2013; Frances 2013, and for related positive ways forward, see Goldberg 2013 and Barnett 2019. I confine this argument to the metaphysics of mind in particular. Other philosophical issues might not have the required features. On some issues, the most capable and up-to-date experts agree. Other issues are sufficiently small to permit mastery of the entire relevant literature. Other disputes might be terminological, or about broad matters of ethical or aesthetic vision, or concern the weighting of approximately incommensurable factors, complicating the issue of what constitutes and justifies disagreement.

9. *An Argument from Lack of Good Method.*

There is no conscious-ometer. Nor should we expect one soon. (See Chapter 10.) Nor is there a material-world-ometer. The lack of these devices hampers the metaphysics of mind.

Samuel Johnson kicked a stone. Thus, he said, he refuted Berkeley's metaphysical idealism.¹³⁰ Johnson's proof convinces no one with a smudge of sympathy for Berkeley, nor should it. Yet it's hard to see what empirical test could be more to the point. Rudolf Carnap imagines an idealist and a non-idealist both measuring a mountain. There is no experiment on which they will disagree.¹³¹ No multiplicity of gauges, neuroimaging equipment, or particle accelerators could be stronger proof against idealism, it seems, than Johnson's kick. Similarly, Smart, in his influential defense of materialism, admits that no empirical test could distinguish materialism from epiphenomenalist substance dualism (according to which immaterial souls exist but have no causal power).¹³² There is no epiphenomenal-substance-ometer.

[Illustration 3 (Caption: Testing a material-world-ometer): Samuel Johnson kicking a stone on a cobbled street in 18th century London, with James Boswell crouching nearby, aiming a "material-world-ometer" at it. Johnson looks confident and Boswell puzzled. The reading on the material-world-ometer is a question mark.]

¹³⁰ Boswell 1791/1980, p. 333.

¹³¹ Carnap 1928/1967, p. 333-334. While accepting Carnap's view that "metaphysical" disputes are largely meaningless because unverifiable, Schlick 1936 emphasizes that there could in principle be empirical evidence of existence without a body, for example through some successful scientific alternative to séances or through observation of your own disembodied existence after death. (In an unfortunate irony, Schlick's article about this was published posthumously, just a month after he was murdered by a Nazi extremist.)

¹³² Smart 1959, p. 155-156.

Why, then, should we be materialists? Smart appeals to Occam's razor: Materialism is simpler. But simplicity is a complex business. Arguably, Berkeley's idealism is simpler than either dualism or materialism. According to Berkeley, all that exists is our minds, God's mind, and the ideas we all share in a common, carefully choreographed dance, with no nonmental entities at all. Solipsism seems simpler yet: just my mind and its ideas, nothing else at all. (Chapter 6 will treat solipsism in more detail.) Anyhow, simplicity is at best one theoretical virtue among several, to be balanced in the mix. Abstract theoretical virtues like simplicity, fecundity, and explanatory power will, I suggest, attach only indecisively, non-compellingly, to the genuine metaphysical contenders. I'm not sure how to argue for this other than to invite you sympathetically to feel the abstract beauty of some of the contending views apart from your favorite. Materialism has its elegance. But so also do the views of Berkeley, Kant, and Chalmers.

If you're willing to commit to materialism, you might still hope for a consciousness-ometer that we could press against a human or animal head to decide among, say, relatively conservative versus moderate versus liberal views of the sparseness or abundance of consciousness in the world. But, as I will argue in Chapter 10, even this is too much to expect in our lifetimes. Imagine a well-informed, up-to-date conservative about consciousness, who holds that consciousness is a rare achievement, requiring substantial cognitive abilities or a specific type of neural architecture. Thus, the conservative holds, consciousness is limited on Earth solely to humans, or solely to the most sophisticated mammals and birds. Imagine also a well-informed, up-to-date liberal about consciousness, who holds that even simple animals like earthworms have some rudimentary consciousness. If this conservative and this liberal disagree about the consciousness of, say, a garden snail, no behavioral test or measure of neural activity is

likely to resolve their disagreement – not unless they have much more in common than is generally shared by conservatives and liberals about the abundance of consciousness. Their disagreement won't be resolved by details of gastropod physiology. It's bigger, more fundamental. Such foundational theoretical disagreement prevents the construction of a conscious-ometer whose applicability could be widely accepted by empirically-oriented materialists.¹³³

Thus I suggest: Major metaphysical issues of mind are resistant enough to empirical resolution that none compel belief on empirical grounds, and this situation is unlikely to change for the foreseeable future. Neither do these issues permit resolution by appeal to common sense, which will rebel against all and is an unreliable guide anyway. Nor do they permit resolution by appeal to broad, abstract theoretical considerations. I see no other means of settling the matter. We have no decisive method for resolving fundamental issues in the metaphysics of mind – indeed, not even any particularly good method.

However, I am not recommending complete surrender. Metaphysical views can be better or worse, more credible or less credible. A view on which people have immaterial souls for exactly seventeen minutes on their eighteenth birthday has no merit by the standards of common sense, empirical science, or theoretical elegance, and it deserves extremely close to zero credence. Despite the shortcomings of common sense, empirical science, and appeals to abstract theoretical virtue as metaphysical tools, if we intend to confront questions in the metaphysics of mind, we have to do our unconfident best with them. It is reasonable to distribute one's credence

¹³³ In Schwitzgebel 2011b, ch 6, I argue similarly for the intractability of the question of how sparse or abundant human experience is – the question of whether, for example, people have constant tactile experience of their feet in their shoes. On this question too, theories diverge radically, and it is virtually impossible for theorists on one side of the gulf to gain non-question-begging leverage against theorists on the other side.

unequally among the four main metaphysical options, and then among subsets of those options, on some combination of scientific, commonsensical, and abstract theoretical grounds.¹³⁴

10. An Argument from Cosmological Dubiety.

If broad-reaching cosmological dubiety is warranted, so too is dubiety about the metaphysics of mind. If we don't know how the universe works, we don't know how the mind fits within it.

I have already mentioned the bizarreness of quantum mechanics and the lack of consensus about its interpretation. Some interpretations treat mentality as fundamental, such as versions of the Copenhagen interpretation on which a conscious observer causes the collapse of the wave function.¹³⁵ The famous physicist Stephen Hawking, appealing to such views, has even said that quantum cosmology implies the backward causation of the history of the universe by our current acts of scientific observation.¹³⁶ Consider also that the standard equations of quantum mechanics aren't relativistic and cannot easily be made so, leading to the well-known apparent conflict between relativity theory and quantum theory.¹³⁷ We don't yet understand fundamental physics.

Consider also the "fine-tuning argument".¹³⁸ If the gravitational constant were a bit higher, stars would be too short lived to permit the evolution of complex life around them. If the

¹³⁴ If you're curious, my own credences are about 55% materialism, 35% compromise/rejection, 5% dualism, 5% idealism.

¹³⁵ Wigner 1961; Faye 2008/2019; Radin, Michel, Galdamez, Wendland, Rickenbach, and Delorme 2012 (though see Walleczek and von Stillfried 2019); Chalmers and McQueen forthcoming.

¹³⁶ Hawking and Mlodinow 2010, p. 140.

¹³⁷ Einstein, Podolsky, and Rosen 1935; Bell 1964; Maudlin 1994/2002.

¹³⁸ See Barrow, Morris, Freeland, and Harper 2008. Against fine-tuning, see Stenger 2011. For a philosophical review of the issue, see Friederich 2017/2018.

gravitational constant were a bit lower, stars would not explode into supernovae, the main source of the heavier elements that enable the complexity of life. Similarly, if the mass of the electron were a bit different, or if the strong nuclear force were a bit different, the universe would also be too simple to support life. In light of our universe's apparent fine-tuning for life, many cosmologies posit either a creator who set the physical constants or initial conditions at the time of the Big Bang so as to support the eventual occurrence of life, or alternatively the real existence of a vast number of universes with different physical constants or conditions. (In the latter case, it would be no surprise that some minority of universes happen to have the right conditions for life and that observers like us would inhabit such rare universes.) If an immaterial entity might have fashioned the physical constants, then we cannot justifiably rest assured the materialism is true. If there might really exist actual universes radically different from our own, perhaps some are so radically different that cognition transpires without the existence of anything we would rightly call material. Then materialism would be at best a provincial contingency.¹³⁹

Another issue is this. If consciousness can be created in artificial networks manipulated by external users – for example, in computer programs run by children for entertainment – and if the entities inside those networks can be kept ignorant of their nature, then there could be entities who are seriously mistaken about fundamental metaphysics and cosmology. Such entities might think they live in a wide world of people like them when in fact they have three-hour lives, isolated from all but their creator and whatever other entities are instantiated in the same temporary environment. I will argue in Chapter 4 that you should not entirely dismiss the

¹³⁹ For readers fussy about modal logic: I am asserting an epistemic possibility concerning actual universes rather than making a metaphysical claim about possible universes.

possibility that you are such an entity. In Chapter 5, I will extend this idea into a form of quasi-Kantian transcendental idealism, according to which we might be irreparably ignorant of the fundamental nature of things. At root, the world might be very different than we imagine and not material at all. Matter might be less fundamental than mind or information or some other ontological type that we can't even conceive or name. These possibilities are, of course, pretty wild. But are they too wild to figure in a list of live cosmological possibilities? Are they more than an order of magnitude more bizarre and dubious than multiverse theory or the typical well-developed religious cosmology? There are no commonsense cosmologies left.

Further support for cosmological dubiety comes from our apparently minuscule perspective. If mainstream scientific cosmology is correct, we have seen only a very small, perhaps an infinitesimal portion of reality. We are like fleas on the back of a dog, watching a hair grow and saying, "Ah, so *that's* how the universe works!"¹⁴⁰

Scientific cosmology is deeply and pervasively bizarre; it is highly conjectural in its conclusions; it has proven unstable over the decades; and experts persistently disagree on fundamental issues. Nor is it even uniformly materialist. If materialism draws its motivation from being securely and straightforwardly the best scientific account of the fundamental nature of things, materialists should think twice. I focus on materialism since it is the leading view right now, as well as my own personal favorite, but similar considerations cast doubt on dualism, idealism, and compromise/rejection views.

11. Conclusion.

¹⁴⁰ Image inspired by Hume 1779/1947, §II, p. 147-149.

Certain fundamental questions about the nature of the mind and its relation to the apparently material world can't, it seems, be settled by empirical science in anything like its current state, nor by abstract reasoning. To address these questions, we must turn to common sense. If common sense, too, is no reliable guide, we are unmoored. Without common sense as a constraint, the possibilities open up, wild and beautiful in their different ways – and once open, they refuse to shut. The metaphysics of mind tangles with fundamental cosmology, and every live option is bizarre and dubious.

The Weirdness of the World

Chapter Four 1% Skepticism

Certainly there is no practical problem regarding skepticism about the external world. For example, no one is paralyzed from action by reading about skeptical considerations or evaluating skeptical arguments. Even if one cannot figure out where a particular skeptical argument goes wrong, life goes on just the same. Similarly, there is no “existential” problem here. Reading skeptical arguments does not throw one into a state of existential dread. One is not typically disturbed or disconcerted for any length of time. One does not feel any less at home in the world, or go about worrying that one’s life might be no more than a dream.

- Greco 2008, p. 109

[W]hen they suspended judgement, tranquility followed as it were fortuitously, as a shadow follows a body.... [T]he aim of Sceptics is tranquility in matters of opinion and moderation of feeling in matters forced upon us.

- Sextus Empiricus c. 200 CE/1994, I.xii, p. 30

I have about a 1% credence in the disjunction of all radically skeptical scenarios combined. That is to say: I am about 99% confident that I am awake, not massively deluded, and have existed for decades in roughly the form I think I have existed, doing roughly the sorts of things I think I have been doing; and I find it about 1% subjectively likely that instead some radically skeptical possibility obtains – for example, that I am a radically deceived brain in a vat, or that I am

currently dreaming, or that I first came into existence only a few moments ago. *Probably* I live in the wide, stable world I think I do, but I'm not more confident of that than I am that this lottery ticket I just bought today, for the sake of writing this sentence, will lose.¹⁴¹

In this chapter, I aim to convince you to embrace a similar 99%-1% credence distribution, on the assumption that you currently have much less than a 1% credence (degree of belief, confidence) in radically skeptical possibilities. Probably, in your daily decisions, you disregard skeptical scenarios entirely! I will also argue that, since 1% is non-trivial in affairs of this magnitude, daily decisions ought sometimes to be influenced by radically skeptical possibilities. The 1% skeptic ought to live life a bit differently from the 0% skeptic.

I don't insist on precisely 1%. "Somewhere around 0.1% to 1%, plus or minus an order of magnitude" will do, or "highly unlikely but not as unlikely as a plane crash".

Part One: Grounds for Doubt

1. Grounds for Doubt: Dreams.

Sitting here alone in a repurposed corner of my son's bedroom amid the COVID-19 pandemic, I am almost certain that I am awake.

I can justify this near certainty, I think, on phenomenological grounds. It's highly unlikely that I would be having experiences *like this* if I were asleep. My confidence might be defensible for other reasons too, as I'll soon discuss. My current experience has both general and specific features that I think warrant the conclusion that I am awake. The general features are two.

¹⁴¹ Update: It lost.

First general feature: I currently have detailed sensory experience in multiple modalities. That is, I currently have a visual sensory experience of the card table holding my computer monitor; of the fan, lamp, papers, and detached webcam cluttering the table; of window slats and backyard trees and grass in the periphery. Simultaneously, I have auditory experience of the tap of my fingers on the keyboard and the repeated chirps of an angry bird. Maybe I also currently have tactile experience of my fingers on the keys, my back against the chair, the lingering taste of coffee in my mouth, a proprioceptive general sense of my bodily posture, and so on.¹⁴² But according to the theories of dream experience I favor, the experience of dreaming is *imagistic* rather than *sensory* and also rather sketchy or sparse in detail, including for example in the specification of color – more like the images that occur in a vivid daydream than like the normal sensory experiences of waking life.¹⁴³ On such views, the experience of dreaming that I am on the field at Waterloo is more like the experience of imagining that I am on the field at Waterloo than it is like standing on the field at Waterloo taking in the sights and sounds; and dreaming that I'm in my son's room is more like the sketchy imagery experience I have while lying in my bed at night thinking about working tomorrow than it is like my current experience of seeing my computer screen, hearing my fingers on the keyboard, and so on.

¹⁴² For discussion of how sparse or abundant our sensory experience is, see Searle 1992; Block 2007; Schwitzgebel 2011b; Dehaene 2014; Cohen, Dennett, and Kanwisher 2016. (Much of the empirical literature on this question is limited by focusing on experience in a single modality (usually vision) rather than on the broader question.) I am assuming that experience is not sparse for me at this moment.

¹⁴³ For defenses of the imagery theory of dream experience, see McGinn 2004; Ichikawa 2009, 2016; Whiteley 2021; and possibly Sartre 1936/1962. However, Thomas 2014 and Windt 2015 argue convincingly that the imagery/perception distinction isn't sharp but instead is a multi-dimensional spectrum along which hallucinations, hypnagogic imagery, and dreams occupy the middle regions. Even if dreams aren't exactly imagistic, as long as most dreams are phenomenologically different enough from the sorts of waking experiences I'm having now, the argument of this paragraph still works.

Second general feature: Everything appears to be mundane, stable, continuous, and coherently integrated with my memories of the past. I seem to remember how I got here (by walking across the house after helping my wife put away the marshmallows), and this memory seems to fit nicely with my memories of what I was doing this morning and yesterday and last week and last year. Nothing seems weirdly gappy or inexplicably strange. I share Descartes' view that if this were a dream, it would probably involve discontinuities in perception and memory that would be evident once I thought to look for them.¹⁴⁴

Some *specific features* of my current conscious experience also bolster my confidence that I'm awake. I think of written text as typically unstable in dream experiences. Words won't stay put. They change and scatter away. But here I am, visually experiencing a stable page. And pain is not as vividly felt as in waking life. But I have just pinched myself and experienced the vivid sting. Light switches don't change the ambient lighting. But I have just flicked the switches, or seemed to, changing the apparent ambient light. If you, reader, are like me, you might have your own favorite tests.¹⁴⁵

But here's the question. Are these general and specific features of my experience sufficient to justify not merely *very high* confidence that I am awake – say 99.9% confidence – but all-out 100% confidence? I think not. I'm not all *that* sure I couldn't dream of a vivid pinch

¹⁴⁴ Descartes 1641/1984, p. 89-90/61-62. I agree with Domhoff 2003 (p. 45, 152-154) that the bizarreness of dreams might tend to be overrated, especially if mundane dreams are less likely to be remembered than bizarre ones. But even if, as Domhoff suggests, dreams are not much more bizarre or discontinuous than ordinary waking relaxed thought, ordinary waking relaxed thought involves frequent scene discontinuities and speculative associations, including hypotheticals inconsistent with previous hypotheticals, in a manner quite different from mundane sensory experience in a stable environment. See Grundmann 2002 for a similar argument. For a review of the literature on dream bizarreness, see Rosen 2018.

¹⁴⁵ See LaBerge and Rheingold 1990 for an extended discussion of “dreamsigns” and methods of “state testing”.

or a stable page. One worry: I seem to recall “false awakenings” in which I judged myself awake in a mundane, stable world. More fundamentally, I’m not all that sure that my favorite theory of dreams is correct. Other dream theorists have held that dreams are sometimes highly realistic and even experientially indistinguishable from waking life – for example Antti Revonsuo, Allan Hobson, Jennifer Windt, and Melanie Rosen.¹⁴⁶ Eminent disagreement! I wouldn’t risk a thousand dollars for the sake of one dollar on denying the possibility that I often have experiences much like this, in sleep. I’m not sure I’d even risk \$1000 for \$250.

But even if I can’t point to a feature of my current experience that seems to warrant 100% confidence in my wakefulness, might there still be a philosophical argument that would rationally deliver 100% confidence? Maybe externalist reliabilism about justification is true, for example. On a simple version of externalist reliabilism, what’s crucial is only that my disposition to believe that I’m awake is hooked up in the right kind of reliable way to the fact that I am awake, such that I wouldn’t now be judging myself awake unless it were truly so.¹⁴⁷ If I accepted that view, and if I also felt absolutely certain, then maybe I would be justified in that absolute certainty.¹⁴⁸ Or here’s a very different philosophical argument: Maybe any successful

¹⁴⁶ Revonsuo 1995; Hobson, Pace-Schott, and Stickgold 2000; Rosen 2019; and to a lesser extent Windt 2015, 2017. Compared to these authors, I have less confidence in the accuracy of people’s dream reports of sensorily realistic dreams (see Schwitzgebel 2011b, Ch. 1) and more doubt about the capacity of the brain to construct fully realistic and detailed sensory experience without the ordinary stream of sensory input. I refer to the mistaken belief that you are having or have had hallucinatory experience X as a *doxastic hallucination*, in contrast with a phenomenal hallucination in which you actually are having or have had experience X (Schwitzgebel 2013b). Reports of stable, richly detailed sensory dream experiences would then, on my preferred view, be doxastic hallucinations.

¹⁴⁷ See Goldman and Beddor 2008/2021 for a review.

¹⁴⁸ See Sosa 2007 for a more subtle version of this argument. I am sympathetic with the critiques of Sosa in Ichikawa 2008; Brown 2009; Ballantyne and Evans 2010; Šuster 2016. But even if I were to think that Sosa’s argument *probably* worked, the central argument of this section should still succeed, as long as it is rational for me to take some skeptical distance from philosophical arguments of this sort and withhold 100% credence on that basis.

referent of “I” necessarily picks out a waking person, so that if I succeed in referring to myself I must indeed be awake.¹⁴⁹ For me at least, such philosophical arguments don’t deliver 100% confidence. The philosophical theories, though in some ways attractive, don’t fully win me over. I’m insufficiently confident in their truth.¹⁵⁰

Is it, maybe, just *constitutive of being rational* that I assume with 100% confidence that I am awake? For example, maybe rationality requires me to treat my wakefulness as an unchallengeable framework (or “hinge”) assumption, in Wittgenstein’s sense.¹⁵¹ I feel some of the allure of that idea. But the thought that such a theory of rationality might be correct, though comforting, does not vault my confidence to the ceiling.

All the reasons I can think of for being confident that I’m awake seem to admit of some doubt. Even stacking these reasons together, doubt remains. To put a number on my doubt suggests more precision than I really intend, but the non-numerical terms of ordinary English have the complementary vice of insufficient clarity. So, with caveats: Given my reflections above, a 90% credence that I’m awake seems unreasonably low. I am much more confident that I’m awake than that a coin flipped four times will come up heads at least once. On the other hand, a 99.999% credence seems unreasonably high, now that I’ve paused to think things through. Neither my apparent phenomenological or experiential grounds nor my dubious

¹⁴⁹ Compare Valberg 2007 on “immanent” dream skepticism. As Valberg emphasizes, such an argument is consistent with the “transcendent” possibility that “THIS” is all a dream. The present section could be accordingly recast.

¹⁵⁰ One way of understanding this last claim is that I have a non-100% higher-order credence that it would be rational to assign 100% credence to the proposition that I’m awake. I might then apply some version of a weighted reflection principle, as in Elga 2013. (See also Lasonen-Aarnio 2014; Silwa and Horowitz 2015; Christensen 2021.) The reasoning of this paragraph fails if some such anti-skeptical consideration ought to give me 100% confidence in my wakefulness, in a way that is immune to undermining by higher-order doubts.

¹⁵¹ Wittgenstein 1951/1969; Coliva 2010; Pritchard 2016.

philosophical theorizing about the nature of justification seems to license so extreme a credence. So I'll split the difference: a 99.9% credence seems about right, give or take an order of magnitude.

To think of it another way: Multiplying a 20% credence that I'm wrong about the general features of dreams by a 20% credence, conditional upon my being wrong about dreams in general, that while dreaming I commonly have mundane working-day experiences like this present experience, yields a 4% credence that I commonly have mundane experiences like these in my dreams – that is, a 96% credence that I don't commonly have experiences like this in my dreams. That seems a rather high credence, really, for that particular theoretical proposition, given the uncertain nature of dream science and the range of expert opinions about it; but suppose I allow it. Once I admit even a 4% credence that this type of experience is common in dreams, however, it's hard for me to see a good epistemic path down to a 0.001% or lower credence that I'm now dreaming.

2. Grounds for Doubt: Simulation Skepticism.

Some philosophers have argued that digital computers could never have conscious experience.¹⁵² There's a chance they are right about that. But they might be wrong. And if they are wrong, it might be possible to create conscious beings who live entirely within simulated environments inside of computers – like Moriarty in Star Trek's "Ship in a Bottle" or the "citizens" who live long, complex lives within powerful, sheltered computers in Greg Egan's

¹⁵² Especially Searle 1980, 1984. On the complexities involved in thinking of the mind as a computer, and stronger versus weaker interpretations of that claim, see Piccinini and Scarantino 2011; Piccinini and Bahar 2013.

novel *Diaspora*.¹⁵³ Some mainstream views about the relation of mind and world – views that, in the spirit of Chapter 3, I don't think we should entirely rule out – imply that it's at least possible for genuinely conscious people to live as computer programs inside entirely artificial, computerized environments.¹⁵⁴ Let's call these entities "sims". *If* sims exist, some of them might be ignorant of their real ontological status. They might not realize that they are sims. They might think that their world is not the computational creation of some other set of entities. They might even think that they live on "Earth" in the early "21st century".

Normally I assume that I'm not an entity instantiated in someone else's computational device. Might that assumption be wrong?

Nick Bostrom argues that we should assign about a one-third credence to being sims of this sort.¹⁵⁵ He invites us to imagine the possibility of technologically advanced civilizations able to cheaply run vastly many "ancestor simulations" that contain conscious, self-reflective, and philosophically thoughtful entities with experiences and attitudes similar to our own. Bostrom suggests that we give approximately equal epistemic weight to three possibilities: (a.) that such technologically advanced civilizations do not arise, (b.) that such civilizations do arise but choose not to run vastly many ancestor simulations, and (c.) that such civilizations do arise and the world as a whole contains vastly many more sims than non-sims. Since in the third scenario the vast majority of entities who are in a subjective and epistemic situation relevantly similar to our own are sims, Bostrom argues that our credence that we ourselves are sims should approximately equal our credence in the third scenario, about one in three.¹⁵⁶

¹⁵³ Echevarria 1993; Egan 1997.

¹⁵⁴ For example, Bostrom 2003; Chalmers 2003/2010, 2010b; Hanson 2018.

¹⁵⁵ Bostrom 2003; Bostrom and Kulczycki 2011.

¹⁵⁶ On similar grounds, Chalmers (forthcoming) suggests that there's at least a 25% chance that we are sims. Among the signs that we may be sims, he suggests, are that we appear

You might object to the starting assumptions, as people would who deny the possibility that consciousness could arise from any kind of digital computer. Or you might press technical objections regarding Bostrom's application of an indifference principle in the final argumentative move.¹⁵⁷ More generally, you might challenge the notion that sims and non-sims are in relevantly similar epistemic situations. Or you might object that to whatever extent we take seriously the possibility that we are sims, to that same extent we undercut our grounds for conjecturing about the future of computation.¹⁵⁸ Or you might object on technological grounds: Maybe Bostrom overestimates the feasibility of cheaply running so many ancestor simulations even in highly advanced civilizations. Legitimate concerns, all.

And yet none of these concerns seems to imply that we should assign *zero* credence to our being sims in Bostrom's sense. What if we assigned a 0.1% credence? Is that plainly too high?¹⁵⁹ Giving the amazing (seeming-)trajectory of computation through the 20th and 21st centuries, and given the philosophical and empirical tenuousness of objections to Bostrom's argument, it seems reasonable to reserve a non-trivial credence for the possibility that the world

to exist relatively early in the universe. We have neither encountered other civilizations nor have we ourselves been able to manufacture conscious sims. Modeling the early universe might be easier than modeling a universe with many interacting civilizations, and consequently disproportionately many sims might find themselves in lonely universes like ours. (A more pessimistic possibility consistent with the same evidence is that technological societies are self-destructive and we're near the end of our run.)

¹⁵⁷ Weatherson 2003. For related technical arguments involving uncertainty about self-location, see the Doomsday Argument (Gott 1993 and subsequent literature) and the Sleeping Beauty problem (Elga 2000 and subsequent literature).

¹⁵⁸ Birch 2013; Crawford 2013.

¹⁵⁹ Opinions about the possibility that we are living in a simulation seem to vary considerably by age. In informal conversation with teenagers and people in their early 20s, I commonly hear credences of five, twenty, fifty percent or more. In contrast, except among the most technophilic, few of my acquaintances over the age of 50 appear to take the possibility seriously.

contains many simulated entities. And if so, it seems reasonable to reserve some non-trivial sub-portion of that credence for the possibility that we ourselves are among those simulated entities.

As David Chalmers has emphasized, the simulation possibility needn't be seen as a *skeptical* possibility.¹⁶⁰ Chalmers analogizes to Berkeleyan idealism. Recall from Chapter 3 that Berkeley holds that the world is fundamentally composed of minds and their ideas, coordinated by God. Material stuff doesn't exist at all. If you and I both see a cup, what's happening is not that there's a material cup out there. Rather, I have visual (and tactile, etc.) experiences as of a cup, and so do you, and so does ever-watching God, and God coordinates things so that all of these experiences match up – including that when we leave the room and return, we experience the cup as being exactly where we left it.¹⁶¹ This is no form of skepticism, Berkeley repeatedly insists. Cups, houses, rivers, brains, and fingers all exist and can be depended upon – only they are metaphysically constituted differently than people normally suppose. Likewise, if the simulation hypothesis is correct, we and our surroundings might be fundamentally constituted by computational processes in high-tech computers but still have reality enough that most of our commonsense beliefs qualify as true.

To qualify as a radically skeptical scenario in the same sense that the dream scenario is a radically skeptical scenario or the “I'm just a brain in a vat” scenario is a radically skeptical scenario, we must be either in a *small* simulation or an *unstable* simulation – maybe a short-term simulation booted up only a few minutes ago in our subjective time (with all our seeming-memories in place, etc.) and doomed for deletion soon, or maybe a spatially small simulation containing only this room or this city, or maybe a simulation with inconstant laws that are due

¹⁶⁰ Chalmers 2003/2010, 2018.

¹⁶¹ Berkeley 1710-1713/1965. For a related view, see the quasi-Kantian transcendental idealism of Chapter 5.

soon for a catastrophic change. Only in simulations of this sort are large swaths of our everyday beliefs about mundane, daily things in fact mistaken, if we accept the solace offered by Berkeley and Chalmers, according to which a firewall shields most everyday beliefs from the wild flames of fundamental metaphysics.

Conditionally upon the chance that I am a sim, how much of my credence should I distribute to the possibility that I'm in a large, stable simulation, and how much should I distribute to the possibility that I'm in a small or unstable simulation? Philosophical advocates of the simulation hypothesis have tended to emphasize the more optimistic, less skeptical possibilities.¹⁶² However, it's unclear what would justify a high degree of optimism. Maybe the best way to develop conscious entities within simulations is to evolve them up slowly in giant, stable sim-planets that endure for thousands or millions or billions of years. That would be comforting. But maybe it's just as easy, or easier, to cut and copy and splice and spawn variants off a template, creating person after person within small sims. Maybe it's convenient, or fun, or beautiful to create countless pre-fab worlds that endure briefly as scientific experiments, toys, or works of art, like the millions of sample cities pre-packaged with the 21st-century computer game SimCity.

Our age is awestruck by digital computers, but we should also bear in mind that simulation might take another form entirely. A simulation could conceivably be created from ordinarily-structured analog physical materials, at a scale that is small relative to the size and power of the designers – a miniature sandbox world. Or the far future might contain technologies as different from electronic computers as electronic computers are from clockwork,

¹⁶² Bostrom 2003, reaffirmed in dialogue with me: Schwitzgebel 2011a; Chalmers 2003/2010, 2018; Steinhart 2014; Hanson 2018.

giving rise to conscious entities in a manner we can't now even begin to understand. Simulation skepticism need not depend entirely on hypotheses about digital computing technology. As long as we might be artificially created playthings, and as long as in our role as playthings we might be radically mistaken in our ordinary day-to-day confidence about yesterday, tomorrow, or the existence of Luxembourg, then we might be sims in the sense relevant to sim-skepticism.

Bostrom's argument for about a one-third credence that we are sims should be salted with caveats both philosophical and technological. And yet it seems that we have positive empirical and philosophical reason to assign *some* non-trivial credence to the possibility that the world contains many sims, and conditionally upon that to the possibility that we are among the sims, and conditionally upon that to the possibility that we (or you, or I) are in a simulated environment small enough or unstable enough to qualify as a skeptical scenario. Multiplying these credences together, we should probably be quite confident that we aren't sims in a small or unstable environment, but it's hard to see grounds for being *hugely* confident of that. Again, a reasonable credence might be 99.9%, plus or minus an order of magnitude.

3. Grounds for Doubt: Cosmological Skepticism.

According to mainstream physical theory, there's an extremely small but finite chance that a molecule-for-molecule duplicate of you (within some arbitrarily small error tolerance) could spontaneously congeal, by chance, from disorganized matter. This is sometimes called the *Boltzmann brain* scenario, after physicist Ludwig Boltzmann, who conjectured that our whole galaxy might have originated from random fluctuation, given infinite time to do so. Wait long

enough and eventually, from thin chaos, by minuscule-upon-minuscule chance, things will align just right. A twin of you, or at least of your brain, will coalesce into existence.¹⁶³

It seems reasonable to assign non-trivial credence to the following philosophical hypothetical: *If* such a Boltzmann brain or “freak observer” emerged, it could have some philosophical and cosmological thoughts, including possibly the thought that it is living on a full-sized, long-enduring planet that contains philosophers and physicists contemplating the Boltzmann brain hypothesis. The standard view is that it is vastly more likely for a relatively small freak system to emerge than for a relatively large one to emerge. If so, almost all freak observers who think they are living on long-enduring planets and who think they will survive to see tomorrow are in fact mistaken. They are blips in the chaos that will soon consume them.

Might I be a doomed freak observer?

Well, how many freak observers exist, compared to human-like observers who have arisen by what I think of as the more normal process of evolution in a large, stable system? If the world is one way – for example, if the universe endures infinitely and is prone to fluctuations even after the stars have long since burned out – then the ratio of freaks to normals might be large as the size of the spatiotemporal region under consideration goes to infinity. If the world is another way – for example, if it is spatiotemporally finite or settles into an unfluctuating state within some reasonable period – there might not be a single freak anywhere. If the world is still another way, the ratio might be 50-50. Do I have any compelling reason to think I’m almost certainly in a world with a highly favorable ratio of freaks to normals? Given the general

¹⁶³ For review and discussion, see Carroll 2021; also Boltzmann 1897; Bostrom 2003; Albrecht 2004; Carroll 2010; De Simone, Guth, Linde, Noorbala, Salem, and Vilenkin 2010; Crawford 2013; Chen 2021; Kotzen 2021. For more detailed replies to common arguments against the Boltzmann brain hypothesis, see Schwitzgebel 2015a; Carroll 2021.

dubiety of (what I think of as) current cosmological theorizing, it seems rash to be hugely confident about which sort of world this is.

But maybe my rational credence here needn't depend on such cosmological ratios. Suppose there are a million doomed freak duplicates of me who think they are "Eric Schwitzgebel" writing about the philosophy of Boltzmann brains, etc., for every one evolved-up "Eric Schwitzgebel" on a large, stable rock. Maybe there's a good philosophical argument that the evolved up Erics should rationally assign 99.9999% or more credence to being non-freaks, even after they have embraced cosmological dubiety about the ratio of freaks to normals. Maybe it's just constitutive of my rationality that I'm entirely sure I'm not a freak; or maybe it's an unchallengeable framework assumption; or maybe there's some externally secured forward-looking reliability, which a freak can't have, that warrants stable-Eric's supreme confidence – but again, as in the dream case, it seems unreasonable to be highly certain such philosophical arguments succeed. Similarly for the "externalist" idea that a bare brain, however structured, couldn't actually entertain the relevant sort of cosmological thoughts.¹⁶⁴ In the face of both

¹⁶⁴ On some "externalist" views (e.g., Putnam 1975, 1981) and views in which thought requires a history of natural functioning (e.g., Millikan 1984, 2017), a bare brain in the void – even if structured exactly like a real human brain – could not have genuine thoughts with the right kind of content. At least two paths are open to the 1% skeptic who is attracted to externalism of this sort, even independently of philosophical doubt about such externalist or historical-functional theories. (1.) We could consider only sufficiently large fluctuations – as large as necessary to encompass the right amount of environment or history. Since there appears to be no limit on the size or energy of possible freak fluctuations, this could even be a whole freak Earth, freak Sun, freak light coming as if from distant stars, and so on. If the minimum fluctuation required for genuine cosmological thoughts includes at least the Sun, the case for a short future life becomes more complicated, since an Earth and Sun system might persist for years in near vacuum, and maybe – *maybe* – the baseline state from which fluctuations arise could be assumed to be near vacuum. However, presumably the "stars" would almost all promptly vanish. (2.) We could allow that a bare freak brain would not have any genuinely cosmological thoughts – but then it's open to wonder how we know that *we* are having such thoughts, with all the necessary external conditions met, rather than whatever variety of thought-simulacra a Boltzmann brain might possess.

cosmological and philosophical doubt, I see no epistemically responsible path to supreme certainty.

Here's another cosmological possibility: Some divine entity made the world. One reason to take the possibility seriously is that atheistic cosmology has not yet produced a stable scientific consensus about the ultimate origin of the universe – that is, about what, if anything, preceded, caused, or explains the Big Bang.

Although it seems possible that some entity intentionally designed and launched the universe, a sober look at the evidence does not compel the conclusion that if such an entity exists it is benevolent. The creator or creators of the universe might be perfectly happy to produce an abundance of doomed freaks. They might not mind deceiving us, might even enjoy doing so or regard it as a moral obligation.¹⁶⁵ Maybe God is a clumsy architect and I am one of a million trial runs, alone in my room, like an artist's quick practice sketch. Maybe God is a sadistic adolescent who has made a temporary Earth so that he can watch us fight like ants in a jar. Maybe God is a giant computer running every possible set of instructions, most of which, after the Nth instruction, produce only chaotic results in the N+1th.¹⁶⁶ (Some of these scenarios overlap the simulation scenarios.) Theology is an uncertain endeavor.

It seems unreasonable to maintain extremely high confidence – 99.9999% or more – that our position in the universe is approximately what we think it is. The Boltzmann brain hypothesis, the sadistic adolescent deity hypothesis, the trial-run hypothesis – these possibilities

¹⁶⁵ I favor the view that if God is omniscient, omnipotent, and intentionally arranged things to influence 21st-century philosophers' cosmological beliefs, God is a deceiver, since the cosmos, and the history of evil and suffering on Earth to date, deceptively suggest the non-existence of such a God. Contrast Descartes' attempt in the *Meditations* (1641/1984) to prove that God is not a deceiver.

¹⁶⁶ I explore these two possibilities in detail in my science fiction stories "Out of the Jar" (Schwitzgebel 2015d) and "THE TURING MACHINES OF BABEL" (Schwitzgebel 2017b).

are only a start. Metaphysical doubt (Chapter 3) engenders cosmological doubt, which opens up a wide range of weird but epistemically possible scenarios.

Either the world is huge and I have an extremely limited perspective on its expanse, beginnings, and underlying structure, or the world is tiny and I'm radically mistaken in thinking it's huge. I ought not be supremely confident I've correctly discerned my position within it. It is rational, I think, to reserve a small credence – again I'd suggest about 0.1% – for epistemically catastrophic cosmological possibilities in which I'm radically wrong about huge portions of what I normally regard as obvious.

4. Grounds for Doubt: Wildcard Skepticism.

These three skeptical scenarios – dream skepticism, simulation skepticism, and cosmological skepticism – are the only specific skeptical worries that currently draw a non-trivial portion of my credence. I think I have *grounds for doubt* in each case. People dream frequently and on some leading theories dreams and waking life are often indistinguishable. Starting from a fairly commonplace set of 21st-century mainstream Anglophone cultural attitudes, one can find positive reasons to assign a small but non-trivial credence to simulation doubts and cosmological doubts.

In contrast, I currently see no good reason to assign even a 0.0001% credence to the hypothesis that aliens envatted my brain last night and are now feeding it fake input. Even if I think such a thing is possible, nothing in my existing network of beliefs points toward any but an extremely tiny chance that it's true. I also find it difficult to assign much credence to the hypothesis that I'm a deluded madman living in a ditch or asylum, hallucinating a workplace and believing I am the not very well-known philosopher Eric Schwitzgebel. No positive argument

gives these doubts much grip. Maybe if I thought I was as immensely famous and interesting as Wittgenstein, I would have reasonable grounds for a non-trivial sliver of madman doubt.

But I also think that if I knew more, my epistemic situation might change. Maybe I'm underestimating the evidence for envatted aliens; maybe I'm overestimating the gulf between my epistemic situation and that of the madman. Maybe there are other types of scenarios I'm not even considering and which, if I did properly consider them, would toss me into further doubt. So I want to reserve a small portion of my credence for the possibility that there is some skeptical scenario that I am overlooking or wrongly downplaying. Call this *wildcard skepticism*.¹⁶⁷

5. *Grounds for Doubt: Conclusion.*

A *radically skeptical scenario*, let's say, is any scenario in which a large chunk of the ordinary beliefs that we tend to take for granted is radically false. I offer the examples above as a gloss: If I'm dreaming or in a small simulation, or if I'm a doomed freak, or if I'm the brief creation of an indifferent deity, then I'm in a skeptical scenario. So also if I'm a recently envatted brain, or if I am the only entity that exists and the whole "external world" is a figment of my mind (Chapter 6), or if the laws of nature suddenly crumble and the future is radically unlike the past. Call the view that no such radically skeptical scenario obtains *non-skeptical realism*.

¹⁶⁷ Compare the "catch-all" hypothesis, or the "problem of unconceived alternatives" in formal theories of evidence Shimony 1970; Earman 1992; Stanford 2006; Wenmackers and Romeijn 2016.

I suggest that it's reasonable to have only a 99% to 99.9% credence in non-skeptical realism, plus or minus an order of magnitude, and thus about a 1% to 0.1% credence that some radically skeptical scenario obtains. This is the position I'm calling 1% skepticism.

I am unconcerned about exact numbers. Nor am I concerned if you reject numerical approaches to confidence. I intend the numbers for convenience only, to gesture toward a degree of confidence much higher than an indifferent shrug but also considerably lower than the nosebleed heights of nearly absolute certainty. If numbers suggest too much precision, I can make do with asserting that the falsity of non-skeptical realism is highly unlikely but not extremely unlikely – a low probability but not nearly as low as the probability of winning the state lottery with a single ticket.

I am also unconcerned about defending against complaints that my view is insufficiently skeptical. There are few genuine radical skeptics. I assume that sincere opposition will come almost exclusively from people much more confident than I in non-skeptical realism. In one sense, this isn't a skeptical position at all. I'm rather confident that non-skeptical realism is true. In conversation, some anti-skeptical philosophers have told me that 99%-99.9% credence in non-skeptical realism is entirely consistent with their anti-skeptical views. However, other anti-skeptical philosophers tell me that they think 0.1% to 1% is far too high a credence that some radically skeptical scenario obtains. The latter are my intended opponents.

You might object that skeptical possibilities are impossible to assess in any rigorous way, that the proposed credence numbers are unscientific, indefensible by statistics based on past measurements, insufficiently determinable – that they even to some extent undermine their own apparent basis (see the next section). Right! Yes! Good and true! Now what? Virtual certainty doesn't seem like the proper reaction. If you don't know how much certainty you should have,

slamming your credence up to 100% seems like exactly the wrong thing to do. Meta-doubt – doubt about proper extent of one’s doubt – is more friend than foe of the 1% skeptic.

Although my thesis concerns skepticism, the reader might notice that I have never once used the word “know” or its cognates in this chapter. Maybe I know some things that I believe with less than 99.9% credence.¹⁶⁸ Maybe I know that I’m awake despite lacking perfect credence in my wakefulness. Conversely, maybe I don’t know that I’m not a brain in a vat despite (let’s suppose) a rational credence in excess of 99.9999%. This chapter thus stands orthogonal to the bulk of the recent Anglophone philosophical writing on radical skepticism, which concerns knowledge rather than rational credence and typically doesn’t address the practical likelihood of particular skeptical scenarios.

If my reasoning in Chapters 1-3 is correct, every big-picture metaphysical/cosmological view is both bizarre and dubious. Radically skeptical scenarios are also bizarre and dubious; but that’s less of an objection to them than it might be if we weren’t already committed to accepting that something bizarre and dubious must be the case. Conversely, if the arguments in the present chapter are sound, they might facilitate investing nontrivial credence in other weird possibilities, such as idealism (Chapters 3 and 5) or group consciousness (Chapter 2). Thus, each of these chapters helps clear the path for the others.

6. Is 1% Skepticism Self-Defeating?

If I’m a Boltzmann brain, then my cosmological beliefs, including my evidence that I might be a Boltzmann brain, are not caused in the proper way by the scientific evidence. If I’m

¹⁶⁸ Vogel 1999; Hill and Schechter 2007; Hetherington 2016. Following Radford 1966, I argue in Myers-Schulz and Schwitzgebel 2013 that knowledge might not even require belief.

in a simulation, my evidence about the fundamental nature of the world, including about the nature of simulations, is dubiously grounded. If I'm dreaming, my seeming memories about the science of dreams and wakefulness might themselves be mere dream errors. The scenarios I've described can partly undermine the evidence that seems to support them.¹⁶⁹

Such apparently self-undermining evidence does not, however, defeat skepticism. In his classic "Apology for Raymond Sebond", Michel de Montaigne compares skeptical affirmation to rhubarb, which flushes everything else out of the digestive system and then itself last.¹⁷⁰

Suppose your only evidence about the outside world is through video feeds. You're confined to a windowless, doorless room of video monitors. Imagine you then discover video-feed evidence that the video feeds are unreliable. Maybe you see video footage that seems to portray you in your room several days ago, as viewed from a hidden camera, and then you see footage that seems to portray people laughing at you, saying "let's keep fooling that dope!" The detailed process of deception is then seemingly revealed in further footage. What would be the proper response to this string of evidence? It wouldn't be to retain high confidence in your video feeds on the grounds that the new video evidence is self-defeating – even though, of course, you are now acutely aware that this new evidence might itself be fake, part of a multi-layered joke. Barring some alternative explanation, the correct response would be uncertainty about your feeds, including uncertainty about the very evidence that drives that uncertainty.

In general it would be surprising if evidence that one is in a risky epistemic position bites its own tail, ouroboratically devouring itself to leave certainty in its place. If something in your

¹⁶⁹ For versions of the self-undermining argument, see Moore 1959; Maitzen 2010; Crawford 2013; Windt 2015.

¹⁷⁰ Montaigne 1580/1595/2003, p. 590. See also Diogenes Laertius 3rd c. BCE/1972 on Pyrrho.

experience suggests that you might be dreaming – maybe you seem to be levitating off your chair – it would presumably not be a sensible general policy to dismiss that evidence on grounds that you might be dreaming it. If something hints that you might be in a simulation, or if scientific cosmological evidence converges toward ever more bizarre and dubious options, the proper response is more doubt, not less.

What is plausibly self-defeating or epistemically unstable is high credence in *one specific skeptical scenario*. Assigning over 50% credence to being a freak observer based on seemingly excellent cosmological evidence probably is cognitively unstable.¹⁷¹ Evidence that this is a dream might be equally compatible with simulation skepticism or cosmological skepticism or madman skepticism. It might be hard to justify high credence in just one of these possibilities. But 1% skepticism is a different matter, since it tolerates a diversity of scenarios and trades only in low credences imprecisely specified.

Part Two: Practical Implications of 1% Skepticism

7. *Should I Try to Fly, on the Off-Chance That This Is a Dream-Body?*

Philosophers sometimes say that skepticism is unlivable and can have no permanent practical effect on those who attempt to endorse it. See the Greco quote in the epigraph, for example. Possibly this is Hume's view too, in Part One of his *Treatise*, where after a game of backgammon he finds that he can no longer take his skeptical doubts seriously.¹⁷² I disagree.

¹⁷¹ For a version of this argument, see Carroll 2010, 2021; for an argument that cognitive instability does not require ascribing a very low credence to unstable positions, see Kotzen 2021. See Rinard 2019 for a related objection to external world skepticism in general, which I believe I can avoid for a related reason: my relatively high credence in non-skeptical realism.

¹⁷² Hume 1740/1978, p. 269.

Like Sextus Empiricus (in the other epigraph quote), I think radical skepticism, including 1% skepticism, need not be behaviorally inert. To explore this idea, I will use examples from my own life, which I have chosen because they are real, as best I can recall them – and thus hopefully realistic – but you may interpret them as hypothetical if you prefer.

I begin with a whimsical case. I was revising the dreaming section of the original essay on which this chapter is based. Classes had just been released for winter break, and I was walking to the science library to borrow more books on dreaming. I had just been reading Evan Thompson on false awakenings.¹⁷³ There, in the wide-open empty path through east campus, I spread my arms, looked at the sky, and added a leap to one of my steps, in an attempt to fly.

My thinking was this: I was almost certainly awake – but only *almost* certainly. At that moment, my credence that I was dreaming was higher than usual for me, maybe around 0.3% (though I didn't conceptualize it numerically at the time). I figured that *if* I was dreaming, it would be thrilling to fly around instead of trudging. On the other hand, if I was not dreaming, it seemed no big deal to leap, and in fact kind of fun – perhaps a bit embarrassing if someone saw me, but no one seemed to be around.

I'll model this thinking with a simple decision theoretic matrix. I don't intend the numbers to be precise, nor do I mean to imply that I was at the time in an especially mathematical frame of mind.

Call dream flying a gain of 100 units of pleasure or benefit or utility, waking leap-and-fail a loss of 0.1 units, continuing to walk in the dream a loss of 1 (since why bother with the trip if it's just a dream), and dreaming leap-and-fail a loss of 1.05 (the loss of the walk plus a little more for the leap-and-fail disappointment) – all relative to a default of zero for walking, awake,

¹⁷³ Thompson 2015, ch. 6.

to the library. For simplicity, I'll assume that if I'm dreaming, things are no overall better or worse than if I'm awake (for example, I can get the books tomorrow). Let's assign a 50% likelihood of successfully flying, conditional upon its being a dream, since I don't always succeed when I try to fly in my dreams.

Here's the payoff matrix:

	dreaming (0.3%)	not dreaming (99.7%)
leap	50% fly: +100 50% not fly: -1.05	-0.1
don't leap	-1	0

The expected value formula, which sums the utilities times the probabilities for each outcome, yields $(0.3\%)(50\%)(100) + (0.3\%)(50\%)(-1.05) + (99.7\%)(-0.1) = +0.05$ units as the expected gain for leaping and $(0.3\%)(-1) + (99.7\%)(0) = -0.003$ units as an expected loss for not leaping. Decision theoretically, given the delightful prospect of flying and the small loss for trying and failing, it made sense to try to fly, even though I thought failure very likely (>99%). Kind of like scratching a free lotto ticket on the off chance.

It was an unusual occasion. Normally more is lost for trying to fly (e.g., embarrassment or distraction from more important tasks). Normally my credence in the dream possibility is lower than 0.3%. And I'm assuming a high payoff for flying. Also, the decision is not repeatable. If I even once try to fly and fail, that will presumably influence my sense of the various probabilities and payoffs. For example, my first leap-and-fail should presumably reduce my credence that I am dreaming and/or my credence that if this is a dream I can fly.

You might say that if the effect of skeptical reflections is to make one attempt to fly across campus, that is a bad result! However, this is a poor objection if it assumes hindsight certainty that I was not dreaming. A similar argument could be mounted against having bought car insurance. A better version of this objection considers the value or disvalue of the type of psychological state induced by 1% skepticism, conditional upon the acknowledged 99%-99.9% subjective probability of non-skeptical realism. If skeptical doubt sufficiently impairs my approach to what I am 99%-99.9% confident is my real life, then there would be pragmatic reason to reject it.

I'm not sure that in this particular case it played out so badly. The leap-and-fail was silly, but also whimsically fun. I briefly punctured my usual professional mien of self-serious confidence. I would not have leapt in that way, for that reason, adding that particular weird color to my day, and now to my recollections of that winter break, if I hadn't been dwelling so much on dream skepticism.¹⁷⁴

8. *Weeding or Borges?*

It was Sunday and my wife and children were at temple. I was sitting at my backyard table, in the shade, drinking tea – a beautiful spring day. Borges' *Labyrinths*, a favorite book, lay open before me. But then I noticed Bermuda grass sprouting amid the daisies. I hadn't done any weeding in several weeks. Should I finally stop procrastinating that chore? Or should I seize current pleasure and continue to defer the weeds? As I recall it, I was right on the cusp between

¹⁷⁴ Maybe if I incorporate such an expectation in the original decision, the value in the not-dreaming/leap cell would become positive. It's not clear, however, that standard decision theory can non-paradoxically incorporate outcomes that reflect one's happiness with the decision process itself.

these two options – and rationally so, let’s assume. I remember this not as a case of weakness of will but rather as a case of rational equipoise.

[Illustration 4 (Caption: Weeding or Borges?): I am in a suburban California backyard on a beautiful day, sitting at a small round table with a cup of tea and a copy of Borges’ *Labyrinths* in my hand but not yet open. My house is behind me, and I am gazing with narrowed brow at a large patch of daisies in front of me. A suburban fence is in the background. The patch of daisies is obviously beset by weeds but not overcome with weeds. In one corner, the scene is fading away to 1s and 0s or alternatively the corner is folded back to reveal 1s and 0s beneath.]

Suddenly, skeptical possibilities came to mind. This might be a dream, or a short-term simulation, or some other sort of brief world, or I might in some other way be radically mistaken about my position in the universe. Weeding might not have the long-term benefits I hoped. I might wake and find that the weeds still needed plucking. I might soon be unmade, I and the weeds dissolving into chaos. None of this I regarded as *likely*. But I figured that if my decision situation, before considering skepticism, was one of almost exact indifference between the options, then this new skeptical consideration should ever-so-slightly tilt me toward the option with more short-term benefit. I lifted the Borges and enjoyed my tea.

Before considering the skeptical scenarios, my decision situation might have been something like this:

Borges: short-term expected value 2, long-term expected value 1, total expected value 3.

Weeding: short-term expected value -1, long-term expected value 4, total expected value 3.

With 1% skepticism, the decision situation shifts to something like this:

	some skeptical scenario obtains (1%)	the world is approximately how I think it is (99%)
Borges	short-term gain: 2 long-term gain: - 50% likely to be 1 (e.g., if the sim endures a while) - 50% likely to be 0	short-term gain: 2 long-term gain: 1
weeding	short-term gain: -1 long-term gain: - 50% likely to be 4 - 50% likely to be 0	short-term gain: -1 long-term gain: 4

Multiplying through, this yields an expectation of +2.995 for Borges and +2.980 for weeding.

Of course I'm simplifying. If I'm a Boltzmann brain, I probably won't get through even one sentence of Borges. If I'm in a simulation, there's some remote chance that the Player will bestow massive pleasure on me as soon as I start weeding. Et cetera. But the underlying thought is, I believe, plausible: To the extent radically skeptical scenarios render the future highly uncertain, they tilt the decisional scales slightly toward short-term benefits over long-term benefits that require short-term sacrifices.

It has been gently suggested to me that my backyard reasoning was just post-hoc rationalization – that I was already set on choosing the Borges and just fishing for some justifying excuse. Maybe so! This tosses us back into the pragmatic question about the psychological effects of 1% skepticism. Would it be a good thing, psychologically, for me to have some convenient gentle rationalization for a dollop more carpe diem? In my own case, probably so. Maybe this is true of you too, if you're the type to read this deep into a book of academic philosophy?

9. *How to Disregard Extremely Remote Possibilities.*

If radical skepticism is on the table, by what rights do I simplify away from extremely remote possibilities? Maybe it's reasonable to allow a $1/10^{50}$ credence in the existence of a Player who gives me at least 10^{50} lifetimes' worth of pleasure if I choose to weed in the Borges scenario. Might my decision whether to weed then be entirely driven by that remote possibility?

I think not, for three reasons.

First, *symmetry*. My credences about such extremely remote possibilities appear to be approximately symmetrical and canceling. In general, I'm not inclined to think that my prospects will be particularly better or worse due to their influence on extremely unlikely deities, considered as a group, if I pull weeds or not. More specifically, I can imagine a variety of unlikely deities who punish and reward actions in complementary ways – one punishing what the other rewards and vice versa. (Similarly for other remote possibilities of huge benefit or suffering, e.g., rising to a 10^{100} -year Elysium if I step right rather than left.) This indifference among the specifics is partly guided by my general sense that extremely remote possibilities of this sort don't greatly diminish or enhance the expected value of such actions. Since it's generally reasonable to reconcile hazy and derivative credences to which one hasn't given much previous thought (e.g., the probability of extremely unlikely right-foot-loving deities) so that they align with credences which one feels more secure and comfortable assessing (e.g., the probability that the long-term expected outcomes are about the same if I step right versus left), I

can use my confidence in the latter to drive my symmetrical assignment to the former, barring the discovery of grounds for asymmetric assignment.¹⁷⁵

Second, *diminishing returns*. Bernard Williams famously argued that extreme longevity would be a tedious thing.¹⁷⁶ I tend to agree instead with John Martin Fischer that extreme longevity needn't be so bad.¹⁷⁷ However, it's by no means clear that 10²⁰ years of bliss is 10²⁰ times more valuable than a single year of bliss. (Similarly for bad outcomes and extreme but instantaneous outcomes.) Value might be very far from proportional to temporal bliss-extension for such magnitudes.¹⁷⁸ And as long as one's credence in remote outcomes declines sharply enough to offset any increasing value in the outcomes, extremely remote possibilities will be decision-theoretically negligible.

Third, *loss aversion*. I'm loss averse: I'll take a bit of a risk to avoid a sure or almost-sure loss (even at some cost to my overall decision-theoretically calculated expected utility if that utility is calculated without consideration of the displeasure of loss).¹⁷⁹ And my life as I think it is, given non-skeptical realism, is the reference point from which I determine what counts as a loss. If I somehow arrived at a 1/10⁵⁰ credence in a deity who would bestow 10⁵⁰ lifetimes of pleasure if I avoided chocolate for the rest of my life (or alternatively a deity who would damn

¹⁷⁵ This last sentence is intended to address Monton's 2019 critique of my symmetry argument as it appeared in Schwitzgebel 2017a.

¹⁷⁶ Williams 1973.

¹⁷⁷ Fischer 1994; Fischer and Mitchell-Yellin 2014. For my specific response to Williams and Fischer, see the "Goldfish Pool" example in Schwitzgebel 2019c, ch. 44 (also fictionally portrayed in Schwitzgebel 2016a).

¹⁷⁸ For example, truly radical life extension seems to require either repeating experiences, which might be good but not additively choiceworthy, or living out ever more divergent experiences, which raises questions about personal identity and species identity, which could undermine the prudential (self-interested) calculus. (See the Goldfish Pool case referenced in note 37.)

¹⁷⁹ The classic treatment is Kahneman and Tversky 1979, which spawned an immense industry of subsequent research.

me to 10^{50} lifetimes of pain if I didn't avoid chocolate), and if there were no countervailing considerations of symmetrical chocolate-rewarding deities, then on a risk-neutral decision-theoretic function it might be rational for me to forego chocolate evermore. But foregoing chocolate would be a loss relative to my reference point; and since I'm loss averse rather than risk-neutral, I might be willing to forego the possible gain (or risk the further loss) so as to avoid that almost certain loss of life-long chocolate pleasure. Similarly, I might reasonably decline a gamble with a 99.99999% chance of death and a 0.00001% chance of 10^{100} lifetimes' worth of pleasure, even bracketing diminishing returns. I might even decide that at some level of improbability – 10^{30} ? – no finite positive or negative outcome could lead me to take a substantial almost-certain loss. And if the time and cognitive effort of sweating over decisions of this sort itself counts as a sufficient loss, then I can simply disregard any possibility where my credence is below that threshold.

These considerations synergize: the more symmetry and the more diminishing returns, the easier it is for loss inversion to inspire disregard. Decisions at credence $1/10^{50}$ are one thing, decisions at credence $1/10^3$ quite another.¹⁸⁰

¹⁸⁰ Monton 2019 provides a helpful review of arguments for “Nicolausian discounting” on which probabilities below a certain threshold are rationally ignored. See also discussions of Pascal’s wager, especially finite-valued versions of it: Pascal 1670/2005; Hájek 2003; Bostrom 2009; Balfour 2021. In reply to my loss-aversion argument, Monton imagines “Bernie” who manages my money by transferring it from Bank A to Bank B for the sole reason that there’s a $1/10^{31}$ chance that he will get all of the money. Monton and I agree that the $1/10^{31}$ chance can rationally be ignored (unless it is iterated, in which case the iterations should be evaluated as a batch). Contra Monton, I hold that loss aversion can deliver the desired result, when coupled with recognition of the costs of cognition. If I spend my time worrying about possible $1/10^{31}$ scams, that’s a sure loss of time and cognitive energy, relative to my reference point of blissfully ignoring massively improbable scams, and I will happily take a small risk of an arbitrarily terrible outcome (e.g., a $1/10^{31}$ risk) to avoid that sure loss. I’d rather continue along my reference point of blissfully ignoring massively improbable scams. Of course, now that Monton has ratted Bernie out, I’ve already paid the cost of thinking about it, so I might as well halt the scam if I can do so cost-free.

10. Descending Through the Fog.

I'm a passenger in a jumbo jet descending through turbulent night fog into New York City. I'm not usually nervous about flying, but the turbulence has me on edge. I estimate the odds of dying in a jet crash with a major U.S. airline to be small – but descent in difficult weather is one of the riskiest situations in flight. So maybe my odds of death in the next few minutes are about one in a million? I can't say those are odds I'm entirely comfortable disregarding. One in a million is on my decision-theoretic radar in a way one in 10^{50} is not.

But then I think: Maybe I'm a short-term sim! Based on the reasoning in Section 2, my credence that I'm a short-term sim should be about one in a thousand. In a substantial proportion of those short-term scenarios, my life will end soon. So my credence that I die today because I'm a short-term sim should be at least an order of magnitude higher than my credence that I die today in a mundane plane crash.

Should I be alarmed or comforted by these reflections?

Maybe I should be alarmed. Once I begin to consider simulation scenarios, my estimated odds of near-term death rise substantially. On the other hand, I'm accustomed to facing the simulation possibility with equanimity. The odds are lowish, and I don't see much to do about it. The descent of the airplane adds only slightly to the odds of my near-term death, on this way of thinking – within rounding error – and in an airplane, as in a simulation, no action seems likely

Monton's alternative YOLO principle is interesting but (a.) unnecessary if the arguments of this section work, and (b.) appears to require adding a new moving part to decision theory: a difference in value between an outcome conceptualized as actual versus as an unrealized possibility. It's unclear how this would work if applied beyond the case of Nicolausian discounting. Of course it's better if good outcomes actually occur, but apart from the obvious sense in which that is true, does winning \$100 somehow have a different value if it's actual than if it's non-actual?

to improve my chances. I should just sit tight. The airplane's turbulent descent has made no material change in my prospects or options.

Such was my thinking several years ago, as I was about to land at La Guardia to give a series of talks on skepticism and consciousness in the New York area. At the time, as I now seem to recall it, I was neither alarmed nor much comforted by these reflections. The chance of imminent death given the simulation possibility felt less emotionally real than what I would have said were the much smaller odds of death in a mundane airplane crash.

It's such a common pattern in our lives – reaching a conclusion based on theoretical reasoning and then failing to be moved by it at a gut level. The Stoic sincerely judges that death is not bad but quakes in fear on the battlefield. The implicit racist judges that her Black students are every bit as capable as her White and Asian students but her habitual reactions and intuitive assessments fail to change.¹⁸¹ Relatedly, academic philosophers might defend their positions passionately at conferences, staking their career and reputation in all sincerity, but feel reluctant to fully endorse those positions in a non-academic context.¹⁸² I'm trying to carry my 1% skepticism out of its academic context. That works nicely when the act is safe, optional, low-stakes, playful – an arm-spreading leap, a morning rereading Borges. But this is something different.

Some philosophical traditions recommend “spiritual exercises” aimed to align one's spontaneous emotional reactions with one's philosophical opinions – for example, in the Stoic

¹⁸¹ Philosophical discussions include Gendler 2008; Frankish 2010; Schwitzgebel 2010; Brownstein 2015/2019; Brownstein and Saul 2016. For a psychology-oriented review, see Gawronski and Brannon 2017.

¹⁸² Goldberg 2013; Lycan 2013; DeRose 2017; Barnett 2019.

tradition, vividly imagining death with equanimity.¹⁸³ Maybe I should do something similar, if I really wish to live as a 1% skeptic? One concern: Any realistic program of spiritual exercises would presumably require substantial time and world stability to succeed, so it's likely to fail unless non-skeptical realism is true; and there's something semi-paradoxical about launching a program that will probably succeed in lowering my credence in non-skeptical realism only if non-skeptical realism is true.

11. A Defense of Agnosticism.

I used to say that my confidence in the non-existence of God was about the same as my confidence that my car was still parked where I left it rather than stolen, borrowed, or towed – a credence of maybe 99.9%, given the safety of my usual parking spots. This credence was sufficient, I thought, to warrant the label atheist. Reflection on skeptical possibilities has converted me to agnosticism.

Cosmological skepticism leaves plenty of room for the possibility of a god or gods – 50%? 10%? – conditional upon accepting it. That puts the 1% cosmological skeptic already in the ballpark of 0.1% credence on those grounds alone. Furthermore, if I'm a sim, the power that the managers of the simulation likely have over me – the power to stop the sim, erase or radically alter me, resurrect me, work miracles – are sufficient that I should probably regard them as gods.¹⁸⁴ And of course plenty of non-skeptical scenarios involve a god or gods (some non-skeptical simulation scenarios, some non-skeptical Big Bang cosmologies). It would be odd to assign a 1% credence to the disjunction of all radically skeptical scenarios, and some substantial

¹⁸³ See Hadot 1995; Stalnaker 2006; Ganeri 2013. Special thanks to my former student Ted Preston whose 2005 PhD dissertation is a cross-cultural exploration of the topic.

¹⁸⁴ I defend this claim further in Schwitzgebel 2019c, ch. 21.

sub-portion of that credence to skeptical scenarios involving gods, while assigning only negligible credence to otherwise similar non-skeptical scenarios involving gods. Furthermore, if I'm going to take substance dualism and metaphysical idealism as seriously as I think I should, after the reasoning in Chapter 3 – rather than just sitting comfortably confident in my default inclination toward scientific materialism – versions of those metaphysical approaches often imply or rely on a god. If the cosmos is weird, it might just be weird enough to be run by one or more deities – which, of course, lots of thoughtful, well-informed people are inclined to think for their own very different reasons. Once I embrace the arguments of this chapter and the last, I can no longer sustain my formerly high level of atheistic self-assurance.

12. *"I Think There's About a 99.8% Chance You Exist"*.

Alone in my backyard or walking across an empty campus, it can seem quite reasonable to me to reserve a 0.1% to 1% credence in the disjunction of all radically skeptical scenarios, and conditionally upon that to have about a 10% or 50% credence in the non-existence of the people whose existence I normally take for granted – my family, my readers, the audience at a talk.

But now I consider my situation before an actual live audience. Can I say to them, sincerely, that I doubt their existence – that I think there's a small chance that I'm dreaming right now, that I think there's a small chance they might merely be mock-up sprites, mere visual input in a small me-centered simulation, lacking real conscious experience? This seems, somehow, even weirder than the run-of-the-mill weirdness of dream skepticism in solitary moments. In fact, I have spoken these words to actual live audiences, and I can testify that it feels strange!

I tried it on my teenage son. He had been hearing, from time to time, my arguments for 1% skepticism. One day several years ago, driving him to high school, a propos of nothing I

said, “I’m *almost* certain you exist.” A joke, of course. How could he have heard it, or how could I have meant it, in any other way?

One possible source of awkwardness is this: My audience would be fully aware that they aren’t mere mock-up sprites, just as I would also invest much less than a 0.1% credence in my being a mindless mock-up sprite. Paraphrasing Descartes, “I think, therefore I am not a mindless mock-up sprite”. Maybe this is even something of which I can be absolutely certain.¹⁸⁵ So it’s tempting to say that the audience would see that my doubts are misplaced.

But in non-skeptical cases, we can view people as reasonable in having substantial credences in possibilities we confidently dismiss, if we recognize an informational asymmetry. The blackjack dealer who sees a 20 doesn’t think the player a fool for standing on a 19. Even if the dealer sincerely tells the player she has a 20, she might think the player reasonable to confess doubt about her truthfulness. So why do radically skeptical cases seem different?

One possible clue is this: It doesn’t feel wrong in quite the same way to say “I think that that *we all* might be part of a short-term simulation”. Being *together* in skeptical doubt seems fine, and in the right context even friendly and fun. Maybe, then, the discomfort arises from an implicitly communicated lack of respect – a failure to treat one’s interlocutor as an equal partner metaphysically, epistemically, ethically? There’s something offensive, perhaps, or inegalitarian or silencing about saying “I’m certain that *I* exist, but I have some doubts about whether you do”.

I feel the problem most keenly around people I love. I can’t doubt that we’re in the world together. It seems wrong – merely a pose, possibly an offensive pose – to say to my dying

¹⁸⁵ In his *Discourse of Method* (1637/1985), Descartes famously says “I think, therefore I am”, but in his later *Meditations on First Philosophy* (1641/1984) he omits the “therefore”, holding that each of “I think” and “I am” are independently certain.

father, in seeming sincerity at the end of a philosophical discussion about death and God and doubt, “I think there’s a 99.8% chance that you exist”. I throws a wall between us. At least I felt that it did, despite my father’s intellectual sophistication, the one time I tried it. He forgave me, I think, more than I have been able to forgive myself.

Can friend-doubting be done a different way? Maybe I could say, “For these reasons, *you* should doubt *me*. And I will doubt you too, just a tiny bit, so that we are doubting together. Very likely, the world exists just as we think it does; or even if it doesn’t, even if nothing exists beyond this room, still I am more sure of you than of anything else”.

Even with this gentler, more egalitarian approach, the skeptical decision calculus might rationally justify a small tilt toward selfish goods over sacrifice for others who might not exist. To illustrate with a simplified model: If, before skeptical reflection I am indifferent between X pleasure for myself versus Y pleasure for you, reducing my credence in your existence to 99.8% might shift me to indifference between $.998 * X$ pleasure for myself versus Y pleasure for you. Is that where 1% skepticism leads?

Maybe not. Selfish choices are often long-term choices, such as for money or career. Kind, charitable engagement is often more pleasant short-term. Also, doubt about the reality of someone you are interacting with might be more difficult to justify than doubt about the future. So if pre-skepticism I am indifferent between Option A with gains for me of X_1 short-term plus X_2 long-term versus Option B with gains for you of Y_1 short-term plus Y_2 long-term, then embracing a 99.8% credence in your existence and a 99.5% credence in the future, I might model the choice as Option A = $X_1 + .995 * X_2$ versus Option B = $.998 * Y_1 + (\text{approx.}) .993 * Y_2$, rationalizing a discount of my long-term gains for your short-term benefit.

In what direction does skeptical doubt tend to move one's character? I'm aware of no direct empirical evidence. But I think of the great humane skeptics in the history of philosophy, especially Zhuangzi and Montaigne. The moral character that shines through their works seems unharmed by their doubts.

If we're evaluating 1% skepticism in part for its psychological effects conditional on the assumption that non-skeptical realism is correct, then there is a risk here, a risk that I will doubt others selfishly or disrespectfully, alienating myself from them. But I expect this risk can be managed. Maybe, even, it can be reversed. In confessing my skepticism to you, I make myself vulnerable. I show you my weird, nerdy doubts, which you might laugh at, or dismiss, or join me in. If you join me, or even just engage me seriously, we will have connected in a way we treasure.

#

I dedicate this chapter to the memory of my father.

The Weirdness of the World

Chapter Five Kant Meets Cyberpunk

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

First, spatial properties depend on our minds. Objects and events, as they are in themselves, independently of us, do not have spatial properties. Things appear to be laid out in space, but that's only because our perceptual faculties necessarily construe 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 features of things as they are in themselves, independently of us, are unknowable to us. We can't achieve positive¹⁸⁶ knowledge of things as they are in themselves through our empirical science, which is limited by being rooted in and contingent upon our perceptual construal of objects as laid out in space. Nor can we achieve positive knowledge of things as they are in themselves by any means that purport to transcend empirical science, such as *a priori* mathematical reasoning, innate insight, or religious revelation.

Transcendental idealism, famously associated with Immanuel Kant, is a historically important alternative to materialism. In Chapter 3, I classed it among the grab bag of “compromise/rejection views” that don't fit neatly into the standard taxonomy. The view is in a

¹⁸⁶ Perhaps we can achieve “negative” knowledge of some sort about things in themselves, such as knowledge that they lack some specific feature or other. For example, if we know that spatiality depends on us, then perhaps we know, negatively, that things are non-spatial when considered as they are in themselves. It's tricky to assess the extent to which negative knowledge might be possible in a transcendental idealist framework. I set the issue aside in this chapter.

sense *idealist* because it treats all spatial (and maybe also temporal and causal) properties of objects as dependent on our minds. However, unlike metaphysical idealism (simply called “idealism” in Chapter 3) transcendental idealism doesn’t commit to a metaphysical picture on which everything that exists is fundamentally mental. Neither is transcendental idealism a materialist or dualist view in the senses of Chapter 3. Notably, contrary to materialism, transcendental idealism denies that the most fundamental properties of things are the types of properties revealed by the physical sciences.¹⁸⁷

Transcendental idealism used to be big. From the late 18th century through the early 20th century, European and North American metaphysicians typically positioned themselves relative to Kant. With the rise of materialism to dominance in the mid-20th century, however, transcendental idealism came to be viewed as mostly a historical relic. Few mainstream Anglophone philosophers or cosmologists would now call themselves transcendental idealists, or indeed idealists of any stripe. I’m rooting for a modest revival. Transcendental idealism should be among the live metaphysical and cosmological options that philosophers take seriously. It deserves a non-trivial portion of our credence. Highlighting and defending its feasibility as a minority position thus fits within the project of exploring the landscape of bizarre but dubious possibilities described in Chapter 3.

In this chapter, 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

¹⁸⁷ As noted in Chapter 3, the exact definition of materialism or “physicalism” is a tricky matter. The criterion here (that the most fundamental properties of things are the types of properties revealed by the physical sciences) is intended only as a necessary condition of materialism such that denying it is sufficient for denying materialism. However, even as a necessary condition it fails, if materialism 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 for the argumentative purposes of this chapter.

true. I will do this by developing an idea popularized by the “cyberpunk” movement in science fiction and which I introduced as part of a skeptical argument in Chapter 4: the idea that we might be living in a computer simulation. The central claim is this. If we are living inside a computer simulation, the fundamental nature of reality might be unknowable to us and very different than we normally suppose. It might be, for example, non-spatial. If we are living inside a computer simulation, spatiality might merely be the way that a fundamentally non-spatial reality is experienced by our minds. Once we grasp the specific though presumably unlikely possibility that we could be living in a computer simulation implemented by a non-spatial system, we can better understand the potential attractiveness of transcendental idealism in its more general form.

This chapter serves four functions in relation to previous chapters. (1.) To the extent they support transcendental idealism, the arguments of this chapter support the Universal Dubiety thesis of Chapter 3: Reason to treat transcendental idealism as a live possibility is reason to have some doubts about other possibilities incompatible with it. (2.) This chapter adds more flesh to the simulation hypothesis as discussed in Chapter 4, illustrating in more detail how the simulation hypothesis might work. (3.) As I will argue near the end of this chapter, transcendental idealism provides grounds for cosmological doubt, further supporting the 1% skepticism thesis of Chapter 4. (4.) Finally, if transcendental idealism is true, the world is bizarre. This claim can be conjoined with claim that materialism is bizarre, dualism is bizarre, and metaphysical idealism is bizarre, to help support the Universal Bizarreness thesis of Chapter 3.

1. Kant, Kant, and Transcendental Idealism.*

According to 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.¹⁸⁸ It is “transcendentally ideal” in the sense that it has no existence independently of our possible experience.¹⁸⁹ 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.¹⁹⁰

Notoriously, these claims invite diverse interpretation. 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, we can treat it as the position of a merely possible philosopher *Kant**.¹⁹¹

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; they lack spatial dimensions like length, breadth, and depth; and they are not extended across spatial or spatiotemporal regions. Spatiality is instead 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 produces 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 can’t

¹⁸⁸ Kant 1781/1787/1998, A26/B42, p. 176-177.

¹⁸⁹ Kant 1781/1787/1998, A28/B44, p. 177.

¹⁹⁰ Kant 1781/1787/1998, A27/B43, p. 177.

¹⁹¹ See Stang 2016 for discussion of the range of recent interpretations of Kant’s metaphysics. I am broadly sympathetic with Allais’s (2015) reading, which draws on recent work in the 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’s Kant. In Chapter 9, I will further explore the connections between transcendental idealism and empirically informed philosophy of perception.

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 such “completely inadequate examples” as colors, taste, and so on, but I believe it can be so illustrated, as long as we are careful not to draw too much from the illustration.¹⁹² 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 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 about *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 concerns the object’s tendency to evoke certain experiences in us.¹⁹³

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

¹⁹² Kant 1781/1787/1998, A29/B45, p. 178. But Kant himself illustrates the view by analogy to secondary qualities in his *Prolegomena* (1783/2004, 4:289, p. 40-41). See also Putnam 1981 and Allais 2015, contra Van Cleve 1995.

¹⁹³ 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. I hope that other not-too-distant views of the nature of tastes and colors could also serve for the present argument.

abstract.) A closer analogy might be having a location somewhere in the manifold of possible tastes (where one possible location might be sweetness 5, sourness 2, saltiness 3, bitterness 0, umami 0¹⁹⁴). 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 bring sweetness 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.

2. *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

¹⁹⁴ As you might expect, this is a simplified characterization of the complex and contentious science of taste. See Erickson 2008; Chen, Gabitto, Pen, Ryba, and Zuker 2011; Beauchamp 2019.

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 this is their situation.

I will now introduce several concepts.

Following David Chalmers, 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’”.¹⁹⁵ 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 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, this book as being a certain

¹⁹⁵ Chalmers 2017, p. 312. The other definitions in this paragraph are also adapted from Chalmers. 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, p. 489, note 2). In reply to the article on which this chapter is based, Chalmers suggests that the view I attribute to Kant might in fact be closer to Schelling's view (Chalmers 2019a, p. 482; see Schelling 1800/1978: “The objective world is simply the original, as yet unconscious, poetry of the spirit”, p. 12). This might be correct if we interpret reality as fundamentally mental (as illustrated by the case of Angel below) but not if we commit only to the weaker claim, as is my intent, that reality is fundamentally non-material, with mentality serving only as an example of the most familiar way of being non-material.

distance from you, 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 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 red-pill out of the Matrix will your empirical and biological bodies be doing the same things. Note that *empirical* is a relational concept. What counts as an empirical object will be different for different people. 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 warehouse. 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 can't travel toward it or away from it using your empirical body. You can't 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's an intuitive sense in which that still-unseen 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 an intuitive sense in which I am right: A treasure chest really is behind that door. Exactly how to make sense of unperceived empirical objects has troubled 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 correct to say that a treasure chest really is 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.¹⁹⁶

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.

3. How to Be a Sim, Fundamentality, and the Noumenal.

As discussed in Chapter 4, Nick Bostrom and several other philosophers have famously argued that we might be *sims* – that is, artificial intelligences within a shared virtual reality coordinated by a computer or network of computers.¹⁹⁷ The crucial difference between this

¹⁹⁶ See for example Berkeley 1710/1965 and Mill 1867.

¹⁹⁷ Bostrom 2003; Chalmers 2003/2010, forthcoming; Steinhart 2014.

scenario and the virtual reality scenarios of Section 2 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 Isaac Asimov's robots or the android Data from *Star Trek*. Whether this is in fact possible or realistic is unclear, depending on the resolution of the various debates about metaphysics and consciousness about which I express doubt throughout this book. In light of such doubts it is, I submit, reasonable to maintain an intermediate credence in the possibility of conscious robots – somewhere between, say, 5% and 95%. The next few sections are conditional upon a non-zero credence in the possibility of artificial computational systems with human-like conscious experiences.

Imagine, then, a conscious robot. Now imagine that it “jacks in” to cyberspace – 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 conscious subject (that is, the experiencer or the self) and its empirical reality would be wholly instantiated in computers. This would be one way to become a sim.

Alternatively, consider the computer game *The Sims*. In this game, artificial people stroll around conducting 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 writers and futurists.¹⁹⁸ 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 conscious subjects and all of their transactions with their simulated environments. At the simulated level are the conscious 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 events in the computer “ground” events in

¹⁹⁸ For example, Egan 1994, 1997; Kurzweil 2005; Chalmers 2010b.

the simulation, while events in the simulation do not similarly ground events in the computer; or 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 as they appear laid out in the spatial manifold of the simulation. The sims might or might not understand that undergirding these phenomena is some more fundamental “noumenal” structure which is not for them a possible object of perception and which remains beyond their access and comprehension.¹⁹⁹

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 its operators occupy 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.

4. Immaterial Computation.

¹⁹⁹ Contra Kant, it might be only contingently the case that the base-level reality is not a possible object of the sims’ perception, but let’s bracket that issue for now.

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 type of “Turing machine” – that, given enough time, has the same ability to solve computational problems as the most powerful supercomputer we can imagine.²⁰⁰

Hilary Putnam remarked that there is nothing about computation that requires it to be implemented in a material substance.²⁰¹ We might, in theory, build a computer out of ectoplasm, out of immaterial soul-stuff. For concreteness, let’s consider an immaterial soul of the sort postulated by Descartes.²⁰² It is capable of thought and conscious experience. It exists in time, and it has causal powers. However, it has no spatial properties such as extension or spatial position. To give it full power, let’s assume that the soul has perfect memory. This needn’t be a *human* soul. Let’s call it Angel.²⁰³

²⁰⁰ Okay, well, it partly depends on whether you can imagine an infinitely accelerating supercomputer or one that performs infinitely many parallel tasks.

²⁰¹ Putnam 1965, p. 43-44.

²⁰² Descartes 1641/1984; 1647/1985.

²⁰³ 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 also possible in souls without bodies, but if necessary we could change to more purely intellectual examples, such as mathematical thoughts. I am also bracketing Descartes’s 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 doing otherwise and if the computational description would have to incorporate that possibility.

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 this stage of the argument. In principle, could a Turing-type computer 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;
- 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 suppose 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 adds or subtracts 1 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 1 to his running total and recalls the pitch that is associated with the updated running total. Instead of moving left, Angel subtracts 1. 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 (i.e., C₄, or C in the fourth octave) is assumed to infinity.

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

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.	

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 1 from his running total. We rely on Angel’s memory to implement the writing and reading: To “write” G_4 when his running total is +2 is to commit to memory the idea that the next time the running count is +2 he will “read” – that is, 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 finite 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 (perhaps non-existent) entity. If there’s something incoherent or otherwise objectionable about Cartesian souls, then imagine, if possible, any entity or process (1.) whose existence is disallowed by materialism (it needn’t be inherently mental) 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.)

5. 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, lasers, 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 enough 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. 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, or 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.²⁰⁴

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, or an ordinary early 21st-century computer, or a 1940s ENIAC 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 these 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 and how deep the modeling of environmental detail²⁰⁵). 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

²⁰⁴ See for example Putnam 1965 on functionalism and probabilistic automata and Chalmers 1996 on the principle of organizational invariance. Compare also with parallel considerations in Chapter 2 on why it's more natural for a materialist to be liberal than restrictive about the types of internal structures that give rise to consciousness if the relevant patterns of outward behavior are present.

²⁰⁵ A budget simulation could save a lot of computational cycles by leaving most environmental details blank until they become relevant to the cognitive processing of a conscious subject, then adding the detail just in time – or even retroactively, while also editing the subject's memory accordingly. This is why I don't accept Dennett's argument (1991, p. 6) against skeptical possibilities that require complex informational modeling of the environment. In fact, Dennett's Stalinesque/Orwellian "multiple drafts" view of consciousness and memory (1991, chapter 5) is very convenient for the skeptic (or transcendental idealist) on exactly this point.

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 could arise from something so simple. But this is where commitment to flexibility about implementation leads us. The bizarreness of this idea is one reason to have some qualms about the assumptions that led us to it; but as I argued in Chapter 3, *all* general theories of consciousness have some highly bizarre consequences, so bizarreness can't in general be an insurmountable objection.

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

6. 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 non-spatial and non-material – something beyond empirical inquiry. Part of the challenge in recognizing the viability of transcendental idealism as a competitor to materialism, I believe, is that the position sounds so vague and mystical that it's difficult to conceive what it might amount to or how it could even possibly be true.

Here is 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 through the tools of

empirical science or by any other means. 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 analogous to the way that (as described in Section 1) our taste experiences help us navigate the edible world. If the world is like that, then transcendental idealism is correct and materialism is false.

We might have excellent *empirical* evidence that everything is material. We might even imagine excellent empirical evidence that consciousness can only occur in entities with brains that have a certain specific biological structure (contra Chapter 2). But all such evidence is consistent with things being very different at a more fundamental level. Artificial intelligences in a virtual 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 our relation to the cosmos is like that of a flea on the back of a dog, watching a hair grow (Chapter 3) – if our perspective is a tiny, contingent, limited slice of a vast and possibly infinite cosmos – then we should acknowledge the possibility that we occupy some bubble or middle layer or weirdly constructed corner, and beneath our empirical reality lies something very different than we ordinarily suppose.

I have articulated a possible transcendental idealism about space. But Kant himself was more radical. He argued that time and causation are also transcendently ideal, not features of things as they are in themselves independently of us. Given the tight relationship between time and space in current physical theory, especially relativity theory, it might be difficult to sustain the transcendental ideality of space without generalizing to time. Furthermore, temporality and

causality might be tightly connected. It's widely assumed, for example, that effects can't precede their causes.²⁰⁶ If so, then the transcendental ideality of causation might follow from the transcendental ideality of time.

The nature of my example relied on the transcendental reality of time: Computation appears to require state transitions, which seems to require change over time.²⁰⁷ Arguably, these changes are also causal: Angel's memory association of +1 with $D^{\#}_5$ caused such-and-such a state transition. I wanted an example that was straightforward to imagine, and the possibility of Angel is sufficient to establish transcendental idealism as defined at the beginning of this chapter.²⁰⁸ However, a transcendental idealism committed to treating time and causation as also dependent on our minds might be more plausible if space, time, and cause are as intimately related as they appear to be. Cartesian souls, as temporal entities with causal powers, would thereby also be transcendently ideal. A noumenon without space, time, or causation, which isn't material but also doesn't conform to our standard conceptions of the mental, is difficult, maybe impossible, to imagine with any specificity. Perhaps we can imagine it only negatively and abstractly. Angel is paint-ball Kant for a prepaid hour. Kant proper is a live ammo home invasion in complete darkness.

Is there any reason to take transcendental idealism seriously as a live possibility? Or might it be a remote, in-principle possibility as negligibly unlikely as being a brain in a vat?

²⁰⁶ Faye 2001/2021.

²⁰⁷ Some have argued that computation can occur in a manner that abstracts away from temporal transitions (Steinhart 2014; Tegmark 2014), but I find this claim difficult to assess and see no need to employ it here.

²⁰⁸ Another possible shortcoming of the Angel example is that on certain versions of structuralism about space, it might turn out that Angel, contra my intentions in designing the example, does in fact have spatial properties. See discussion in Schwitzgebel 2019a; Chalmers 2019b.

(See Chapter 4 on the difference between negligible and non-negligible doubts.) I see four reasons.

First, materialism faces problems as a philosophical position, including in the difficulty articulating what it is to be “material” or “physical”, in the widespread opinion that it could never adequately explain consciousness, and in the fact (as argued in Chapters 2 and 3) that all well-worked out materialist approaches commit to highly bizarre consequences of one sort or another.²⁰⁹ Pressure against materialism is pressure in favor of an alternative position, and transcendental idealism is a historically important alternative position.

Second, as Nick Bostrom and others have argued, and as I argued in Chapter 4, 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, as I have also argued, it is reasonable for us to have substantial skepticism about the correct metaphysical picture of the relation of mind to world (Chapter 3) and more generally about our position in the cosmos (Chapter 4). 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 the universe, and is not even uniformly materialist. We have seen, perhaps, only a minuscule portion of it.

Fourth, most or all of what we know about material things (apart from what is knowable innately, or *a priori*, or transcendently, or through mystical revelation) depends on how those material things affect our senses. But things with very different underlying properties, call them A-type properties versus much weirder and less comprehensible B-type properties, could

²⁰⁹ On the first point, see note XXX above. On the second, see, e.g., Chalmers 1996.

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.²¹⁰

9. *Transcendental Idealism and Skepticism.*

Defenders of the possibility that we live in a simulated virtual reality, including Bostrom, Chalmers, and Steinhart, have tended to emphasize that this needn't be construed as a skeptical possibility.²¹¹ Even if we are trapped in the Matrix by evil supercomputers, ordinary things like cups, hands, 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, their 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 blanket rebuttal of radically skeptical scenarios, it is too optimistic. If we are living in a simulation, there's no compelling reason to believe that it must be a large, stable simulation. It might be a simulation run for only two

²¹⁰ For versions of this fourth consideration in favor 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 fundamental reality is relatively simple, then we can infer that the A-type properties are more likely. Or maybe, if we had good reason to think that fundamental reality isn't weird, we could reject the B-type in favor of the A-type. Or....

²¹¹ Bostrom 2011; Chalmers 2012, 2017; Steinhart 2014. Similarly, and contrary to the ideas of this section, Kant argues that transcendental idealism has anti-skeptical consequences, as I discuss briefly in Chapter 4.

subjective hours before shutdown. It might be a simulation containing only you in your room reading this book. If that's what's going on beneath appearances, then much of what you probably think you know 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 ways that negate huge swaths of our supposed knowledge.

Consider these two simulation scenarios, both designed to undercut the durable stability assumption.

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 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 and long enough to address their research questions, awe potential mates, or 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 infinitely 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 anyone 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 majority of them host conscious subjects only briefly, soon lapsing into disorganization.²¹²

[Illustration 5 (Caption: God stumbles over the power cord): A sleepy person holding a cup of coffee is about to trip over the power cord of a computer in a dingy, cluttered office. On the computer monitor is planet Earth as seen from space and the caption (still on the monitor) "sim-Earth, year 2024". A stack of books has the following titles visible: "How to Run a Sim", "Godhead for Mortals", "Programming People", and "Kritik der reinen Vernunft".]

If our empirical knowledge about simulations is any guide, most simulations are small scale. If our empirical knowledge is no guide, we should be even more at sea. If we leave simulation scenarios behind, generalizing transcendental idealism as recommended in Section 8, then the noumenal is almost completely incomprehensible. Our empirical reality might then be subject to whims and chances far beyond our ken. The Divine might stumble over the power cord at any moment, ending us all. Or even worse, $\alpha_\phi \mathcal{U}$ might $\dashv \mathbb{V}^\infty$.²¹³ Transcendental idealism explodes, or ought to explode, anti-skeptical certainty that we understand the origin, scope, structure, and stability of the cosmos.

²¹² The cosmological literatures on Boltzmann brains and the anthropic principle are relevant here, e.g., Barrow and Tipler 1986; Bostrom 2002; Carroll 2010. I explore a version of the Random scenario in Schwitzgebel 2017b.

²¹³ Please forgive the inadequate notation.

The Weirdness of the World

Part Two: The Size of the Cosmos

The Weirdness of the World

Chapter Six Experimental Evidence for the Existence of an External World

with Alan Tonnies Moore

It occurs to me to wonder whether anything exists besides my stream of conscious experience. *Radical solipsism*, I'll say, is the view that my consciousness is the only thing that exists in the universe.²¹⁴ There are no material objects, no other minds, not even a hidden unconscious side of myself – no “external world” of any sort at all. *This* (here I gesture inwardly at my sensory, emotional, and cognitive experiences) is all there is, nothing more. What a strange and lonely metaphysical picture! I'd like some evidence or proof of its falsity.

You might think – if you exist – that any desire for proof is foolish. You might think it plain that I could never *show* radical solipsism to be false, that I can only *assume* that it's false, that any attempt to prove solipsism wrong would inevitably turn in a circle. You might think, with (my seeming memory of) Wittgenstein, that the existence of an external world is an unchallengeable framework assumption necessary for any inquiry to make sense – that it's a kind of philosophical disease to want to rationally refute solipsism, that I might as well hope to establish the validity of logic using no logical assumptions.²¹⁵

I'll grant that I might be philosophically sick. But it's not entirely clear to me, at least not yet, that I can't find my cure from within the disease, by giving my sick mind exactly the proof it wants.

²¹⁴ If the cosmos contains more than one universe, then solipsism as defined here is consistent with the existence of other entities in other universes. *Cosmic solipsism* would deny even that.

²¹⁵ Wittgenstein 1950-1951/1970.

1. *Historical Prelude.*

As first blush, the historical evidence – or what I think of as a the historical evidence – invites pessimism about the prospects of satisfactory proof. The two most famous attempts to cure radical solipsism from within come from Descartes, in his *Meditations*, and from Kant, in his “Refutation of Idealism”.²¹⁶ Neither succeeds. Descartes’s proof of the external world requires accepting, as an intermediate step, the dubious claim that the thought of a perfect God could only arise from a being as perfect as God.²¹⁷ Kant’s proof turns on the assertion that I cannot be “conscious of my own existence as determined in time” or conscious of change in my representations unless I perceive some permanent things that actually exist outside of me. However, he offers no clear argument for this assertion. Why couldn’t a sense of representational change and of my determination in time arise innately, or from temporally overlapping experiences, or from hallucinatory experiences as if I saw things that exist outside of me?²¹⁸ Most philosophers today, it seems, regard as hopeless all such attempts to prove solipsism false using only general logic and solipsism-compatible premises about one’s own conscious experience.

So we might, with David Hume, yield to the skeptic, granting them the stronger argument, then turn our minds aside awhile, play some backgammon, and go on living and philosophizing just as before, only avoiding the question of radical solipsism.²¹⁹ Or we might, with G. E. Moore, defend the existence of an external world by means of solipsism-*incompatible*

²¹⁶ Descartes 1641/1984; Kant 1781/1787/1998, B 274-279, 326-328.

²¹⁷ See discussion in Broughton 2002; Nolan and Nelson 2006; Wagner 2014.

²¹⁸ See discussion in Guyer 1987; Stroud 1984, 1994; and especially Chignell’s objections to Dicker’s reconstruction (Dicker 2008, 2011, 2012; Chignell 2010, 2011).

²¹⁹ Hume 1740/1978.

premises that beg the question by assuming the falsity of what we aim to prove: Here is a hand, here is another, therefore there are external things. What, you want stronger proof than *that*?²²⁰ Or we might, with Wittgenstein, try to undercut the very desire for proof. However, none of these responses seems preferable to actually delivering a non-question-begging proof if such a proof is discoverable. They are all *fallback maneuvers*. Another type of fallback maneuver can be found among recent “contextualist” and “reliabilist” epistemologists who concede to the radical solipsist that we can’t know that the external world exists once the question of its existence has been raised in a philosophical context, while insisting that we nonetheless have ordinary knowledge of the mundane facts of practical life.²²¹

The historical landscape has been dominated by those two broad approaches. The first approach aims high, hoping to establish with certainty, in a non-question-begging way, that the external world really does exist. The second approach abandons hope of a non-question-begging proof, seeking in one way or another to make us comfortable with its absence.

But there is a third approach, historically less influential, that deserves further exploration. Its most famous advocate is Bertrand Russell. Russell writes:

In one sense it must be admitted that we can never *prove* the existence of things other than ourselves and our experiences.... There is no logical impossibility in the supposition that the whole of life is a dream, in which we ourselves create all the objects that come before us. But although this is not logically impossible,

²²⁰ Moore 1939.

²²¹ For example Williams 1991; DeRose 1995; Lewis 1996; Dretske 2003 (assuming that their remarks about skepticism in general apply also to the specific case of radical solipsism), though see DeRose 2017 and Ichikawa 2017 on combining contextualism with a Moorean approach. Relatedly, responses to skepticism that assume that I already believe that the external world exists, such as Sosa’s (2000), “safety” response, do not apply to the present case since I have suspended belief.

there is no reason whatsoever to suppose that it is true; and it is, in fact, a less simple hypothesis, viewed as a means of accounting for the facts of our own life, than the common-sense hypothesis that there really are objects independent of us, whose action on us causes our sensations.²²²

Russell also states that certain experiences are “utterly inexplicable” from the solipsistic point of view and that the belief in objects independent of us “tends to simplify and systematize our account of our experiences”. For these reasons, he says, the evidence of our experience speaks against solipsism, at least as a “working hypothesis”.²²³ Russell aims lower than do Descartes and Kant, and partly as a result his goal seems more plausibly attainable. Yet Russell also promises something that Hume, Wittgenstein, and Moore do not: a non-question-begging positive argument against solipsism. It’s a middle path between certainty and surrender or refusal.

Unfortunately, Russell’s argument has two major shortcomings. One is its emphasis on simplicity. The most natural way to develop the external world hypothesis, it seems, involves committing to the real existence of billions of people, many more billions of artifacts, and naturally occurring entities vastly more numerous even than that, of many types, manifesting in complex and often unpredictable patterns. It’s odd to say that such a picture of the world is simpler than radical solipsism.²²⁴

²²² Russell 1912, p. 22-23, emphasis in original.

²²³ Russell 1912, p. 23-24, and 1914, p. 103-104.

²²⁴ On the complex issue of the nature of simplicity, see Sober 1975; Vogel 1990; Zellner et al. 2001; Peacocke 2004; Cowling 2013; Schaffer 2015.

The second shortcoming is the unconvincing, gestural nature of Russell's supporting examples. What is it, exactly, that is "utterly inexplicable" for the solipsist? It's a cat's seeming hunger after an interval during which the cat was not experienced:

If [the cat] does not exist when I am not seeing it, it seems odd that appetite should grow during non-existence as fast as during existence. And if the cat consists only of sense-data, it cannot be *hungry*, since no hunger but my own can be a sense-datum to me. Thus the behaviour of the sense-data which represent the cat to me, though it seems quite natural when regarded as an expression of hunger, becomes utterly inexplicable when regarded as mere movements and changes of patches of colour, which are as incapable of hunger as a triangle is of playing football.²²⁵

To this example, Russell appends a second one: that when people seem to speak "it is very difficult to suppose that what we hear is not the expression of a thought".²²⁶

But are such experiences really *so* explicable for the solipsist? Consider hallucinations or dreams, which arguably can involve apparent hungry cats and apparent human voices without, behind them, real feline hunger or real human minds independent of my own. Russell considers this objection but seems satisfied with a cursory, single-sentence response: "But dreams are more or less suggested by what we call waking life and are capable of being more or less accounted for on scientific principles if we assume that there really is a physical world".²²⁷ The inadequacy of this response is revealed by the fact that the solipsist could say something similar: If I assume *solipsism* then I can account for the appearance of the cat and the interlocutor as imperfect

²²⁵ Russell 1912, p. 23, emphasis in original.

²²⁶ Russell 1912, p. 24.

²²⁷ Russell 1912, p. 24.

projections of myself upon my imagined world, grounded in what I know about myself through introspection. Such an explanation is sketchy to be sure, but so also is the present scientific explanation of dreaming and hallucination. At best, Russell's argument is seriously underdeveloped.

Although I am dissatisfied with Russell's particular argument, as I sit here (or seem to) with my solipsistic doubts, I still feel the attraction of the general approach. The core idea I want to preserve is this: Although in principle (contra Descartes and Kant) all the patterns in my experience are compatible with the nonexistence of anything beyond that experience, I can still evaluate two competing hypotheses about the origins of those experiences: the solipsistic hypothesis, according to which all there is in the universe is those experiences themselves, and the external world hypothesis, which holds that there is something more. I can consider these hypotheses not by the standards of airtight philosophical proof beyond doubt but rather as something like scientific hypotheses, with epistemic favor accruing to the one that does the better job accounting for my experiential evidence. Although Russell's remarks are too cursory, that doesn't speak against the general project. A more patient effort might still bear fruit.²²⁸

²²⁸ More recently, BonJour 1985, 2003; Vogel 1990, 2005, 2008; Peacocke 2004; and McCain 2014 have also argued against radical skepticism on explanationist grounds. See Beebe 2009 and McCain 2012 for reviews. Vogel, for example, asserts that the skeptic will be forced to choose between leaving their experiences unexplained and deploying ad hoc explanations: "Niceties aside, the fact that something is spherical explains why it behaves like a sphere (in its interactions with us and with other things). If something that is not a sphere behaves like one, this will call for a more extended explanation" (Vogel 1990, p. 663-664). Like Russell, however, these remarks remain sketchy. Furthermore, discussions tend to focus on "evil deceiver" skepticism rather than radical solipsism. Chalmers (2012, 2018) provides an interesting structural argument against a global deceiver scenario, but he does not engage with radical solipsism. Karl Popper (1983, p. 83-84, and 1994, p. 107) focuses explicitly on solipsism, and his example of being unable to write the plays of Shakespeare resembles the examples in Experiments 1 and 3 – and yet, like Russell, his treatment is sketchy and doesn't adequately consider possible replies by the solipsist.

In place of Russell's vaguely scientific appeal, I will try an *actual empirical test* of the two hypotheses. I will generate specific, conflicting empirical predictions from radical solipsism and from non-skeptical external world realism, pitting the two hypotheses against each other in what I hope is a fair way, and then I will note and statistically analyze the experimental results. After each experiment, I will consider the limitations of the research and what alternative hypotheses might explain the findings.

In other words, I aim to do some solipsistic science. There is no contradiction in this. Skepticism about the external world is one thing, skepticism about scientific methodology quite another. I aim to see whether, from assumptions and procedures that even a radical solipsist can accept, I can generate experimental evidence for the existence of an external world.

2. *Ground Rules.*

This project needs some rules, laid out clearly in advance. I don't hope to prove something from nothing. The skeptic's position is unassailable if the opponent must prove all the premises of any potential argument. That forces either infinite regress or a logical circle.²²⁹ In this chapter, I aim to refute not every possible skeptical position but only radical solipsism. I aim only to move from solipsism-compatible premises to an anti-solipsistic conclusion. That would be accomplishment enough! To my knowledge, this has never been done rigorously and well.

²²⁹ These are the first two horns of the "Agrippan trilemma", the third of which is – as we do here – stopping one's attempts at proof (Comesaña and Klein 2001/2019). Pihlström 2020 argues that considering solipsism is valuable because it forces us to consider such facts about the limits of argumentation and the nature of philosophical commitment when argumentation falls short.

Accordingly, for this project, I don't plan to entertain any more than ordinary degrees of doubt about solipsism-compatible versions of scientific method, or deduction, or medium-term memory, or introspective self-knowledge. Specifically, I will permit myself to assume the following:

- introspective knowledge of sensory experience and other happenings in the stream of experience;
- memories of past experiences from the time of the beginning of the series of experiments but not before;
- concepts and categories arrived at I-know-not-how and shorn of any presupposition of grounding in a really existing external world;
- the general tools of reason and scientific evaluation to the extent those rules avoid assumptions about affairs beyond my stream of experience.

Drawing only on those resources, I will try to establish, to a reasonable standard of scientific confidence, the existence of an external world.

My reasons for not drawing on memories of yesterday are two. First, I am curious whether, from a radical skeptical break in which I cut away the past and all things external, I can build things back up. Second, if my memory of previous days does serve me correctly, one of the things it tells me is that it itself is highly selective and unsystematic and that free recall of what is salient and readily available is not usually good scientific method.

If solipsism implied that I had complete control over my stream of experience, I could easily refute it. I could, for example, take in my hands a deck of cards (or at least seem to myself to do so) and resolutely *will* that I draw a queen of clubs. Then I might note the failure of the world to comply. In fact, I have now attempted exactly this, with an apparent seven of spades as

my result. Unfortunately for the prospects of such an easy proof, solipsism has no such voluntaristic implications and thus admits of no such antivoluntaristic refutation – contra certain remarks by John Locke, George Berkeley, and Johann Fichte.²³⁰ Such *compliant world solipsism* is a flimsy cousin of the more robust version of solipsism I have in mind.

To contemplate this last issue more clearly, I close my eyes – or rather, not to assume the existence of my body, I do something that seems to be a closing of the eyes. What I visually experience is an unpredictable and uncontrollable array of colors against a dark gray background, the *Eigenlicht*.²³¹ This uncompliant *Eigenlicht* is entirely compatible with solipsism as long as I conceptualize the patterns it contains as nothing but patterns in, or randomness in, my stream of experience – that is, patterns governed by their own internal coherences rather than by anything further behind them. The unpredictability and uncontrollability of these visual patterns no more compels me to accept the existence of nonexperiential entities than irresolvable randomness and unexplained laws in the material world, as I conceptualize it, would compel me to accept the existence of immaterial entities behind the material ones.

I could ensure the impossibility of refuting solipsism by insulating it in advance against making any predictions that conflict with the predictions of the external world hypothesis. But there are only two ways to do this, both unscientific. Radical solipsism could avoid making predictions in conflict with the external world hypothesis by predicting nothing at all. This would make it unempirical and unfalsifiable. Such a hypothesis would either not belong to

²³⁰ Locke 1689/1975, IV.xi; Berkeley 1710/1965, §29 ff.; Fichte 1797/2000, §2. Descartes also entertains the idea of such a proof but ultimately notes its inadequacy: 1641/1984, *Meditations* 3 and 6.

²³¹ See Schwitzgebel 2011b, ch. 8, for a historical review of the fascinating introspective psychological literature on the *Eigenlicht* and the radical changes in its description by psychologists from different eras and with different theories.

science or belong to science but always be less well confirmed by incoming data than hypotheses that make risky predictions that are borne out.²³² Or it could avoid making predictions in conflict with the external world hypothesis by being jury-rigged post-hoc so that its predictions precisely match the predictions of the external world hypothesis without generating any new organic predictions of its own. This would make it entirely derivative on its competitor, adding no value but needless theoretical complexity – again, clearly a failure by ordinary standards of scientific evaluation.

For purposes of this project, radical solipsism is best understood as a view that can accommodate uncontrollable sensory patterns like the *Eigenlicht* but which also naturally generates *some* predictions at variance with the predictions of the external world hypothesis. In particular, as I will discuss in more detail shortly, radical solipsism is naturally interpreted as predicting the nonexistence of evidence of entities that can outpredict or outreason me or persist in detail across gaps in my sensations and memory of them.

I will now describe three experiments, all conducted in one uninterrupted episode on a single day. The first experiment provides evidence for the existence of an entity better than me at calculating prime numbers. The second provides evidence for the existence of an entity with detailed features that are forgotten by me. The third provides evidence of the existence of an entity whose cleverness at chess frustratingly exceeds my own. In all cases, I will evaluate the scientific merits of a solipsistic interpretation of the same experiences.²³³

²³² The classic source on falsifiability is Popper 1935/1959/2002. A similar idea is also central to Bayesian confirmation theory (e.g., Earman 1992; Talbott 2001/2016).

²³³ Stanislaw Lem has a character perform an experiment similar to the first of these, in his science fiction novel *Solaris* (1961/1970, p. 50-51). For discussion of the advantages and shortcomings of Lem's experiment, see Schwitzgebel 2014a.

To the extent possible, the remainder of this chapter, apart from the final concluding section, reflects real thoughts on the day of experimentation, with subsequent modifications only for clarity. To fit all of these thoughts into the span of a single day, I drafted a version of the material below in the present tense using dummy results based on pilot experiments. I entered into the experiment with the intention of genuinely thinking the thoughts below with real data as the final results came in.

3. Experiment 1: The Prime Number Experiment.

Method. I have prepared for this experiment by programming Microsoft Excel to calculate whether a four-digit number is prime, displaying “prime” next to the number if it is prime and “nonprime” otherwise. Then I programmed Excel to generate arbitrary numbers between 1000 and 4000, excluding numbers divisible by 2, 3, or 5. Or rather, I should say, without assuming the existence of a computer or Excel, this is what I now seem to remember having done. Version A of this experiment will proceed in four stages, if all goes as planned. First, I will generate a fresh set of 20 new qualifying four-digit numbers. Second, I will take my best guess which of those 20 are prime, allotting approximately two seconds for each guess. Third, I will paste the set of 20 into my seeming prime number calculator function, noting which are marked as prime by the seeming-machine. Finally, by laborious manual calculation, I will determine which among those twenty numbers actually are prime. Version B will proceed the same way, except using Roman numerals.

My hypothesis is this: If nothing exists in the world apart from my stream of conscious experience, then the swift, seemingly Excel-generated answers should not be statistically more accurate than my best guesses. For if they were more accurate, that would suggest the existence

of something more capable of swift, accurate prime-number detection than my own solipsistically-conceived self.

Results. I have just now completed the experiment as described. The main results are displayed in Figure 1. For Version A, my best guesses yielded an estimate of 11 primes. In most cases this felt like simple hunchy guessing, though 3913 did pop out as nonprime. The apparent Excel calculation also yielded an output of 11 primes. Manual calculation confirmed the seeming-machine results in 19 out of 20 cases.²³⁴ In contrast, manual calculation confirmed my best-guess judgment in only 11 of the 20 cases. The difference in accuracy between 19/20 and 11/20 is statistically significant by Fisher's exact test (manually calculated), with a two-tailed p value of $< .02$.²³⁵ In other words, the apparent computer performed so much better than I did, that it is statistically very unlikely that that the difference in our success rates was entirely due to chance. For Version B, again both my best guesses and the apparent Excel outputs yielded 11 estimated primes, and again manual calculation confirmed the apparent Excel outputs in 19 of the 20 cases,²³⁶ while manual calculation confirmed my best guesses in 13 of the 20 cases. The difference in accuracy between 19/20 and 13/20 is marginally statistically significant by Fisher's exact test (manually calculated), with a two-tailed p value of approximately $.05$.²³⁷

²³⁴ Subsequent examination of my notes suggests a long-division error on my part and thus that the apparent Excel output was correct after all.

²³⁵ Here and elsewhere I used some approximations to ease calculation, always approximating in the statistically conservative direction of overestimating the p value.

²³⁶ Subsequent examination of my notes suggests an error in my manual conversion from Roman to Arabic numerals.

²³⁷ Subsequent analysis with a statistical program yields $p = .044$. For Version A, subsequent analysis yields $p = .008$.

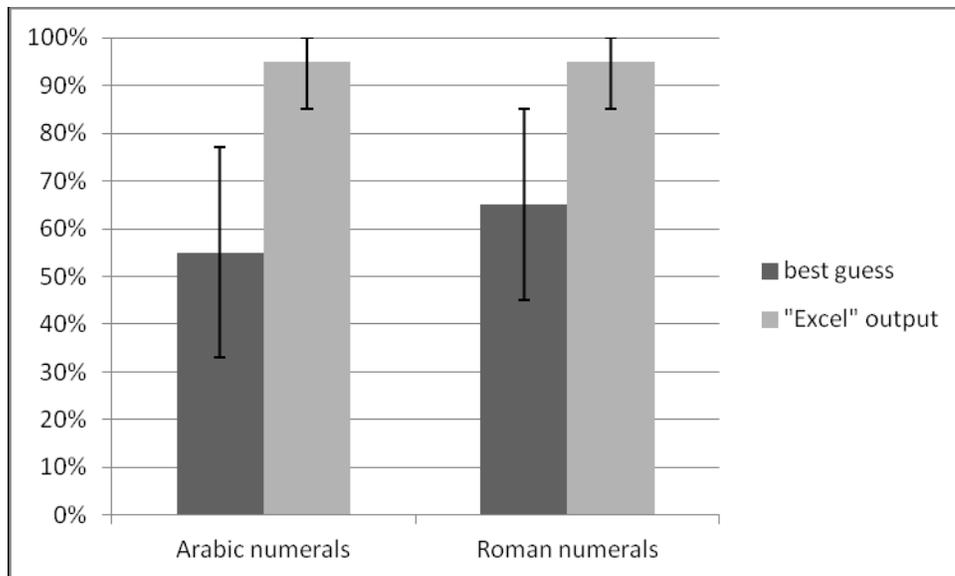


Figure 1: Accuracy of prime number estimates, as judged by manual calculation, for my best guesses before calculating, compared to the apparent outputs of an Excel spreadsheet programmed to detect primes. Error bars are manually calculated 95% confidence intervals using the normal approximation and a ceiling of 100%.

Discussion. I believe the most natural interpretation of these results is that something exists in the external world that has calculation capacities exceeding my own. Although I was able to manually confirm most of the answers, I couldn't do so swiftly, nor recognize the primes correctly as they arose. The best explanation of the impressive 19/20 success rate seems to be that, in the instant that I seemed to drag down the Excel function, something outside of my solipsistically-conceived stream of experience performed calculations of which I was not conscious.

As previously mentioned, I am setting aside skeptical concerns about memory within the span of the experiment (what if the world was created two seconds ago?) and introspection (what

if I delusionally misjudged my intentions and sensory experiences?²³⁸). My aim, as I've emphasized, is not to apply radically skeptical standards generally, but rather to employ normal standards of science insofar as they can be employed by someone open-minded about radical solipsism. I also set aside concerns about whether this seeming computer really *calculates* versus being designed to produce outputs interpretable by users as the results of calculation. Either way, the results suggest the existence of someone or something with prime-number detection capacities exceeding those present in my solipsistically conceived stream of experience. Even if that thing is only my own unconscious mind, bent on tricking my conscious mind into misinterpreting my experimental results, radical solipsism as I've defined it would be false, since radical solipsism denies the existence of *anything* outside of my stream of experience.

As I reflected earlier, solipsism can readily allow that the stream of experience contains patterns within it as long as those patterns are not caused by anything behind that experience. The anti-solipsistic interpretation of the experimental results thus turns crucially on the question of whether the outcome of this experiment might plausibly be only a manifestation of such solipsistic patterns of experience. How plausible would such a pattern of experiences be, given solipsistic assumptions? What should I expect patterns of experience to look like if solipsism is true?

²³⁸ From within the perspective of this project, I intend my judgments not to be perceptual judgments about the outside world but rather introspective judgments about my experiences of the outside world. In fact, I think this distinction is ontologically fraught (Schwitzgebel 2012b) and psychologically difficult to sustain (Schwitzgebel 2005). Furthermore, I'm skeptical about the accuracy of introspective judgments that aren't derived from knowledge of the outside world (Schwitzgebel 2011b, ch. 7). I disregard all these qualms here.

These are hard questions to answer. And yet I don't want to be *too* hard on myself. I'm looking only for scientific plausibility, not absolute certainty.

Examining my experience now, one typical type of pattern is this: When I do something that feels like shifting my eyes to the left, my experienced visual field seems to shift to the right – a fairly simple law of experience, a simple way in which two experiences might be directly related with no compelling explanatory need of a nonexperiential intermediary. Likewise, when I seem to see a cylindrical thing – this water bottle here – and then seem to reach out to touch it, I seem also to have tactile experiences of something cylindrical. This more complex pattern is not fully expressible by me, but still it seems a fairly straightforward set of relationships among experiences. It's tempting to think that there must be a genuine mind-independent material cylinder that unifies and structures these experiences across sensory modalities. But if I am to be genuinely open-minded about solipsism, I must admit that the existence of a radically new ontological type (“mind-independent thing”) is a heavy cost to pay to explain this relationship among my experiences. The visual and tactile experiences might be related to each other by an admittedly somewhat complex set of intrinsically experiential laws, as suggested by John Stuart Mill in contemplating a similar example.²³⁹ Similarly, when I close my eyes, there are regularities – the visual field changes radically in roughly the way I expect but also gains some unpredictable elements. Solipsism can also allow the existence of unrecognized patterns in the relationships among my experiences. For example, afterimages of bright seeming-objects might be in perfect complementary colors (red objects always followed by green afterimages, etc.) even if I don't realize that fact. There might also be discernable but as-yet-undiscovered regularities

²³⁹ Mill 1867. I see no reason to assume, as Laurence Lafleur (1952) assumes in his attempted refutation, that a solipsism requires that all the laws of experience are known or even knowable.

governing the flight of colors I experience when my eyes are closed. All of this seems plausibly explicable by solipsistic laws that relate experiences directly one to another. The core question is: Could laws of broadly this sort suffice to explain my experimental results?

To explain the results of Experiment 1 solipsistically via such unmediated laws of experience, something like the following would have to be true. There would have to be an unmediated relationship between, on the one hand, the visual experience of, for example, the numeral “2837” in apparent black Calibri 11-point font in an apparent Excel spreadsheet, accompanied by the inner speech experience of saying to myself, with understanding, ‘twenty-eight thirty-seven’ in English, and, on the other hand, the visual experience of suddenly seeing “prime” in the matched column to the right if the number is prime and seeing “nonprime” otherwise. At first blush, such an unmediated relationship strikes me as at least a little strange. On radical solipsism, I’m inclined to think, it should be some feature of the visual experience that drives the appearance of “prime” or “nonprime” on the seeming screen. But what feature could that be? The visual experience of the numeral “1867” (prime) doesn’t seem to share anything particularly more in common with the visual experience of “3023” (also prime) than with “1367” (nonprime) – and still less with “MMCCCXXXIII”. What seems to unify the pattern of results is instead a semantic feature, primeness of the represented number, something that is neither on the sensory surface nor, seemingly, present anywhere else in my conscious mind before the seeming computer receives the instruction to calculate. Such a semantic property strikes me as an unlikely candidate to drive unmediated regularities in a solipsistic universe. How would it find traction? “Display ‘prime’ if and only if the squiggles happen to represent a prime number in a recognizable numeral system” seems too complex or arbitrary to be a fundamental law. The structural pattern seems too disconnected from the surface

complexities of my experience at the moment to be driven directly by those surface complexities. Solipsism more naturally predicts, I think, what I did in fact predict on its behalf: that just as I might expect in a dream, the assortment of “prime” and “nonprime” should be unrelated to actual primeness except insofar as I have guessed correctly.

And yet, given my acknowledged bias against solipsism, it would be imprudent for me to leap too swiftly from this one experiment to the conclusion that solipsism is false. Maybe this is exactly the sort of law of experience that I should expect on a solipsistic worldview, if I treat that view with appropriate charity? Maybe solipsism can be developed to accommodate laws that directly relate unrecognized semantic properties of numeral experiences to semantic properties of English orthography experiences, or something like that. One awkward result does not a decisive refutation make. So I have some further experiments in mind.

4. Experiment 2: Two Memory Tests.

Method. I am currently having a visual experience of an apparent person. I am inclined to think of this apparent person as my graduate student collaborator, Alan. I have arranged for this seeming “Alan” to test my memory. In the first test, he will orally present to me an arbitrary-seeming series of 20 three-digit numbers. He will present this list to me twice. I will first attempt to freely recall items from that list, and then I will attempt to recognize those items from a list of 40 items, half of which are new three-digit combinations. The second test will be the same procedure with 20 sets of three letters each and with a two-minute enforced delay to further impair my performance. In both cases, I expect that seeming-Alan will tell me that my memory has been imperfect. He will then, if all goes according to plan, tell me that he is visually re-presenting the original lists. If solipsism is true, nothing should exist to anchor the advertised

visual “re-presentation” to the earlier orally presented lists, apart from my own memory. In those cases, then, where my memory has failed, the supposed re-presentation should not contain a greater number of the originally presented elements than would be generated by chance. I shouldn’t be able to step into the same stream twice except insofar as I can remember that stream (though maybe I could have the illusory feeling of having done so). The contents of experience should not have a fixity that exceeds that of my memory, because nothing exists beyond my own experience to do the fixing. At least, this seems to me the most straightforward prediction from within the solipsistic worldview. Thinking momentarily as a solipsist, this is what I expect to be the nature of things.

The final move in the experiment will be to test whether the re-presented list does indeed match the original list despite the gap in my memory. You (my imagined reader or interlocutor) might wonder, how is such a test possible? How could a solipsist distinguish *genuine* constancy across a gap of memory from a sham *feeling* of constancy? The method is this: Seeming-Alan will state the procedure by which he generated the seemingly arbitrary lists. By (seeming) prior arrangement, he will have used a simple arithmetic procedure for the numbers and a simple semantic procedure for the letters (such as the decimal expansion of a four-digit fraction and a simple transformation of a familiar text). I should then be able to test whether the later-presented full lists of 20 three-element items are indeed consistent with generation by the claimed procedures. If so, this will suggest that the original lists were also generated by those same procedures. It will do so, if all goes well, because (as I will later estimate) there’s only a very small chance that two arbitrary lists of 20 three-element items would have several items in common – the several items I will presumably remember across the temporal gap – unless they were generated by the same procedure. For example, the decimal expansion of $1/6796$ is

.000147145379635.... If both my memory and the “re-presented list” begin “147”, “145”, “379”, and if both lists contain only other triples from the first 63 decimal places of the digital expansion of $1/6796$, then it is overwhelmingly likely that both lists were generated from that expansion rather than randomly or arbitrarily. This would then allow me to infer that the entire re-presented list does indeed match the entire original list despite my failure to recall some items across the interval – in conflict with the no-same-stream prediction of the solipsistic hypothesis.

Results. The main results are displayed in Figure 2. According to seeming-Alan, in the number test, I correctly recalled 7 of the 20 three-digit items (with no false recall), and I accurately recognized 14 of the items. In the letter test, I correctly recalled 8 of the 20 three-letter items (with no false recall) and accurately recognized 18. The generating patterns, he claims, were the decimal expansion of $1/2012$, excluding the initial zeroes, and the most famous lines of Martin Luther King, Jr.’s famous “I Have a Dream” speech, skipping every other letter and excluding one repeated item. In both cases I manually confirmed that the re-presented lists conformed to the purported generation procedure. Manual application of the two-tailed Fisher’s exact test shows my recall to match significantly less well to the re-presented lists than do the manually confirmed results (both p ’s $< .001$). At a $p < .05$ (i.e., 5%) confidence level, the recognition results are statistically significant for the three-digit items but not the three-letter items.

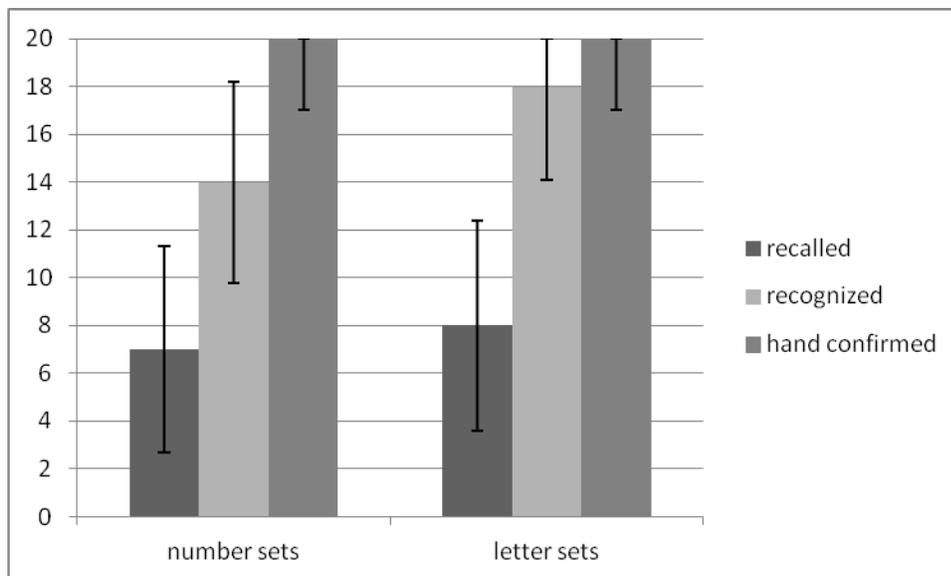


Figure 2: Number correct out of 20 as judged by comparison with the lists “re-presented” by seeming-Alan, for my recall guesses, my recognition-test guesses, and my manual confirmation of the purported generation procedure. Error bars for “recalled” and “recognized” are manually calculated 95% confidence intervals using the normal approximation. Error bars for the ceiling results use the “rule of three”.

I found myself trying hard in both memory tasks. Since I’m inclined to believe that the external world does exist and thus that some other people might read what I now appear to be writing, I was motivated not to come across as sieve-brained. This created a substantial incentive to answer correctly, which I did not feel as strongly in Experiment 1.

Discussion. I believe that the most natural interpretation of these results is that something existed in the external world that retained the originally presented information across my gap of memory.

Alternative interpretations are possible as always in science. The question is whether I should find those alternative explanations scientifically attractive, considering the truth or falsity

of solipsism as an open question. Can solipsism accommodate the results naturally, without the strained, ad hoc excuses that are telltale sign of a theory in trouble?

Might there have been an direct, unmediated connection between the original auditory experience of “IAE”, “DEM”, etc., and the later visual experience of those same letter arrangements? If I am willing to countenance causation backward in time, then a neat explanation offers itself: I might have solipsistically concocted the generating patterns at the moment that seeming-Alan seemed to be informing me of them, and then those generating patterns might have backwardly caused my initial auditory experiences and guesses. However, temporally backward causation seems a desperate move if my aim is to apply normal scientific standards as I conceptualize them. A somewhat less radical possibility is temporally gappy cross-modal causation from auditory experience at the beginning of the experiment to visual representation later. However, this requires, in addition to the somewhat uncomfortable provision of temporally gappy causation, a further seeming implausibility, similar to that in Experiment 1, of an unmediated law of experience that operates upon semantic contents of those experiences that are unrecognized as such by me at the time of the law’s operation. In this case, the relevant semantic contents would be nothing as elegant as primeness but rather the decimal expansion of one divided by the current calendar year,²⁴⁰ excluding the initial zeroes, and the English orthography of the words of a familiar-seeming speech, skipping alternate letters and excluding a repeated triplet.

The thought occurs to me that some of the laws of external-world psychology, as I conceive of it, are also weird and semantical. For example, an advertisement might trigger a

²⁴⁰ Yes, this experiment was conducted back in 2012, and the article on which this chapter was based was published in 2015. Perhaps I should refresh my confidence by performing these experiments again.

tangentially associated memory. But the crucial difference is this: In the case of external-world psychology, I assume that the semantic associations, even if not conscious, are grounded in mundane facts about neural firing patterns and the like. A bare solipsistic tendency to create and then recreate, unbeknownst to myself, the same partial orthography of a familiar speech while meanwhile being unable to produce that partial orthography when I consciously try to do so – well, that’s not impossible perhaps, but neither does it seem as natural a development of solipsism as does the view that the stability of experience should not exceed the stability of memory.

My argument would be defeated if I could have easily found some simple scheme, post hoc, that could generate twenty items including exactly those seven recalled numbers and eight recalled letter sets. My anti-solipsistic interpretation requires that there be only one plausible generating scheme for each set; otherwise there is no reason to think the unrecalled items would be out the same on the initial and subsequent lists. So, then, what are the odds of a post hoc fit of seven or more items from each set? Fortunately, the odds are very low – about one in a million, given some plausible assumptions and the mathematics of combination.²⁴¹

Perhaps, then, the best defense for solipsism, if it’s not to collapse into a general radical skepticism about short-term memory or scientific induction or arithmetic, is temporally gappy cross-modal forward causation grounded in unrecognized weird semantic features of the relevant

²⁴¹ Consider the odds of hitting one specific familiar sentence or one specific number-generating sequence. Give 1000 possible three-digit numbers, equally likely, or 1000 equally likely letter sets, the odds of exactly seven matches among 20 generated items are $(1/10^3)^7$ times 20-choose-7 possible arrangements (about 10^6), i.e., approximately one in 10^{15} . The odds of 8 or more matches adds only negligibly to this probability. Even if we assume a billion possible simple generating schemes along roughly the lines I recall having suggested to Seemingly-Alan, the odds of a chance match of at least seven of 20 to any one of those billion schemes are about one in a million (i.e., $10^9/10^{15} = 1/10^6$).

experiences. I'm inclined to think this is a somewhat awkward position for the solipsistic hypothesis. But maybe I'm still being too unsympathetic to solipsism? Maybe I should have expected that scientific laws would be strange like this in a solipsistic world and rather unlike the scientific laws I think of as characteristic of the natural sciences and naturalistic psychology? So I have planned for myself one final experiment of a rather different sort.

5. Experiment 3: Defeat at Chess.

Method. Seeming-Alan tells me that he is good at chess. I believe that I stink at chess. Thus, I have arranged to play 20 games of speed chess against seeming-Alan, with a limit of approximately five seconds per move. If solipsism is true, nothing in the universe should exist that has chess-playing practical reasoning capacities that exceed my own, and so I should not experience defeat at rates above statistical chance when directing all of my conscious efforts on behalf of one color. Figure 3 displays the procedure, as presented to me by a seeming camera held by a seeming Philosophy Department staff member.



Figure 3: The procedure of Experiment 3. Photo credit: seeming-Gerardo-Sanchez.

Results. Seeming-Alan defeated me in 17 games of 20, with one stalemate. Seventeen out of 19 is statistically higher than 50% with a p value of $< .001$ (manually calculated).

Discussion. Might I have hoped to lose, so as to generate results confirming my preferred hypothesis that the external world exists? Against this concern, I reassure myself with the following thoughts. If it as an *unconscious* desire to lose, that appears to imply the existence of something besides my own stream of experience, namely, an unconscious part of my mind, and thus radical solipsism as I have defined it would be false. If it was, instead, a *conscious* desire to lose, I should have been able to detect it, barring an unusual degree of introspective skepticism. What I detected instead was a desire to win as many games as I could manage, given

my background assumption that if Alan actually exists I would be hard pressed to win even one or two games. Playing with what I experienced as determination and energy, I found myself forcefully and repeatedly struck by the impression that the universe contained a practical intelligence superior to my own and bent on defeating me. The most natural scientific explanation of this pattern in my experience is that that impression was correct.

Does it matter that, if the external world exists in something like the form I think it does, some chess-playing machines could have defeated me as handily? I don't see why it should. Whether the strategy and intentionality is manifested directly by a human being or instead through the medium of a human-programmed chess machine – or in any other way – as long as that strategy and intentionality surpasses my own conscious mind, the solipsistic hypothesis faces a substantial explanatory challenge. It can try to address this challenge by appealing to intrinsic relationships among experiences of mine – relationships among seeming chess moves countered by other seeming chess moves whose power I only recognized in retrospect – but the more elegant and satisfying explanation of the results would appear to be the existence of a competing goal-directed intelligence.

Could I then maybe just abandon the pursuit of explanation? Could I just say it's a regularity unexplained, end of story? But why settle so quickly into explanatory defeat when the existence of a competing intelligence seems so readily available as an explanation? Simply shrugging, when I'm not forced to do so, runs contrary to the exploratory, scientific spirit of this exercise.

I might easily enough *dream* of being consistently defeated in chess. Maybe some dreamlike concoction of a seeming chess master could equally well explain my pattern of experience without need of an external world? But dreams of this sort, as I seem to remember,

differ from the present experiment in one crucial way: They are vague on the specific moves or unrealistic about those moves. In the same way, I might dream of proving Fermat's last theorem. Such cases involve dreamlike gappiness or irrationality or delusive credulity – the types of gappiness or irrationality or delusive credulity that might make me see nothing remarkable in discovering I am my own father or in discovering a new whole number between 5 and 6. Genuine dream doubt might involve doubt about my basic rational capacities, but if so, such doubts outrun simply solipsistic doubts.²⁴² Even if I am dreaming or in some other way concocting for myself an imaginary chess master, radical solipsism would still be defeated if the dream explanation implies that there is some unconscious part of my mind skilled at chess and on which I unwittingly draw to select those startlingly clever chess moves. I can easily imagine a world in which I am regularly defeated at chess, but relying on the resources only of my chess-inept conscious mind, I can no more specifically imagine the truly brilliant play of a chess expert than I can specifically imagine a world in which twenty four-digit numbers are all, in a flash, properly marked as prime or nonprime. If I consistently experience genuinely clever and perceptive chess moves that repeatedly exploit flaws in my own conscious strategizing, moves that I experience as surprising but retrospectively appreciate, it feels hard to avoid the conclusion that something exists that exceeds my own conscious intelligence in at least this one arena.

6. Conclusion.

²⁴² Perhaps it's worth noting that if we add such doubts to the dream argument in Chapter 3, it strengthens rather than weakens the argument for at least a smidgen of dream doubt by undercutting the trustworthiness of my reasoning in favor of non-skeptical realism. However, I did not want to rely on that aspect of dreams in that stage of the argument.

When I examine my stream of experience casually, nothing in it seems to compel the rejection of solipsism. My experience contains seeming representations of outward objects. It follows patterns unknown to me and that resist my will. But those basic facts of experience are readily compatible with the truth of radical solipsism. Once I find myself with solipsistic doubts, G. E. Moore's confident "here is a hand" doesn't help me recover. But neither do ambitious proofs in the spirit of Descartes and Kant seem to succeed. I could try to reconcile myself to the impossibility of proof, but that seems like giving up.

Fortunately, the external world hypothesis and the solipsistic hypothesis appear to make different empirical predictions under certain conditions, at least when interpreted straightforwardly. The external world hypothesis predicts that I will see evidence of theoretical reasoning capacities, property retention, and practical reasoning capacities exceeding those of my narrowly conceived conscious self, while solipsism appears to predict the contrary. I can then scientifically test these predictions and avoid begging the question by using only tools that are available to me from within a solipsistic perspective.

The results come out badly for solipsism. To escape my seemingly anti-solipsistic experimental results requires either adopting other forms of radical skepticism in addition to solipsism (for example, about memory, even over the short duration of these experiments) or adopting increasingly ad hoc, strained, and convoluted accounts of the nature of the laws or regularities connecting one experience to the next.

Did I really need to do science to arrive at this conclusions, though? Maybe instead of running formal experiments, I simply could have consulted long-term memory for evidence of my frustration by superior intelligences and the like? Surely so. And thus maybe also even before conducting these exercises I implicitly relied on such evidence informally to support my

knowledge that the external world exists. Indeed, it would be nice to grant this point, since then I can rightly say that I have known for a long time that the external world exists. Nevertheless, the present procedure has several advantages over attempts to remember past frustrations and failures. For one thing, it achieves its goal despite conceding more to the skeptic from the outset, for example, unbelief in yesterday. For another, it more rigorously excludes chance and confirmation bias in evidence selection. And for still another, it forces me to consider, starkly and explicitly, the best possible alternative solipsistic explanations I can devise to account for specific, concrete pieces of evidence – giving solipsism a chance, I hope a fair chance, to strut its stuff, if stuff it has.

Perhaps it's worth noting that the best experiments I could concoct all involved pitting my intelligence against another intelligence or against a device created by another intelligence – a device or intelligence capable of generating semantic or strategic patterns that I could appreciate only retrospectively. Whether this is an accidental feature of the experiments I happened to design or whether it reflects some deeper truth, I am unsure.²⁴³

I conclude that the external world exists. I can find no good grounds for doubt that something transcends my own conscious experience, and when I do try to doubt, the empirical evidence makes the doubt difficult to sustain. However, the arguments of this chapter establish almost nothing about the metaphysical character of the external world, apart from its capacity to

²⁴³ Maybe Fichte (1797/2000) and Hegel (1807/1977) wouldn't be surprised by this feature of my findings. See discussion in Beiser 2005 and Stern 2012. However, I share Stern's concerns about the force of these responses to solipsism as historically developed. Russell 1914, in partial contrast with Russell 1912, appears to accept the existence of an external world on testimonial grounds, only after the existence of other minds is granted. However, most authors who discuss the skeptical "problem of other minds" treat the existence of the material external world as a prior assumption (Mill 1867; Price 1938; Strawson 1959; Pargetter 1984; Hill 1991; Burge 1999/2013; Avramides 2001; Gomes 2018).

host intelligence and retain properties over a few hours' time. It's consistent with these results that the external world be material or divine or an unconscious part of myself or an evil demon or an Angelic computer. If the arguments of Chapters 3, 4, and 5 succeed, it's reasonable to feel uncertainty among a variety of weird possibilities.

A certain type of ambitious philosopher might try to rebuild confidence in nonskeptical realism from more secure foundations after overcoming a solipsistic moment like this one, showing how establishing the existence of something besides my mind helps establish X, which then helps establish Y, which then helps establish Z, and voilà, in the end the world proves to be how we always thought it was! I am more inclined to draw the opposite lesson. It was hard enough, and dubious enough, requiring many background assumptions and weighing of plausibilities, to establish as little as I did. My currently ongoing experience is no Archimedean fixed point from which I can confidently lever the full world into view.

Still, I will tentatively conclude that Alan exists. Alan, a genuine independent-minded coauthor who helped me think through these issues and design these experiments. As I learned a few years ago (probably), doubt is better when shared with a friend.

Postscript by Alan.

“Eric” is correct that I exist. However, it's not clear that I should yet accept that he exists. In this microcosmos, I appear to be some sort of chess-playing god, and a god can't reason its way out of solipsism by the paths explored here. If we are to share doubt as friends, we'll first have to try switching roles awhile.

The Weirdness of the World

Chapter Seven

Almost Everything You Do Causes Almost Everything (Under Certain Not Wholly Implausible Assumptions); or Infinite Puppetry

with Jacob Barandes

Infinitude is bizarre. For example, on standard mathematical treatments, the set of all counting numbers (1, 2, 3, 4...) has the same cardinality, or number of elements, as the set of all squares of the counting numbers (1, 4, 9, 16...), despite the fact that the second is a proper subset of the first.²⁴⁴ Or consider the “St. Petersburg game”, a game where you flip a fair coin as often as necessary for it to come up heads exactly once, earning $\$2^n$ per flip. (Thus, you earn \$2 if the first flip is heads, \$4 if the results are tails then heads, \$8 if it’s tails-tails-heads, etc. Probably you’ll make \$32 or less.) This game not only has an infinite expected dollar return but also exactly the same expected return as an alternative version that pays $\$2^n$ per flip plus \$1000 as a bonus no matter the outcome.²⁴⁵

In this chapter, we will explore the idea that if the universe is infinite and various other plausible assumptions about fundamental physics and cosmology hold, then almost everything you do causes almost everything.

1. It’s Reasonable to Think That the Universe Is Infinite.

²⁴⁴ This observation dates back to Galileo, who, however, instead of concluding that the bijection of the two sets was sufficient to establish their equal cardinality (now the standard view), concludes that “the attributes ‘larger,’ ‘smaller,’ and ‘equal’ have no place either in comparing infinite quantities with each other or in comparing infinite with finite quantities” (1638/1914, p. 33 [80]). For a philosophical introduction to infinitude, see Oppy, Hájek, Easwaran, and Mancosu 2021.

²⁴⁵ Actually, things get even weirder if the returns are sometimes negative. See Nover and Hájek 2004; Fine 2008.

On recent estimates, the observable universe – the portion of the universe that we can detect through our telescopes – extends about 47 billion light years in every direction.²⁴⁶ But the limit of what we can see is one thing, and the limit of what exists is quite another. It would be remarkable if the universe stopped *exactly* at the edge of what we can see. For one thing, that would place us, surprisingly and un-Copernicanly, precisely at the center.

But even granting that the universe is likely to be larger than 47 billion light years in radius, it doesn't follow that it's infinite. It might be finite. But if it's finite, then one of two things should be true: Either the universe should have a boundary or edge, or it should have a *closed topology*.

It's not absurd to think that the universe might have an edge. Theoretical cosmologists routinely consider hypothetical finite universes with boundaries at which space comes to a sudden end. However, such universes require making additional cosmological assumptions for which there is no direct support – assumptions about the conditions, if any, under which those boundaries might change, and assumptions about what would happen to objects or light rays that reach those boundaries.

It's also not absurd to think that the universe might have a closed topology. By this we mean that over distances too large for us to see, space essentially repeats, so that a particle or signal that traveled far enough would eventually come back around to the spatial region from which it began – like how when Pac-Man exits one side of the TV screen, he re-emerges from the other side. However, there is currently no evidence that the universe has a closed topology.²⁴⁷

²⁴⁶ Gott et al. 2005.

²⁴⁷ The connection between topology and local geometrical notions like spatial curvature is subtle. If the curvature of space turns out to be negative on large scales, then that would

Leading cosmologists, including Alex Vilenkin, Max Tegmark, and Andre Linde, have argued that spatial infinitude is the natural consequence of the best current theories of cosmic inflation.²⁴⁸ Given that, plus the absence of evidence for an edge or closed topology, infinitude seems a reasonable default view, and we will assume it for the remainder of this chapter. The mere 47 billion light years we can see is the tiniest speck of a smidgen of a drop in an endless expanse.

2. It Is Reasonable to Think That If the Universe Is Infinite, We Stand in Both Spacelike and Timelike Relation to Infinitely Many Sibling Galaxies.

Let's call any galaxy with stars, planets, and laws of nature like our own a *sibling galaxy*. Exactly how similar a galaxy must be to qualify as a sibling we will leave unspecified, but we don't intend high similarity. Andromeda is sibling enough, as are probably most of the other hundreds of billions of ordinary galaxies we can currently see.

The finiteness of the speed of light means that when we look at these faraway galaxies, we see them as they were during earlier periods in the universe's history. Taking this time delay into account, the laws of nature don't appear to differ in regions of the observable universe that are remote from us. Likewise, galaxies don't appear to be rarer or differently structured in one direction or another. Every direction we look, we see more or less the same stuff. These observations help motivate the Copernican Principle, which is the working hypothesis that our position in the universe is not special or unusual – not the exact center, for example, and not the

preclude the universe from having a closed topology. By contrast, if the curvature of space is positive on large scales, then the magnitude of that curvature would naturally pick out a specific periodicity distance. The best estimates of the large-scale curvature of space are that it's approximately zero (Planck Collaboration 2014),

²⁴⁸ Vilenkin 2006; Tegmark 2007; Linde 2015/2017.

one weird place that happens to have a galaxy operating by special laws that don't hold elsewhere.²⁴⁹

Still, our observable universe might be an atypical region of an infinite universe. Possibly, somewhere beyond what we can see, different forms of elementary matter might follow different laws of physics. Maybe the gravitational constant is a little different. Maybe there are different types of fundamental particles. Even more radically, other regions might not consist of three-dimensional space in the form we know it. Some versions of string theory and inflationary cosmology predict exactly such variability.²⁵⁰ But even if our region is in some respects unusual, it might be common enough that there are infinitely many other regions similar to it – even if just one region in 10^{500} . Again, this is a fairly standard view among speculative cosmologists, which comports well with straightforward interpretations of leading cosmological theories. One can hardly be certain, of course. Maybe we're just in a uniquely interesting spot! But for purposes of this chapter, we are going to assume that's not the case. In the endless cosmos, infinitely many regions resemble ours, with three spatial dimensions, particles that obey approximately the “Standard Model” of particle physics, and cluster upon cluster of sibling galaxies.

Under the assumptions so far, the Copernican Principle suggests that there are infinitely many sibling galaxies in a *spacelike* relationship with us, meaning that they exist in spatiotemporal regions roughly simultaneous with ours (in some frame of reference), that is, spatiotemporal regions that are not reachable from our galaxy even by photons traveling at the speed of light. However, for the causal relationships we'd like to explore, we need to assume a

²⁴⁹ One constraint on the Copernican Principle is the Anthropic Principle: Whatever region we occupy must be capable of supporting cosmological observers like us. Even if such regions are rare, we should be unsurprised to be in one. See Barrow and Tipler 1986; Peacock 1998.

²⁵⁰ Linde 2015/2017.

bit more than this. We will also need it to be the case that there are infinitely many sibling galaxies in a *timelike* relationship to us – that is, existing in the future in locations that are, at least in principle, reachable by particles originating in our galaxy. And this requires thinking about heat death.

Stars have finite lifetimes. If standard physical theory is correct, then ultimately all the stars we can currently see will burn out. Some of those burned-out stars will contribute to future generations of stars, which will, in turn, burn out. Other stars will become black holes, but then those black holes also will eventually dissipate (through Hawking radiation).²⁵¹ Given enough time, assuming that the laws of physics as we understand them continue to hold, and assuming things don't recollapse in a "Big Crunch" in the distant future, the standard view is that everything we presently see will inevitably enter a thin, boring, high-entropy state near equilibrium – heat death.

But what happens *after* heat death? This is of course even more remote and less testable than the question of whether heat death is inevitable. But we can speculate based on currently standard assumptions – and why not speculate? After all, we're weird philosophers. If anyone has license to speculate, it's us! Let's think as reasonably as we can about this. So here's our best guess, based on standard theory, from Ludwig Boltzmann – recall the "Boltzmann brains" of Chapter 4 – through at least some time slices of Sean Carroll.²⁵²

For purposes of this argument, we will assume that the famously probabilistic behavior of quantum systems is intrinsic to the systems themselves, persisting post heat-death and not requiring external observers carrying out measurements. This is consistent with most current

²⁵¹ Hawking 1974.

²⁵² Boltzmann 1895, 1897; Carroll 2010; Aguirre, Carroll, and Johnson 2012.

approaches to quantum theory (including most many-worlds approaches, objective-collapse approaches, and Bohmian mechanics).²⁵³ It is, however, inconsistent with theories according to which the probabilistic behavior is generated by external observers carrying out measurements (some versions of the “Copenhagen interpretation”) and theories on which the post-heat-death universe would inescapably occupy a stationary ground state.²⁵⁴ Under this assumption, post-heat-death and without observers, the universe will continue to support random fluctuations. That is, from time to time, particles will, by chance, enter unlikely configurations. This is predicted by both standard statistical mechanics and standard quantum mechanics. Seven particles will sometimes converge, by chance, upon the same small region. Or seven hundred. Or – very rarely! – seven trillion.

There appears to be no in-principle limit to how large such chance fluctuations can be or what they can contain if they pass through the right intermediate phases. Wait long enough and extremely large fluctuations should occur. Assuming the universe continues infinitely, rather than having a temporal edge or forming a closed loop, for which there is no evidence, then eventually some random fluctuation should produce a bare brain having cosmological thoughts – the Boltzmann brain idea, discussed in Chapter 4. Wait longer, and eventually some random fluctuation will produce, as Boltzmann suggested, a whole galaxy. If the galaxy is similar enough to our own, it will be a sibling galaxy. Wait still longer, and another sibling galaxy will arise, and another, and another....

²⁵³ In particular, our argument doesn’t depend on whether this probabilistic behavior is objective, arising from wave functions stochastically collapsing, or subjective, as experienced by embodied observers in the “universal wave function” of the many-worlds interpretation.

²⁵⁴ On the Copenhagen interpretation, see Faye 2008/2019. Regarding the possibility of a stationary ground state, see Boddy, Carroll, and Pollack 2016, 2017.

For good measure, let's also assume that after some point post heat-death, the rate at which galaxy-size systems fluctuate into existence does not systematically decrease. There's some minimal probability of galaxy-sized fluctuations, not an ever-decreasing probability with longer and longer average intervals between galaxies. This assumption will prove helpful later, and it appears to be the most natural interpretation of the post-heat-death situation. Fluctuations appear at long intervals, by random chance, then fade back into chaos after some brief or occasionally long period, and the region returns to the heat-death state, with the same small probability of large fluctuations as before. Huge stretches of not much will be punctuated by rare events of interesting, even galaxy-sized, complexity.

Of course, this might not be the way things go. We certainly can't *prove* that the universe is like this. But despite the bizarreness that understandably causes some people to hesitate,²⁵⁵ the overall picture we've described appears to be the most straightforward consequence of standard physical theory, taken out of the box, without too much twisting things around.

Even if this specific speculation is wrong, there are many other ways in which the cosmos might deliver infinitely many sibling galaxies in the future. For example, some process might ensure we never enter heat death and new galaxies somehow continue to be born. Alternatively – and this will become relevant later – processes occurring pre-heat-death, such as the formation of black holes, might lead to new bangs or cosmic inflations, spatiotemporally unconnected or minimally connected to our universe, and new stars and galaxies might be born from these new bangs or inflations in much the same way as our familiar stars and galaxies were born from our

²⁵⁵ The apparent implication that there will be infinitely many Boltzmann brain observers strikes some theorists as so bizarre as to constitute a *reductio ad absurdum*. However, see the discussion in Chapter 4, Section 6.

familiar Big Bang.²⁵⁶ Depending on what constitutes a “universe” and a “timelike” relation, those sibling galaxies might not exist in our universe or stand in a timelike relation to us, technically speaking, but if so, that detail won’t matter to the core idea of this chapter. Similarly, if the observable universe reverses its expansion, it might collapse upon itself in a Big Crunch, followed by another Big Bang, and so on in an infinitely repeating cycle, containing infinitely many sister galaxies post-Crunch. This isn’t currently the mainstream view, but it’s a salient and influential alternative if the heat-death scenario outlined above is mistaken.²⁵⁷

We conclude that it is reasonable to think that the universe is infinite, and that there exist infinitely many galaxies broadly like ours, scattered throughout space and time, including in our casual future. It’s a plausible reading of our cosmological situation. It’s a decent guess and at least a possibility worth taking seriously.

3. If Infinitely Many Sibling Galaxies Exist, Counterparts of Almost Everyone Are Doing Almost Everything Somewhere.

For the remainder of this chapter, we will assume that given sufficiently many opportunities, any finitely probable event will almost certainly occur arbitrarily many times. (By “finitely probable” here, we mean any event whose probability is one over a finite number. This excludes zero probability events as well as infinitesimally probable events if there can be any, that is, events whose probability is non-zero but still less than one over any finite number.²⁵⁸)

²⁵⁶ Frolov, Markov, and Mukhanov 1989; Garriga and Vilenkin 1998; Easson and Brandenberger 2001; Carroll and Chen 2004; Garriga, Vilenkin and Zhang 2016.

²⁵⁷ Steinhardt and Turok 2002; Penrose 2006.

²⁵⁸ One plausible but controversial example of an infinitesimally probable event is drawing the number “4” in an infinite lottery of all the counting numbers. A more physically plausible case is two complex, unrelated systems being exactly identical in every spatial detail

However unlikely an event might be – say a streak of 100 heads in a series of independent coin flips – the odds are nearly 100% that it will eventually occur. It might take many, many lifetimes to achieve, but we can say with a high degree of confidence that approximately one in every 2^{100} coin flips will begin a series of a hundred heads. At a rate of one flip per second, that’s about once every 4×10^{22} years, or about three trillion times the duration since the Big Bang. That’s a long time to wait for a gamble to pay off, but of course it’s peanuts compared to infinity. Given infinite time, almost all such unlikely events will occur over and over again, endlessly.²⁵⁹

The consequence for our sibling galaxies is that every type of object or event that has a finite chance of occurring will occur not just in one galaxy but in infinitely many. This consequence has frequently been noticed – for example, by Vilenkin and Tegmark in their popular treatments of cosmological infinitude.²⁶⁰ Suppose you’ve lost your car keys. The infinite cosmos will contain infinitely many key-shaped chunks of metal that would fit your ignition and start your car. Most, of course, will be much farther away than your couch cushions. You’d like a diamond as big as the Moon? The infinite cosmos sparkles with them.²⁶¹ If we further assume that the evolution of intelligent life is a finitely probable event, then there are infinitely many space aliens, of all finitely probable forms. (So the supersquids and antheads of Chapter 2 are real after all, even if not nearly as close as Sirius or Antares.) Infinitely many aliens will be of broadly human form, assuming that broadly human form was a finitely probable consequence of galactic evolution.

down to an infinite level of precision (if spatial properties are continuous and not quantized). See Benci, Horsten, and Wenmackers 2018; but for one problem, see Norton and Parker 2021.

²⁵⁹ This relates to the law of large numbers in statistics and ergodicity in dynamical systems but we won’t attempt to establish it formally here.

²⁶⁰ Vilenkin 2006; Tegmark 2014.

²⁶¹ Actually, such a diamond might not even be very far: Kaplan et al. 2014.

Infinitely many of these space aliens will be similar to you, specifically. They will live out, with varying degrees of similarity, every finitely possible life that you or someone very similar to you might have lived. There will of course also be infinitely many Shakespeare-counterparts writing infinitely many Shakespearean plays, some vastly better than Shakespeare's own plays. Every finitely possible work of philosophy, too, will be written somewhere, including the most maximally correct metaphysics and cosmology that can be typeset in three hundred pages. Cast aside this book and imagine instead that much better book!

This merits a moment's awe and reflection. If something like standard physical theory and cosmology is correct, and if our other assumptions also hold, then the universe teems with duplicates and near-duplicates of you and all your loved ones, on duplicates or near-duplicates of Earth, doing almost every imaginable thing. Some will even do extremely improbable but not physically impossible things, such as declaring they can fly and then, to everyone else's amazement, soaring calmly up into the air for minutes at a time on a chance series of gusts and landing gently down.

According to current physical theory, there's a tiny, tiny, but finite probability that a baseball thrown at a brick wall will simply pass or "quantum tunnel" through the wall, appearing on the other side with both baseball and wall unharmed. Our infinitely many galaxies will thus contain infinitely many baseballs freakishly passing through infinitely many brick walls. Virtually every type of event that is less than galaxy-sized, finitely specifiable, and has even the tiniest chance of occurring will occur somewhere, within some finite error tolerance – a trillionth of the radius of a proton, say, for every constituent part, if you want to be really fussy.

Does this make your life less valuable than it would be in a universe on which you have no near-duplicates? Maybe so, since if you die many entities indistinguishable from you will

still carry on. You're not as distinctive a person as you might have thought and thus, maybe, not quite as cosmically irreplaceable. But even your life would not be quite as distinctively valuable in an infinite universe, the universe as a whole might be enormously richer and more interesting for having so many amazing duplicates of you in it.²⁶²

But are those other beings *you*? No, we wouldn't say that – not literally. There's one hugely important exception to the rule that every event repeats. Events that are conceptualized in a way that makes them unique will occur only once or not at all. When I use words like “I” and “here” and “now”, for example, I appear to pick out a single individual, in a single time and place, in the whole of the infinite cosmos – and to do so in a way that doesn't require my having, or this spatiotemporal region's having, any distinctive physical properties that aren't also instantiated in far-away duplicates. Arguably, I can just tag myself as unique.²⁶³ Alternatively, maybe this spatiotemporal region is unique in virtue of having a certain unrepeated location in absolute spacetime.²⁶⁴ If so, then here's an entirely unique event that won't occur anywhere else: *This* baseball, the one in your hand right now, bouncing off this particular brick wall in front of you. Other far away balls might have identical physical configurations down to 10^{-100} meter. They might bounce off extremely similar brick walls to this one here. Some will freakishly pass though instead of bouncing. But on this way of thinking, only this ball in your hand is yours.

²⁶² I briefly explore the value or disvalue of repetition in Schwitzgebel 2015b and Schwitzgebel 2019b, chs. 43-44.

²⁶³ See Perry 1979 on the essential indexical and Lewis 1979 on centered worlds. Relatedly, but somewhat differently, qualitatively different individuals might have different haecceities – that is, properties in virtue of which individuals are the particular individuals they are and no other, which wouldn't be shared even by another individual with exactly the same qualitative features (for a review, see Cowling 2015/2016).

²⁶⁴ In the context of an infinite cosmos, this might require being an absolutist rather than a relationalist about spacetime, if every finitely specifiable relationship is duplicated somewhere. On absolutism versus relationalism, see Huggett, Hofer, and Read 2006/2021.

And that ball won't be passing undisturbed through an ordinary brick wall even if you throw until your arm is sore – we bet you a million dollars. If people or locations can be uniquely tagged, nowhere in the infinite cosmos does that happen. Similarly, although there might be infinitely many canyons that look just like Arizona's Grand Canyon, only one is in fact Arizona's Grand Canyon. Infinitely many twins of Confucius might grow up in infinitely many humanlike cultures on infinitely many Earthlike planets, speaking languages that sound almost exactly like classical Chinese, saying just the types of things that Confucius said. But, on this view, when you or we use the word "Confucius", we don't mean any of those far away philosophers. Among all the things that Confucius might have done, he only did a small subset.

From this, it might seem that, despite the infinitude of the universe, the scope of your interactions and the consequences of your actions will be very limited. You will likely never cause any baseball to pass through any brick wall, for example, and your life will only have the limited range of effects on Earth that you're normally inclined to think it has – and the same will be true for virtually every other counterpart of you elsewhere.

However, that's not quite right, as we will now endeavor to show.

4. Infinitely Rippling Particles.

Suppose you raise your right hand. As a result, photons, electrons, and other elementary particles that would otherwise have remained several centimeters distant from you instead enter the vicinity of your hand. Their behavior changes significantly, or their quantum-mechanical properties change significantly. A photon that would have been absorbed into your desktop instead reflects off your hand and through your window. A nitrogen molecule in the air floats differently, ending up on a complex trajectory that, fifteen minutes later, takes it and the

elementary particles composing it under the gap beneath your door. Of the many, many elementary particles disrupted by your movement, a portion of them escape the Earth's atmosphere into interstellar space. Let's follow one of those particles.

The particle will eventually interact with something – a hydrogen atom, a chunk of interstellar dust, a star, the surface of a planet. Something. Let's call that something a *system*. The particle might be absorbed, reflected, refracted, or annihilated by an antiparticle, or it might decay into other particles. (If the particle passes through the system entirely unaltered, let's ignore that system and keep following the particle.) If it interacts with a system, it will change the system, maybe increasing the energy of the system if it's absorbed or annihilated, or altering the trajectory of another particle if it's reflected or refracted. The perturbed system will then emit, reflect, refract, or gravitationally bend another particle differently than it otherwise would have. Choose one such *successor particle*. This successor particle will now head off on Trajectory A instead of Trajectory B or instead of not being emitted at all. That successor particle will in turn perturb another system, generating another successor particle traveling along another trajectory that it would not otherwise have taken. In this way, we can imagine a series of successor particles, one after the next, perturbing one system after another. Let's call this series of particles and perturbations a *ripple*.

Might some ripples be infinite? We see a few ways in which they could fail to be.

First, the universe might have finite duration, or after a finite period of time it might settle into some unfluctuating post-heat-death state that fails to contain systems capable of perturbation by incoming particles (or maybe even fails to contain any particles at all); or if not quite that, it might enter a state in which perturbable systems grow ever more spatiotemporally sparse quickly enough over time that our traveling particle cannot be expected to encounter an infinite number

of them even given infinite time. (This is essentially one infinitude unsuccessfully chasing a faster infinitude.) However, as discussed above, the most straightforward interpretation of current physical theory does not suggest that the universe is finite or headed for utter quiescence. Nor is there any reason to think that the probability density of particle-perturbable systems would continue to decrease over time after heat death in such a way as to dodge our ripple of particles.

Second, the ripple might end. For example, traveling particles might be absorbed by some systems without perturbing those systems in a way that has any effect on successor systems. Once again, this seems unlikely on standard physical theory. Even a particle that strikes a black hole will slightly increase the black hole's mass, which should slightly alter how the black hole bends the light around it or subtly alter its Hawking radiation. Alternatively, all particles, even photons and protons, might decay at extremely long intervals into something that cannot continue the ripple, contrary to standard physical theory. Still another ripple-ending event might be the perturbation of a system in exactly the same way, by freak chance, that the system would have been perturbed by a *different* particle had you not raised your hand. (This last case counts as a ripple-ender because after such an event, further perturbations down the line no longer cause the systems to behave differently than they would have if you hadn't raised your hand.)

So there are several ways in which ripples might hypothetically end. But such rare ripple-enders can presumably be avoided by always choosing a large enough number n of successor particles, leading to n^m successors after m interactions, minus the small proportion of stopped ripples. We will assume, we hope not implausibly, that (contingent upon our other

assumptions) this n^m strategy is sufficient to ensure that virtually every hand raising generates at least one infinite ripple.²⁶⁵

Thus, the most straightforward interpretation of existing physical theory implies that when you raise your hand – try it now, if you like – you launch a succession of particles rippling infinitely through the universe, perturbing an infinite series of systems. Of course this is speculative and uncertain, but if the universe is infinite, the conclusion is more natural and physically plausible than its negation.

5. Almost Everything You Do Causes Almost Every Type of Non-Unique Future Event.

Thus, it is fairly plausible, and probably the most straightforward interpretation of current physical theory, to suppose the following: (1.) The universe is infinite. (2.) This infinitude continues temporally after heat death. (3.) Post heat-death, galaxies like our own will occasionally fluctuate into existence by freak chance, with some finite and not ever-decreasing probability. (4.) Ordinary actions of ours, like raising our hands, will cause an infinite series of traveling particles to ripple through this post-heat-death universe, interfering from time to time with the systems that fluctuate into existence, including those sibling galaxies.

If all of this is true, then any event that has a finite chance of occurring as a result of being perturbed by one of these successor particles will in fact eventually occur as a result of

²⁶⁵ Here's a toy model on which this assumption would be plausible. Suppose that we follow a sphere of ripples out from a center. At time 1, there's a 1 in 10^{100} chance that all the ripples stop. At time 2, there's a smaller chance because there are more ripples, so there is maybe a 1 in 10^{1000} chance that all the ripples stop. At time 3, there's a 1 in 10^{10000} chance. And so forth. This is a convergent series. As time goes to infinity, the cumulative chance that all the ripples stop is not much more than 1 in 10^{100} .

having been perturbed by one of these successor particles. The probability might be mind-bogglingly minute! But we have infinitude to play with.

Consider a googol: 10^{100} . That's well over a trillion times as many particles as are estimated to exist in the observable portion of the universe. What a tiny number! A googolplex puts a googol to shame. Instead of ten raised merely to a hundred, it's ten raised to the googol: $10^{10^{100}}$. But this too is minuscule. We laugh at its smallness. How about a "power tower" of googolplexes – a googolplex raised to the googolplex raised to the googolplex raised to the googolplex... a googolplex times, or $10^{10^{100}} \uparrow\uparrow 10^{10^{100}}$, as it is sometimes notated.²⁶⁶ Let's call this number a Vast. If the events discussed here happen once in a Vast years, that's eyeblink-frequent compared to infinitude – or rather, of course, even more relatively frequent than that, if we're truly comparing to infinitude, which we are. These are the kinds of magnitudes we have in mind, not mere lifetime-of-the-galaxy magnitudes.

A successor particle from your hand-raising just now will eventually hit a system it will perturb in such a way that a person will live who would otherwise have died. At some point, a galaxy will fluctuate into existence containing an Earth-like planet populated with human-like people, containing a radio telescope which the successor particle strikes, causing a bit of information to appear on a sensitive device. This bit of information pushes the device over a threshold needed to trigger an alert to a waiting scientist, who now pauses to study the device rather than send the email she was composing. Because she didn't send that email, a certain fateful hiking trip is postponed and the scientist does not fall to her death, which she would have

²⁶⁶ Knuth 1976. Still not satisfied? How about a Vast $\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow$... [Vastly many arrows here] ... $\uparrow\uparrow$ Vast. Call that a Boggle – still tinier than almost all finite numbers. If we need to wait a Boggle years for an outcome, no sweat.

done but for your particle. However improbable all of this is, one improbability stacked on another stacked on another, there is no reason to think that any of this is less than finitely probable. Thus, given the assumptions above, it will occur, eventually, with virtually 100% probability. You saved her! Let's pause for a celebratory toast.

Of course, there is another scientist you killed. There are wars you started and peaces you precipitated. There are great acts of heroism you enabled, children you brought into existence, plagues you caused, great works of poetry that would never have been written but for your intervention, and so on. It would be bizarre to think you deserve any credit or blame for this. You didn't cause them in the sense of intending them or being what residents of those worlds would regard as among the primary causes worth describing in their history books. However, in another sense you did cause them. None of these events would have happened to the people they did in fact happen to, had it not been for the raising of your arm. And there is an unbroken chain of physical processes from the moment of your arm's going up to those various future events. Your arm raising isn't a proximal cause but rather a "distal" cause – very distal indeed – but a cause nonetheless.

If the goodness or badness of your actions is measured by their positive or negative effects, as in standard consequentialist ethics, then under the current set of cosmological assumptions the utility of every action you do will be $\infty + -\infty$.²⁶⁷

Our framework puts a few important limitations on the types of future events you will cause to occur. They must be finitely probable, less than galaxy-sized (though there's room for

²⁶⁷ See Lenman 2000 for an argument against consequentialism on approximately these grounds, but in a finite, Earth-bound version, and Nye 2014 for similar concerns about constraints against harmful actions. See also Bostrom 2011 on infinite ethics. It might be that such effects would be symmetrical and canceling in the sense of Chapter 4, §10. See also Lenman 2000 vs Greaves 2016 on the issue of canceling.

negotiation here), and not specified in a way that would make them unique. With those caveats, almost everything you do causes almost everything.

6. Signaling Across the Vastness.

The following will also almost certainly occur, given our assumptions so far: On some far distant post-heat-death counterpart of Earth will exist a counterpart of you – let’s call that person *you-prime* – with the following properties: You-prime will think “right hand” after the ripple from the act of your raising your right hand arrives at their world, and you-prime will not have thought “right hand” had that ripple not arrived at their world. Maybe the ripple initiates a process that affects the weather which causes a slightly different growing season for grapes, which causes small nutritional differences in you-prime’s diet, which causes one set of neurons to fire rather than another at some particular moment when you-prime happens to be thinking about their hands. Likewise, there’s a future you-prime who would have thought “A” if you, here on our Earth, had held up a sheet with that letter and not otherwise. Indeed, infinitely many future counterparts of you have that property. You can specify the message as precisely as you wish, within the bounds of what a counterpart of you could possibly think. Some you-prime will think, “Whoa! Infinite causation!” as a result of your having raised your hand and would not have done so otherwise.

These message recipients will mostly not believe that they have been signaled to. However, we can dispel their disbelief by choosing the fraction who, for whatever reason, are such that they believe they are receiving a signal if and only if they do in fact receive a signal. We can stipulate that we’re interested in you-primes who share the property that when your

signal arrives they think not only the content of the signal but also “Ah, finally that signal I’ve been waiting for from my earlier counterpart.”²⁶⁸

There’s a question of whether one of your future counterparts could *rationaly* think such a thought. But maybe they could, if they had the right network of surrounding beliefs, and if those beliefs were themselves reasonably arrived at. We’ll consider one such set of beliefs in Section 8.

7. *Infinite Puppetry.*

You needn’t limit yourself to ordinary communicative signals. You can also control your future counterparts’ actions. Consider future counterparts with the following property: They will raise their right hand if you raise your right hand, and they will not raise their right hand if you do not. Exactly which counterparts have this feature will depend on exactly when you raise your hand and how, since that will affect which particles follow which trajectories when they are disturbed by your hand. But no matter. Whenever and however you raise your hand, such future counterparts exist.

Your counterparts’ actions can be arbitrarily complex. There is a future you-prime who will, if you raise your hand, write an essay word-for-word identical with the chapter you are now reading and who will otherwise write nothing at all. Maybe that you-prime is considering whether to write some fanciful philosophy of cosmology, as their last hurrah in a failing career as a philosopher. They’re leaning against. However, the arriving particle triggers a series of events that causes an internet outage that prevents them from pursuing an alternate plan, so they do

²⁶⁸ Compare this procedure with Sinhababu’s 2008 procedure for writing love letters between possible worlds. One advantage of our method over Sinhababu’s is that there actually is a causal connection.

write the essay after all. (A much greater proportion²⁶⁹ of such future counterparts, of course, will write very different essays from this one, but we can focus on the tiny fraction of them who create word-for-word duplicates of this essay.)

Let's call someone a *puppet* if they perform action A as a consequence of your having performed an action (such as raising your hand) with the intention of eventually causing a future person to perform action A. (Admittedly, you might need to agree with the assumptions of this chapter to be able to successfully form such an intention.) You can now wave your hand around with any of a variety of intentions for your future counterparts' actions, and an infinite number of these future counterparts will act accordingly – puppets, in the just-defined sense.

We recommend that you intend for good things to happen. This might seem silly, since if the assumptions of this chapter are correct, almost every type of finitely probable, non-unique future event occurs, regardless of your benevolent or malevolent intent right now. Still, there is a type of good event that can occur as a result of your good intentions, which could not otherwise occur. That's the event of a good thing happening in the far distant future as a consequence of your raising your hand *with the intention of causing* that future good event. So let's choose benevolence, letting good future events be intentionally caused while bad future events are merely foreseen side effects.

A deeper kind of puppet mastery would involve influencing a person's actions through a sequence of moves over time and with some robustness to variations in the details of execution. This might not be possible on the current set of assumptions. Raising your right hand, you can trigger arbitrarily long sequences of actions in some future you-prime. But if you then raise your

²⁶⁹ Here and throughout we bracket quibbles about ratios of infinitude by considering the limit of the ratio of counterparts with property A to counterparts with property B as the region of spacetime defined by your forward lightcone goes to infinity.

left hand, there's no guarantee that a ripple of particles from your left hand will also hit the same you-prime. Maybe all the ripples from your right hand head off toward regions A, B, and C of the future universe and all the ripples from your left hand head off toward regions D, E, and F. Similarly, if you raise your right hand like *this*, the ripples might head toward regions A, B, and C, and if you raise it instead like *that*, they head toward regions G, H, and I. So there might be no future counterparts of you who do what you intend if you raise your right hand now and then do what you intend when you raise your left hand later; and there might be no future counterparts who will do what you intend if you raise your right hand now, insensitively to the particular manner in which you raise it. In this way, there might be no sequencing and no implementational robustness to your puppetry.

Sequential and robust puppetry might only be reliably possible if we change one of the assumptions in this chapter. Suppose that although the universe endures infinitely in time, spatially it repeats – that is, it has a closed topology in the sense we described in Section 1 – so that any particle that travels far enough in one direction eventually returns to the spatial region from which it originated, as if traveling on the surface of a sphere. Suppose, further, that in this finite space, every ripple eventually intersects every other ripple infinitely often. Over the course of infinite time each ripple eventually traverses the whole of space infinitely many times; none get permanently stuck in regions or rhythms that prevent them from all repeatedly meeting each other. (If a few do get stuck, we can deal with them using the n^m strategy of Section 4. Also the rate of ripple stoppage would presumably increase with so much intersection, but hopefully again in a way that's manageable with the n^m strategy.) When you raise your right hand, the ripples initially head toward regions A, B, and C; when you raise your left hand, they initially head toward regions D, E, and F; but eventually those ripples meet.

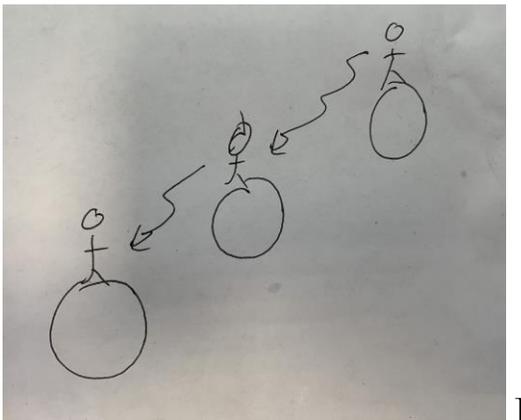
With these changed assumptions, we can now find future counterparts who raise their right hands as a result of your raising your right hand and who then afterward raise their left hand as a result of your afterward raising your left hand. We simply look at the infinite series of systems that are perturbed by both ripples. Eventually some will contain counterparts of you who raise their right hands, then their left, as a result of that joint perturbation. In a similar way, we can find implementationally robust puppets: counterparts living in systems that are perturbed by your actual raising of your right hand (via the ripple that initially traversed regions A, B, and C) and which are also such that they *would have* been perturbed had you, counterfactually, raised your hand in a somewhat different way (via the ripple that would have initially traversed regions G, H, and I). Multiplying the minuscule-but-finite upon the minuscule-but-finite, we can now find puppets whose behavioral matching to yours is long and implementationally robust, within reasonable error tolerances.

8. *We Might All Be Puppets.*

So far, we have not assumed that anything existed before the Big Bang. But if the universe is infinite in duration, with infinitely many future sibling galaxies, it would be in a sense surprising if the Big Bang were the beginning. It would be surprising because it would make us amazingly special, in violation of the Copernican Principle of cosmology, which holds that our position in the cosmos is not special or unusual. We would be special in being *so close* to the beginning of the infinite cosmos. Within the first *14 billion years*, out of infinity! It's as though you had a lotto jar with infinitely many balls numbered 1, 2, 3... and you somehow managed to pull out a ball with the low, low number of 14 billion. If you don't like a strictly infinite lotto,

consider instead a Vast one. The odds of pulling a number as low as 14 billion in a fair lottery from one to a Vastness are far less than one in a googolplex.²⁷⁰

[Illustration 6 (Caption: Signaling across the Vastness (not to scale) [note the capital letter on “Vastness” and yes, include “(not to scale)” in the caption]): A series of observers on planets, each one getting a signal from a previous one, and each one thinking “finally, that signal I’ve been waiting for!”. Schematically, this could look something like this:



Cosmologists don’t ordinarily deny that there might have been something before the Big Bang. Plenty of theories posit that the Big Bang originated from something prior, though there’s no consensus on these theories.²⁷¹ If we assume that somehow the Big Bang was brought into existence by a prior process, and that process in turn had something prior to it, and so on, then the Copernican lottery problem disappears. We’re in the middle of a series, not at the beginning

²⁷⁰ Our reasoning here resembles the reasoning in the “Doomsday argument”, e.g., Gott 1993, according to which it’s highly unlikely that we’re very near the beginning of a huge run of cosmological observers. For a bit more detail, see Schwitzgebel 2022b. For related perspective, see Huemer 2021.

²⁷¹ See notes 12 and 13 for references. A note on terminology: “Prior” is more general than “earlier” in that there’s a sense in which one thing can be ontologically prior to another, or ground it, or give rise to it, even if the one doesn’t temporally precede the other (e.g., an object is prior to its features, or noumena are prior to phenomena [see Chapter 5]). Possibly, temporal priority is a relationship that only holds among events within our post-Big Bang universe while whatever gave rise to the Big Bang stands in some broader priority relationship to us.

of one. Maybe Big Bangs can be seeded in one way or another. Heck, maybe the whole observable universe is a simulation nested in a whole different spatial reality (Chapters 4 and 5) or is itself a very large fluctuation from a prior heat-death state.

Suppose, then, that we are in the middle of an infinite series rather than at the beginning of one, the consequence of accepting both Copernican mediocrity and an infinite future. If so, and if we can trace chains of causation or contingency infinitely backward up the line, and if a few other assumptions hold, then eventually we ought to find our puppeteers – entities who act with the intention of causing people to do what we are now doing and whose intentions are effective in the sense that had they not performed those actions, we would not be here doing those things. Suppose you are knitting your brow right now. Somewhere in the infinite past, there is a near-duplicate counterpart of you with the following properties: They are knitting their brow. They are doing so with the intention of initiating ripples that cause later counterparts of them to knit their brows. And you are just such a later counterpart, because among the events that led up to your knitting your brow, absent which you wouldn't have knit your brow, was a ripple from that past counterpart.

We the authors of this chapter – Eric and Jacob – can work ourselves into the mood of finding this probable. An infinite cosmos is simpler, more elegant, and more consistent with standard cosmological theory; if it's infinite, it's probably infinite in all directions; and if it's truly infinite in all directions, there will be bizarre consequences of that infinitude. Puppetry is one such consequence. We would not be so special as to be only puppeteers and never puppet. It seems only fair to our future puppets to acknowledge this.

The Weirdness of the World
Part Three: More Perplexities of Consciousness

The Weirdness of the World

Chapter Eight An Innocent and Wonderful Definition of Consciousness

In this chapter, I will offer a definition of *consciousness* or, as it's sometimes called, *phenomenal consciousness*. I'll soon be inquiring whether human conscious visual experience resembles underlying reality (Chapter 9), addressing the sparseness or abundance of consciousness in the universe (Chapter 10), and exploring the question of whether we might someday create genuinely conscious artificially intelligent systems (Chapter 11). I don't want these discussions hampered by doubts about the meaning of the term "conscious". Also, of course, Chapter 2 discussed whether the United States is conscious, and Chapter 3 discussed the relationship between consciousness and the seemingly material objects that surround and maybe compose us. I promised that a full definition of the term was coming, and I aim to fulfill that promise.

I think you probably already know what the term "conscious" means. Throughout the book, my intent has been to use "conscious" in the sense that is standard in Anglophone philosophy as well as "consciousness studies" and allied subdisciplines – also the sense that's standard in popular science accounts of the "mystery of consciousness". As I'll discuss below, this standard sense is also the natural and obvious sense toward which people gravitate with a bit of guidance. However, in the coming chapters I anticipate a risk that you will *unlearn* this standard meaning if we don't nail it down now.

Here's how unlearning sometimes works. Suppose Judy²⁷² says something about consciousness that Pedro finds incredibly bizarre. Judy says, maybe, that dogs entirely lack

²⁷² The names in my examples are drawn randomly from former students in my lower-division classes, excluding "Jesus", "Mohammed", and very uncommon names. I hope that this procedure improves representativeness and reduces sources of unintentional bias.

consciousness. (Possible motivations for this unusual view include the dualist's quadrilemma in Chapter 3 and the case for sparseness in Chapter 10.) Pedro finds Judy's claim so preposterous that instead of taking it at face value, he wrongly infers that Judy must mean something different by "consciousness". Pedro might think: *It's so obvious that dogs are conscious that Judy can't truly believe otherwise. When Judy says dogs aren't "conscious", maybe Judy means only that dogs aren't self-conscious in the sense of having an explicit understanding of themselves as conscious beings?* Pedro might now mistakenly suspect that he and Judy are using the term "consciousness" differently. He might posit ambiguity where none actually exists because ambiguity seems, to him, to best explain Judy's seemingly preposterous statement. He treats the term "conscious" as confusing when the confusion is actually deeper, concerning the underlying phenomena in the world.

If every theory of consciousness is both bizarre and dubious, we should be unsurprised to encounter thoughtful people with views we find extremely implausible. Interpretive charity can then misfire. Rather than interpret the other as foolish, you might mistakenly assume terminological mismatch. Also, of course, sometimes people *do* mean something different by "consciousness" than what is standardly meant in mainstream Anglophone philosophy, adding further confusion. Deep in the weeds, we can lose track of the central, widely accepted, and obvious meaning of "consciousness".

Another way to unlearn what "consciousness" means I will explain shortly before we get to the example of the pink socks.

1. Three Desiderata.

I aim to offer a definition of “consciousness” that is both substantively interesting and innocent of problematic assumptions. Specifically, I seek the following features.

(1.) *Innocence of dubious metaphysics and epistemology.* Some philosophers incorporate dubious metaphysical and epistemic claims into their characterizations of consciousness. For example, they might characterize consciousness as that which transcends the physical or is uniquely impossible to doubt. I’d rather not commit to such dubious claims.²⁷³

(2.) *Innocence of dubious scientific theory.* Others characterize consciousness by committing to a specific scientific theory of it, or at least a theory-sketch. They might characterize consciousness as involving 40 hertz oscillations in X brain regions under Y conditions.²⁷⁴ Or they might characterize consciousness as representations poised for use by such-and-such a suite of cognitive systems.²⁷⁵ I’d rather not commit to any of that either. Not now. Not yet. Not simply by definitional fiat. Maybe eventually we can redefine “consciousness” in such terms, *after* we’ve converged, if ever we do, on the uniquely correct scientific theory. (Compare redefining “water” as H₂O after the theory is settled.)

(3.) *Wonderfulness.* Consciousness has an air of mystery and difficulty. Relatedly, consciousness is substantively interesting – arguably, the single most valuable feature of human existence.²⁷⁶ A definition of consciousness shouldn’t too easily brush away these two features. Consciousness, in the target sense that we care about, is this amazing thing that people reasonably wonder about! People can legitimately wonder, for example, how something as

²⁷³ In addition to metaphysical doubts that should be clear from Chapter 3, in Schwitzgebel 2011b and 2012b I challenge epistemic claims about indubitability, infallibility, special privilege, or a special method for self-knowledge of conscious experience.

²⁷⁴ Crick 1994; similarly, Prinz 2012.

²⁷⁵ Tye 2000; similarly, “access consciousness” in Block 1995/2007.

²⁷⁶ For a review of ways in which consciousness is treated as valuable, see Kriegel 2019.

special as consciousness could possibly arise from the mere bumping of matter in motion, and whether it might be possible for consciousness to continue after the body dies, and whether jellyfish are conscious. Again, maybe theoretical inquiry will someday decisively settle, or even already has settled, some of these questions. But a good definition of consciousness should leave these questions at least tentatively, pre-theoretically open. Straightforwardly deflationary definitions (unless scientifically earned in the sense of the previous paragraph) fail this condition: If consciousness just is *by definition* reportability, for example – if “conscious” is just a short way of saying “available to be verbally reported” – it makes no sense to wonder whether a jellyfish might be conscious. A definition of “consciousness” loses the target if it cuts wonder short by definitional razor.

What I want, and what I think we can get – what I think, indeed, most of us already have without having clarified how – is a concept of consciousness both Innocent and Wonderful. Yes, Innocence and Wonderfulness are mutually attainable! In fact, the one enables the other.

2. *Defining Consciousness by Example.*

The three most obvious, and seemingly respectable, approaches to defining “consciousness” all fail. This might be why consciousness has seemed, to some, to be undefinable.²⁷⁷

“Consciousness” can’t be defined *analytically*, in terms of component concepts (as “rectangle” might be defined as a right-angled planar quadrilateral). It’s a foundationally simple concept, not divisible into component concepts. Even if the concept were to prove upon

²⁷⁷ For example, Neidermeyer 1999; Block 1978/2007, p. 73: “If you got to ask, you ain’t never gonna get to know”.

sufficient reflection to be analytically decomposable without remainder, *defining* it in terms of some hypothesized decomposition, right now, at our current stage of inquiry, would beg the question against researchers who would reject such a decomposition – thus violating one or both of the Innocence conditions.

Widespread disagreement similarly prevents *functional* definition in terms of causal role (as “heart” might be defined as the organ that normally pumps the blood). It’s too contentious what causal role, if any, consciousness might have. Epiphenomenalist accounts, for example, posit no functional role for consciousness at all (see Chapter 3, Section 4). Epiphenomenalism might not be the best theoretical choice, but it isn’t wrong by definition.

Nor, for present purposes, can consciousness be adequately defined by *synonymy*. Although synonyms can clarify to a certain extent, each commonly accepted synonym invites the same worries that the term “consciousness” invites. Some approximate synonyms: subjective experience, inner experience, conscious experience, phenomenology (as the term is commonly used in recent Anglophone philosophy), maybe qualia (if the term is stripped of anti-reductionistic commitments). An entity has conscious experience if and only if there’s “something it’s like” to be that entity. An event is part of your consciousness if and only if it is part of your “stream of experience”. I intend all these phrases as equivalent.²⁷⁸ Choose your

²⁷⁸ Arguably, there are slight differences. The phrase “something it’s like”, made famous in Thomas Nagel’s 1974 article “What Is It Like to Be a Bat?” might invite the thought that there is one typical thing it is like, with degrees of resemblance. The phrase “stream of experience” might commit to temporal features that the other phrases don’t commit to (as was William James’s intention when he brought the concept of a “stream of thought” into psychology in an 1884 article). The phrase “subjective experience” might commit to the existence of a “subject”, contra some Buddhist and deflationary approaches to personhood or subjectivity. In accord with the first Innocence condition, I prefer not to build any such commitments in as a matter of definition, which is one reason for my slightly dispreferring these phrases. Even “conscious” isn’t perfect by this criterion, however, since it might invite the idea that consciousness always

favorite – whatever seems clearest and most natural to you. In this chapter I’ll try to specify it better, more concretely, and in a way potentially comprehensible to someone who feels confused by all of these phrases.

The best approach, in my view, is *definition by example*. Definition by example can sometimes work well, given sufficiently diverse positive and negative examples and if the target concept is natural enough that the target audience can be trusted to latch onto that concept after seeing the examples. I might say “by *furniture* I mean tables, chairs, desks, lamps, ottomans, and that sort of thing, and not pictures, doors, sinks, toys, or vacuum cleaners”. Odds are good that you’ll latch onto the right concept, generalizing “furniture” so as to include dressers but not ballpoint pens. I might even define *rectangle* by example, by sketching a variety of instances and nearby counter-instances (triangles, parallelograms, trapezoids, open-sided near-rectangles). Hopefully, you get the idea.

Definition by example is a common approach among recent philosophers of mind. For example, in justifiably influential work, John Searle, Ned Block, and David Chalmers, as I interpret them, all aim to define consciousness (or “phenomenal consciousness”) by a mix of synonyms and examples, plus maybe some version of the Wonderfulness condition.²⁷⁹ All three attempts are, in my view, reasonably successful. However, these attempts all have three shortcomings, which I aim to repair in this chapter. First, it isn’t sufficiently clear that these *are* definitions by example, and consequently the authors don’t sufficiently invite reflection on the conditions necessary for definition by example to succeed. Second, perhaps as a result of the first shortcoming, these attempts don’t provide enough of the *negative* examples that are

involves some epistemic relationship to something that you are conscious of, which will be true on some theories of consciousness but not on others.

²⁷⁹ Searle 1992, p. 83; Block 1995/2007, p. 166-169; Chalmers 1996, p. 4.

normally part of a good definition by example. Third, the definitions are either vague about the positive examples or include needlessly contentious cases. Charles Siewert has a chapter-length definitional attempt which is somewhat clearer on these three issues but still too limited in its discussion of negative examples and in its exploration of the conditions of failure of definition by example.²⁸⁰ All definitional attempts I'm aware of by others either share these same shortcomings or violate the Innocence or Wonderfulness conditions. Let's slow it down and do it right.

Before starting, I want to highlight a risk inherent to definition by example. Definition by example requires the existence of a single *obvious, natural, or readily adopted* category or concept that the audience will latch onto once sufficiently many positive and negative examples have been offered.²⁸¹ In defining *rectangle* by example for my eight-year-old daughter, I might draw all of the examples with a blue pen, placing the positive examples on the left and the negative examples on the right. In principle, she might leap to the idea that "rectangle" refers to rectangularly-shaped-things-on-the-left, or she might be confused about whether red figures could also be rectangles, or she might think I'm referring to the spots on the envelope rather than the figures. But that's not how the typically enculturated eight-year-old human mind works. The definition succeeds because I know she'll latch onto the intended concept rather than some less obvious concept that also fits the cases.

²⁸⁰ Siewert 1998, ch. 3.

²⁸¹ Although I believe that consciousness is an obvious category, definition by example will also probably tend to pick out natural kinds (e.g., tiger) over more haphazard collections (see, for example, the work on "reference magnetism", growing from Lewis 1983 and 1984) as well as categories friendly to classificational practice either in general or in one's subculture (see, for example, work on "fast-mapping", growing from Carey 1978, on how children quickly acquire concepts, sometimes after a single example). In addition to the obviousness of consciousness, consciousness might be a natural kind and the concept of consciousness might be readily adopted. If so, all the better for definition by example.

Defining “consciousness” by example requires that there be *only one* obvious, natural, or readily adopted category that fits the examples. I do think there is probably only one such category in the vicinity, at least once we do some explicit narrowing of the candidates. In Section 3, I’ll discuss concerns about this assumption.

Let’s begin with positive examples. The word “experience” is sometimes used non-phenomenally (“I have twenty years of teaching experience”). However, in English it often refers to consciousness in my intended sense. Similarly for the adjective “conscious”. I will use these terms in that way now, hoping that when you read them they help you grasp the relevant examples. However, I will not always rely on these terms. They are intended to point you toward the cases rather than as (possibly circular or synonymous) components of the definition.

Sensory and somatic experiences. If you aren’t blind and you think about your visual experience, you will probably find you are having some visual experience right now. Maybe you are visually experiencing black text on a white page. Maybe you are visually experiencing a computer screen. If you press the heels of your palms firmly against your closed eyes for several seconds, you will probably notice a swarm of bright colors and shapes, called phosphenes. All of these visual goings-on are examples of conscious experiences. Similarly, you probably have auditory experiences if you aren’t deaf, at least when you stop to think about it. Maybe you hear the hum of your computer fan. Maybe you hear someone talking down the hall. If you cup a hand over one ear, you will probably notice a change in the ambient sound. If you stroke your chin with a finger, you will probably have tactile experience. Maybe you’re feeling the pain of a headache. If you sip a drink right now, you’ll probably experience the taste and feel of the drink in your mouth. If you close your eyes and consider the positions of your limbs, you might have

proprioceptive experience of your bodily posture, which might grow vaguer if you remain motionless for an extended period. You might feel sated or hungry, energetic or lethargic.

Conscious imagery. Maybe there's unconscious imagery, but if there is, it's doubtful that you'll be able to reflect on an instance of it at will. Try to conjure a visual image – of the Eiffel Tower, maybe. Try to conjure an auditory image – of the tune of “Happy Birthday”, for example. Imagine how it would feel to stretch your arms back and wiggle your fingers. You might fail in some of these imagery tasks but hopefully you'll succeed in at least one, which you can now treat as another example of consciousness.

Emotional experience. Presumably, you've had an experience of sudden fear on the road, during or after an accident or near-accident. Presumably, you've felt joy, anger, surprise, disappointment, in various forms. Maybe there is no unified core feeling of “fear” or “joy” that is the same from instance to instance. No matter. Maybe emotional experience is nothing but various somatic, sensory, and imagery experiences. That doesn't matter either. Recall some occasions when you've vividly felt what you'd call an emotion. Add these to your list of examples of consciousness.

Thinking and desiring. Probably you've thought to yourself something like “what a jerk!” when someone has treated you rudely. Probably you've found yourself craving a dessert. Probably you've stopped to plan, deliberately in advance, the best route to the far side of town. Probably you've discovered yourself wishing that Wonderful Person X would notice and admire you. Presumably not all our thinking and desiring is conscious in the intended sense, but presumably any instances you can now vividly remember or create were or are conscious. Add these to the stock of positive examples. Again, it doesn't matter if these experiences aren't clearly differentiated from other types of experience discussed above.

Dream experiences. In one sense of “conscious” we are not conscious when we dream. But according to both mainstream scientific psychology and the commonsense understanding of dreams, dreams are conscious experiences in the intended sense, involving sensory or quasi-sensory experiences, or maybe instead only imagery, and often some emotional or quasi-emotional component like dread of the monster chasing you.

Other people. Bracketing radical skepticism about other minds (see Chapter 6), we normally assume that other people also have sensory experiences, imagery, emotional experiences, conscious thoughts and desires, and dreams. Count these too among the positive examples.

Negative examples. Not every event in your body belongs to your stream of conscious experiences. You presumably have no conscious experience of the growth of your fingernails, or of the absorption of lipids in your intestines, or of the release of growth hormones in your brain – nor do other people experience such things in themselves. Nor is everything that we ordinarily classify as mental conscious. Two minutes ago, before reading this sentence, you probably had no conscious experience of your disposition to answer “twenty-four” when asked “six times four”. You probably had no conscious experience of your standing intention to stop for lunch at 11:45. You presumably have no conscious experience of the structures of very early auditory processing. Under some conditions, if a visual display is presented for thirty milliseconds and then quickly masked – a very quick flash across your computer screen – you may have no conscious experience of it whatsoever, even if it later subtly influences your behavior. Nor do you have sensory experience of everything you know to be in your sensory environment: no visual experience of the world behind your head, no tactile experience of the smooth surface of your desk that you see but aren’t currently touching. Nor do you literally experience other

people's thoughts and images. Possibly, dreamless sleep involves a complete absence of conscious experiences.

Consciousness is the most obvious thing or feature that the positive examples possess and the negative examples lack, the thing or feature that ordinary people without any specialized training will normally notice when presented with examples of this sort. I do think that there is one very obvious feature that ties together sensory experiences, imagery experiences, emotional experiences, dream experiences, and conscious thoughts and desires. They're all *conscious experiences*. None of the other stuff is experienced (lipid absorption, your unactivated background knowledge that $6 \times 4 = 24$, etc.). I hope it feels to you that I have belabored an obvious point. Indeed, my argumentative strategy relies upon this obviousness.

Don't try to be too clever and creative here! Of course you could invent a *new and non-obvious* concept that fits the examples. You could invent some "Cambridge property" like *being conscious and within 30 miles of Earth's surface* or *being referred to in a certain way by Eric Schwitzgebel in this chapter*. Or you could pick out some scientifically constructed but non-obvious feature like accessibility to the "central workspace" or in-principle-reportability-by-a-certain-type-of-cognitive-mechanism. Or you could pick out a feature like the disposition to *judge* that you are having conscious experiences. None of these is the feature I mean. I mean the *obvious* feature, the thing that smacks you in the face when you think about the cases. That one!

Don't try to analyze it yet. Do you have an analysis of "furniture"? I doubt it. Still, when I talk about furniture you know what I mean and you can sort the positive and negative examples pretty well, with some borderline or doubtful cases. Do the same thing with consciousness. Save the analysis, reduction, and metaphysics for later.

Maybe scientific inquiry and philosophical reflection will reveal all the positive examples to have some set of functional properties in common or to be reducible to certain sorts of brain processes or whatever. Unifying features can also be found for “rectangle” and presumably “furniture”. This doesn’t prevent us from defining by example while remaining open-minded and noncommittal about theoretical questions that might be answered in later inquiry.

3. Contentious Cases and Wonderfulness.

Some consciousness researchers think that conscious experience is possible without attention – for example, that you have constant tactile experience of your feet in your shoes even though you rarely attend to your feet or shoes. Others think consciousness is limited mostly or entirely to what’s in attention.²⁸² Some researchers contend that conscious experience is exhausted by sensory, imagery, and emotional experiences, while others contend that consciousness comes in a wider range of uniquely irreducible kinds, possibly including imageless thoughts, an irreducible feeling of self, or feelings of agency.²⁸³

I have avoided committing on these issues by restricting the examples in Section 2 to what I hope are uncontentious cases. I did not, for example, list a peripheral experience of the feeling of the feet in your shoes among the positive examples, nor did I list non-conscious knowledge of the shoe-pressure on your feet among the negative examples. This leaves open the possibility that there are two or more fairly obvious categories that fit with the positive and negative examples and differ in whether they include or exclude such contentious cases. For

²⁸² In Schwitzgebel 2011b, ch. 6, I review the literature on this topic. See also Prinz 2012 and the extensive literature arising in reaction to Block 2007 and 2011.

²⁸³ For example, Titchener 1915 and Prinz 2012 defend versions of the first view, while Kriegel 2015 and Hurlburt (Hurlburt and Schwitzgebel 2007; Heavey and Hurlburt 2008) defend versions of the second view.

example, if conscious experience substantially outruns attention, both the intended category of *consciousness* and the narrower category of *consciousness-along-with-attention* adequately fit the positive and negative examples.

Similarly, consciousness might or might not always involve some kind of reflective self-knowledge, some awareness of oneself *as* conscious.²⁸⁴ I intend the concept as initially open on this question, prior to careful introspective and other evidence.

You might find it introspectively compelling that your own stream of conscious experience does, or does not, involve constant experience of your feet in your shoes, or reflective self-consciousness, or an irreducible sense of agency. Such confidence, in my view, is often misplaced, as I've argued at length elsewhere.²⁸⁵ But regardless of whether such confidence is misplaced, the intended concept of “consciousness” does not build in, *as a matter of definition*, that consciousness is limited (or not) to what's in attention or that it includes (or fails to include) phenomena such as an irreducible awareness of oneself as an experiential subject. If it seems to you that there are two obvious concepts here, one that commits definitionally on such contentious matters and another that leaves such questions open to introspective and other types of evidence, my intended concept is the less committal one. This is in any case probably the more obvious concept. We can *argue* about whether consciousness outruns attention. It's not normally antecedently stipulated.

It is likewise contentious what sorts of organisms are conscious – a question I will explore at length in Chapter 10. Do snails, for example, have streams of conscious experience?

²⁸⁴ As suggested by, for example, by “higher-order” views (Rosenthal 2005; Carruthers and Gennaro 2001/2020) and related self-representational views (Kriegel 2009). Among those who deny this are those drawn to “abundant” views of the sort discussed in Chapter 10, including panpsychist views (Seager, ed., 2020).

²⁸⁵ Especially Schwitzgebel 2011b.

If I touch my finger to a snail's tentacle, does the snail have visual or tactile phenomenology? If *consciousness* just meant "sensory sensitivity" we would have to say yes. If *consciousness* just meant "processes reportable through a cognitively sophisticated faculty of introspection" we would have to say no. I intend neither of those concepts but rather a concept that doesn't settle the question as a straightforward matter of definition. Again, I think this is probably the more typical concept to latch onto in any case.

It is this openness in the concept that enables it to meet the Wonderfulness condition. One can *wonder* about the relationship between consciousness and reportability, wonder about the relationship between consciousness and sensory sensitivity, wonder about the relationship between consciousness and any particular functional or biological process. One can wonder about idealism, dualism, an afterlife – any of a variety of weird views at odds with mainstream scientific approaches. Definition by example isn't completely devoid of potentially problematic background assumptions, as I'll discuss in the next section. But it's innocent of the types of assumptions that overcommit on these issues. In consciousness studies, as elsewhere in life, innocence enables wonder.

Wonder needn't be permanent. Maybe a bit of investigation will definitively settle these questions. Wonder is compatible even with the demonstrable mathematical impossibility of some of the epistemically open options. Before doing the calculation, you can wonder if and where the equation $y = x^2 - 2x + 2$ crosses the x axis. The Wonderfulness condition as I intend it here requires no insurmountable epistemic gap, only a moment's epistemic breathing space.

I submit that there is exactly one concept, perhaps blurry-edged, that is obvious to nonspecialists, fits the positive and negative examples, leaves the contentious examples open, and permits wonder of the intended sort. That is the concept of consciousness.

4. Problematic Assumptions?

Some philosophers have argued that consciousness, or phenomenal consciousness, does not exist. Keith Frankish is the most notable recent advocate, but others include Paul Feyerabend, Jay Garfield, François Kammerer, and maybe early Patricia Churchland.²⁸⁶ The argument is always a version of the following: The ordinary concept of (phenomenal) consciousness ineliminably contains some dubious metaphysical or epistemic assumption at odds with the mainstream scientific or materialist world conception. In this way, “consciousness” is like “ghost” or “Heavenly spirit” or “divine revelation”. You can’t get the immaterial metaphysics out of the concept of a ghost or spirit; you can’t get the religious epistemology out of the concept of divine revelation. Therefore, all these terms must go; there is no such thing.

Frankish in particular offers a lovely list of dubious commitments that consciousness enthusiasts sometimes commit to. Let me now disavow all such commitments. Conscious experiences need not be simple, nor ineffable, nor intrinsic, nor private, nor immediately apprehended. They need not have non-physical properties, or be inaccessible to “third-person” science, or be inexplicable in materialist vocabulary. My definition did not, I hope, commit me on any such questions (which was exactly *Desideratum 1*). My best guess – though only a guess (see Chapter 3) – is that all such claims are false, if intended as universal generalizations about consciousness. (However, if some such feature turns out to be present in all of the examples and

²⁸⁶ Feyerabend 1963; Churchland 1983; Garfield 2015; Frankish 2016; Kammerer 2021. I agree with Niikawa 2021 that eliminativism or illusionism only denies the existence of phenomenal consciousness on the assumption that the term “phenomenal consciousness” is not as innocent as I try to make it in this chapter (see also Schwitzgebel 2020a). In his reply to the article on which this chapter was based, Frankish, for example, agrees that consciousness in the innocent sense that I have defined it does exist and constitutes a “neutral explanandum” for consciousness science (Frankish 2016, p. 277).

thus, by virtue of its presence, in some sense indirectly or implicitly built into the definition by example, that's okay. Such indirect commitments to features that are actually universally present shouldn't implausibly inflate our target. This is different, of course, from commitment to features *falsely believed* to be universally present.)

My definition did commit me to a fairly strong claim about ordinary non-specialists' thinking about the mind: that there is a single obvious category that people will latch on to once provided the positive and negative examples. This is, however, a potentially empirically testable psychological commitment of a rather ordinary type rather than commitment to a radical or peculiar view about the metaphysics or epistemology of consciousness.

I also committed to realism about that category. This category is not empty or broken but rather picks out a feature that (most of) the positive examples share and the negative examples presumably lack. If the target examples had nothing important in common and were only a hodgepodge, this assumption would be violated. This realism is a substantive commitment. It's a plausible one I hope, supported by both intuitive considerations and various competing scientific models that attempt to explain what functional or neurophysiological systems underlie the commonality.²⁸⁷

The Wonderfulness condition involves a mild epistemic commitment in the *neighborhood* of immateriality or irreducibility. The Wonderfulness condition commits to its being not straightforwardly obvious as a matter of definition how consciousness relates to cognitive, functional, or physical properties. This commitment is entirely compatible with the

²⁸⁷ Here I'm considering failure of commonality among the positive examples. Another possible definitional threat would be if the putative negative examples failed to be negative, as in some versions of panpsychism. In this case, we might still salvage the concept by targeting the feature that the positive examples have and that the negative examples are *falsely assumed* to lack.

view that a clever argument or compelling empirical evidence could someday show, perhaps even already has shown (though not all of us appreciate it), that consciousness is reducible to or identical to something functional or physical.

After being invited to consider the positive and negative examples, someone might say, “I’m not sure I understand. What *exactly* do you mean by the word ‘consciousness?’” At this point, it is tempting to clarify by making some metaphysical or epistemic commitments – whatever commitments seem most plausible to you. You might say, “our conscious experiences are those events with which we are most directly and infallibly acquainted” or “conscious properties are the kinds of properties that can’t be reduced to physical properties or functional role”. Please don’t! At least, don’t build these commitments into the definition. Such commitments risk introducing doubt or confusion in people who aren’t sure they accept such commitments.

We have arrived at the second way to unlearn the meaning of “consciousness” that I mentioned at the start of this chapter. If you say “by ‘consciousness’ I mean such-and-such that inevitably escapes physical description”, then your hearer, who maybe thinks nothing need inevitably escape physical description, might come to wonder whether you mean something different from “conscious” than they do. And maybe you do, now, mean something different if you aim to truly and ineliminably build that feature into your concept as an essential part. Thus the definition of consciousness grows murky, weedy, and equivocal.

Here’s a comparison. You are trying to teach someone the concept “pink”. Maybe their native language doesn’t have a corresponding term (as English has no widely used term for pale green). You have shown her a wide range of pink things (a pink pen, a pink light source, a pink shirt, pictures and photos with various shades of pink in various natural contexts). You’ve

recalled some famously pink things like cherry blossoms and flamingos. You've pointed out some non-pink things as negative examples (medium reds, pale blues, oranges, etc.). It would be odd for them to ask, "So, do you mean this-shade-and-mentioned-by-you?" or "Must pink things be less than six miles wide?" It would also be odd for them to insist that you provide an analysis of the metaphysics of pink before she accepts it as a workable concept. You might be open about the metaphysics of pink.²⁸⁸ It might be helpful to point, noncommittally, to what some people have said ("Well, some people think of pink as a reflectance property of physical objects"). But lean on the examples. If your friend isn't colorblind or perverse, there's something obvious that the positive instances share, which the negative examples lack, which normal people will normally latch onto well enough if they don't try too hard to be creative and if you don't confuse things by introducing dubious theses. This is a perfectly adequate way to teach someone the concept *pink*, well enough that your friend can now confidently affirm that pink things exist (perhaps feeling baffled how anyone could deny it), sorting future positive and negative examples in more or less the consensus way, except perhaps in borderline cases (e.g., near-red) and contentious cases (e.g., someone's briefly glimpsed socks). The concept of *consciousness* can be approached in the same manner.

²⁸⁸ The metaphysics of color is a notoriously tricky issue! For a review, see Maund 1997/2019.

The Weirdness of the World

Chapter Nine The Loose Friendship of Visual Experience and Reality

According to the United States Code of Federal Regulations, Title 49, Ch. 5, §571.111, S5.4.2, pertaining to convex mirrors on passenger vehicles,

Each convex mirror shall have permanently and indelibly marked at the lower edge of the mirror's reflective surface, in letters not less than 4.8 mm nor more than 6.4 mm high the words "Objects in Mirror Are Closer Than They Appear".²⁸⁹

See Figure 1 for an example.

²⁸⁹ See <https://www.govinfo.gov/content/pkg/CFR-2019-title49-vol6/pdf/CFR-2019-title49-vol6-sec571-111.pdf> [accessed Jan. 8, 2021].



Figure 1. “OBJECTS IN MIRROR ARE CLOSER THAN THEY APPEAR”. [Photo by me. This image can be converted to black and white, if desirable to reduce printing costs. Either convert the current image to black and white if that’s possible without impairing its legibility or consult with me about substituting another image. Crucial to the image is the text “OBJECTS IN MIRROR ARE CLOSER THAN THEY APPEAR”, so that must remain legible, or even be enhanced if possible.]

In this chapter, I will argue that the phenomenologists at the U.S. Department of Transportation have it wrong. For skilled drivers, objects in convex passenger-side mirrors are not closer than they appear. If I’m right about that, consequences follow for our understanding

of the relationship between visual experience and reality. The relationship, though friendly, is in some respects loose.

Consider your visual experience right now. To what extent is reality like *that*? As discussed in Chapter 5, Kant holds that, fundamentally, mind-independent reality doesn't much resemble our experience of it. We have no good reason to think that things as they are in themselves, independently of us, even have shapes and sizes. Spatial properties depend essentially on the structure of our senses.

Others argue that the relationship is closer. Locke, for example, holds that although color is not a feature of things as they are in themselves, shape is such a feature.²⁹⁰ Your experience of blue does not resemble the vibrations of light that cause that experience in you – no more than your experience of sweetness resembles the chemical structures in ice cream. But, according to Locke, shape and size are different. Suppose you are looking at a cube in something like normal conditions. The visual experience you have of cubicalness – of edges and corners arranged in such-and-such a manner, faces of equal size, positioned at right angles to each other – in some important respect resembles the external cube out there in front of you. Although we paint the world with color and sweetness and painfulness that do not intrinsically belong to things as they exist independently of us, our experience of shape and size – except in cases of illusion – is an orderly geometric transformation of objects' real, mind-independent shapes and sizes.

I will argue that the relationship need not be so orderly. It is instead, probably, highly contingent and skill-dependent. My argument will not take us all the way to Kant, but it will

²⁹⁰ Locke 1960/1975. For a variety of positions on the historical and contemporary debate about primary and secondary qualities, see Nolan, ed., 2011.

suggest that the association between visual experience and mind-independent external reality is more like a loose friendship than a straightforward geometrical transformation.

1. *“Objects in Mirror Are Closer Than They Appear”.*

Are objects in convex passenger-side car mirrors closer than they appear? Here are three possible answers:

1. Why yes, they are!
2. No, objects in the mirror are not closer than they appear, but that’s because instead of being closer than they appear they are larger than they appear. The convex mirror distorts size rather than distance.
3. No, objects in the mirror are not closer than they appear, nor are they larger than they appear. If you’re a skilled driver, the car behind you, as seen with the aid of a convex, passenger-side mirror, is just where it appears to be and just the size it appears to be.

I will now present three considerations in favor of Answer 3.

One reason to favor Answer 3 is that there seem to be no good geometrical grounds to privilege Answer 1 over Answer 2 or vice versa. The car subtends fewer degrees of visual arc in the convex mirror than it would in a flat mirror and appears smaller in flat projection on the printed page (Figure 2). But this could be construed either as a distortion of size or as a distortion of distance. Maybe the car looks 3 feet tall and 30 feet away, approaching at one relative speed. Or maybe the car looks 6 feet tall and 60 feet away, approaching at a different relative speed. The car might look illusorily distant, similarly to how things viewed through the wrong end of a telescope look farther away than they are. Or it might look illusorily small,

similarly to how things viewed through a magnifying glass look larger than they are. Geometrically, nothing seems to favor one way of thinking about the convex car mirror distortion over the other. It also seems unjustified to select some specific compromise, such as that the car looks 4.5 feet tall and 45 feet away. If Answers 1 and 2 each problematize the other, that provides at least some initial reason to favor Answer 3.



Figure 2. The view in flat (left) versus convex (right) driver-side mirrors. Image detail from Wierwille et al. 2008. [This image can be converted to black and white, if desirable to reduce printing costs. Either convert the current image to black and white if that's possible without impairing its legibility or consult with me about substituting another image.]

Another reason to favor Answer 3 over Answers 1 and 2 is that skilled drivers are undecieved. Decades of auto safety research show that practiced drivers do not misjudge the distance of cars in convex passenger-side mirrors. People don't misjudge when asked oral questions about the distance, nor do they misjudge when they make lane-shifting decisions in actual driving. Instead, they seem to skillfully and spontaneously use the mirror to accurately

gauge the real distance of the cars behind them.²⁹¹ Presumably (though I'm not aware of its having been explicitly tested) they also accurately judge car size. No one mistakes the looming F-150 for a new type of Cooper Mini. Of course, people familiar enough with an illusion might be entirely undeceived by it. But I hope you'll agree that there's something happier about a position that avoids the proliferation of undeceiving illusions, if an alternative theory is available.²⁹²

A third type of evidence for Answer 3 is introspective. Get in a car. Drive. Think about how the cars behind you look, in both the driver-side and passenger-side mirrors. Try adjusting the two mirrors so that they both point at the same car behind you when your head is centrally positioned in the cabin. Does the car look closer when your eyes are aimed at the flat driver-side mirror than when your eyes are aimed at the convex passenger-side mirror? Based on my own messing around, I'm inclined to say no, the cars look the same distance in both mirrors. There is more than one way in which a car can look like it's 6 feet tall and 60 feet behind. There's a flat-driver-side-mirror way, and there's a convex-passenger-side-mirror way.

2. *The Multiple Veridicalities View.*

Combined, the three considerations above create some pressure in favor of what I'll call the *Multiple Veridicalities* view of sensory experience. According to Multiple Veridicalities, there is more than one distinct way that an object can appear to us in vision (or any other sensory modality) completely truthfully, accurately, and without distortion.²⁹³ Two or more qualitatively

²⁹¹ De Vos 2000; Wierwille et al. 2008.

²⁹² I explore theoretical puzzles about the possibility of undeceiving illusions in Schwitzgebel 2011b, ch. 5.

²⁹³ "Inverted spectrum" cases are an interesting comparison point. On the inverted spectrum thought experiment, it is at least in principle possible that someone might experience

different visual experiences – one that looks like *this* (the way the car behind me looks in the driver-side mirror) and another that looks like *that* (the way the car behind me looks in the passenger-side mirror) can equally well visually present, without distortion, exactly the same set of visually available distal properties (the car’s height, shape, speed, relative distance, and so on). These two different experiences each represent, with complete faithfulness, the same state of affairs as transformed through different intervening media.

You might think that convex mirrors distort what they present while flat mirrors are faithful (or at least more faithful). You might analogize to warped funhouse mirrors, which we normally regard as distortive. In one, you look short and fat, in another tall and thin, while another stretches your legs and shortens your chest. You look wrong. You look like you have properties you don’t in fact have, like long, skinny legs. That’s why funhouse mirrors are *fun*.

On the Multiple Veridicalities view, matters aren’t so simple. Not all non-flat mirrors distort. Some are perfectly veridical. It’s a matter of what you’re used to, or what you expect, or how skillfully and naturally you use the mirror. If there is no temptation to misjudge, if you can act smoothly and effectively with no more need of cognitive self-correction than in the flat-mirror case, if accurate representations of the world arise smoothly and swiftly, and if you’d say everything looks natural and correct, then there is no need to posit distortion simply because the mirror’s surface is non-flat.

green when looking at canonically red things and vice versa. If this possibility were combined with a form of color realism, it might be possible to have multiple veridicalities in color experience, my experience of red and the invert’s experience of green both equally veridically representing the redness of a ripe tomato. For a careful discussion of the inverted spectrum thought experiment and its argumentative uses, see Byrne 2004/2020. See also discussions of the possibility of left-right reversed worlds and vertically stretched “El Greco worlds” in Hurley 1998; Lee 2006; Thompson 2010; Chalmers 2019b. To be clear, my discussion does not assume the complete functional or representational identity of the contrasting veridical experiences, since the media (convex vs. flat reflection) are represented differently.

Imagine another culture, Convexia, where mirrors are always manufactured convex. Maybe glass is expensive, and convexity helps reduce manufacturing cost by allowing larger viewing angles to be presented in a more compact mirror that uses less glass. Maybe the amount of convexity is always the same, as a matter of regulation and custom, so that people needn't re-adjust their expectations for every mirror. Convexians use their mirrors much as we do – for grooming in bathrooms, to see what's behind them while driving, and to decorate their homes. Now you arrive in Convexia with your flat mirror. Will the Convexians gaze at it in amazement, saying “finally, a mirror that doesn't distort!” Will they rush to install it in their cars? Doubtful! Presumably, this new type of mirror will require some getting used to. Might the Convexians even at first think that it is your flat mirror that distorts things, while their convex mirrors display the world accurately?

If the Multiple Veridicalities view can be sustained for convex rearview mirrors, we can consider how far it might generalize. In my early 40s, I switched from single-correction prescription eyeglasses to progressive lenses. At first, the progressive lenses seemed to warp things in the bottom half of the visual field, especially when I shook my head or my glasses slid halfway down my nose. Now I don't experience them as distorting the world at all. In fact, I sometime switch back and forth between progressives and single-correction eyeglasses for different purposes, and although I notice differences in the clarity of objects at different distances, and although the geometry of refraction works differently for my different pairs of glasses, objects never look warped and action remains fluid.

The case of the half-submerged oar has spawned a minor philosophical subliterature, going back to Plato's *Republic*.²⁹⁴ Most people in our culture appear ready to say that an oar half submerged in water "looks bent". If the Multiple Veridicalities view extends to this case, then to a sufficiently skilled rower, the half-submerged oar looks straight, not bent. It looks just like a straight oar should look when half submerged. We might imagine the rower to be accustomed to rowing amid shallow rocks, so that knowledge of the exact angle and position of the oar becomes swift and second nature while rowing. Such a rower might be shocked and stunned to see an oar designed to bend precisely at the point of contact with the water, cleverly arranged so that it always bends so as to "look straight" to a completely naive eye in the rower's position. Faced with such a trick oar, our expert rower would immediately recognize that it was bent, and as the oar moved, our rower would immediately notice its strange deformations of shape. Now, you might object that an ordinary half-submerged oar must look bent even to an expert rower because of some simple geometric fact, such as that a photograph from the rower's point of view would depict the oar as other than a straight line on the page. To this, I reply that the assumption that flat pictures accurately depict our visual experience is both culturally specific and geometrically flawed.²⁹⁵ To see the beginning of the general geometrical issue, consider the troubles with printing wide-angle or panoramic photographs on a flat page.

Generalizing further, it seems that at least in principle people could adjust to any systematic geometry of reflection or refraction, as long as it isn't too complicated.²⁹⁶ Maybe if you saw yourself in a funhouse mirror day and night, you'd slowly lose your temptation to think

²⁹⁴ Plato 4th c. BCE/1992, 602c. Other examples: Montaigne 1580/1595/2003, ch. 14 ("taste of good and evil"); Berkeley 1710-1713/1965, 3rd dialogue; Ayer 1958; Austin 1962; Marr 1982/2010; Brogaard 2014; Maddy 2017.

²⁹⁵ See Schwitzgebel 2011b, ch. 2, for extended discussion of this point.

²⁹⁶ Though see Kohler 1962 for a classic discussion of some possible limitations.

it distorts. Maybe if you always wore a fisheye lens over one eye, you'd come to regard that view as perfectly accurate and natural. Maybe if you were a god whose eye was a huge sphere encompassing the Earth, always gazing in toward Earth's inhabitants, your spherical visual geometry would seem most perfect and divine.

3. *Inverting Lenses.*

Let's consider a famous case from history of psychology: inverting lenses.

Inverting lenses were first tried by George Stratton in the late 19th century.²⁹⁷ Stratton covered one eye and then presented to the other eye a field of view rotated 180 degrees so that top was bottom and left was right. In his primary experiment, he wore this lens for the bulk of the day over the course of eight days, and he gives detailed introspective reports about his experience. Stratton adapted to his inverting lenses. But what does *adapting* amount to?

The simplest possibility to conceptualize is this: After adaptation, everything returns to looking just the way it did before. Let's say that pre-experiment you gaze out at the world and see a lamp. Let's call the way things look, the way the lamp looks, before you don the inverting glasses, *teavy*. Now you don the glasses and at first everything seems to have rotated 180 degrees. Let's call that visual experience, the way the lamp looks to you now, *toovy*. According to a simple view of adaptation, if you wear the glasses for eight days, things return to looking *teavy* – perhaps at first slowly, unstably, and disjointedly. After adaptation, things look the same way they would have looked had you never donned the glasses at all (ignoring details about the frame, the narrower field of view, minor imperfections in the lenses, etc.). This is the way that

²⁹⁷ Stratton 1896, 1897a, 1897b, 1897c.

adaptation to inverting lenses is sometimes described, including by influential scholars such as Ivo Kohler, James Taylor, Susan Hurley, and Alva Noë.²⁹⁸

However, there is another possibility – I think a more interesting and plausible one. That’s the possibility that after donning the lenses, things continue to look toovy throughout the adaptation process, but you grow accustomed to their looking toovy, so that you lose the normative sense that this is a wrong or misleading way for things to look. The lamp no longer looks “upside-down” in the normative sense, that is to say the evaluative sense, of looking like the wrong side is on top, but it retains its “upside-down” look in the non-normative sense that the visual experience is the reverse of what it was before you put on the inverting lenses. To the adapted mind, there would now be two ways in which a lamp might look to be right-side up: the without-glasses teavy way and the with-glasses toovy way.

Maybe toovy changes too, as one accommodates, becoming toovy-prime, more richly layered with meaning and affordances for action. I don’t mean to deny that. It’s a tricky phenomenological debate to what extent our visual experience contains within it features like something looking to be the kind of thing that “affords” comfortable sitting or the kind of thing that could easily be stopped from tipping over if I reached out my hand like this.²⁹⁹ Such knowledge might be embedded in visual experience itself, or it might instead arise only in post-visual cognition. On this issue, I take no stand. The important thing for my current project is that the experience doesn’t return to teavy.

²⁹⁸ Kohler 1951/1964; Taylor 1962; Hurley and Noë 2003.

²⁹⁹ Recent advocates include Nanay 2011; Prosser 2011; Siegel 2014; Gallagher 2017 – a tradition drawing from Merleau-Ponty 1945/2012 and Gibson 1979/2015. See Siegel and Byrne 2017 for a pro and con discussion of the extent to which perceptual experience is rich with affordances and other properties.

[Illustration 7 (Caption: Three ways of experiencing the same visual scene): Three views of the same living room scene from the point of view of an observer: We see the observer's legs on an ottoman or footrest and maybe a hand on an armrest, so that it's really clear that the point of view is a person's eye. (Compare this famous drawing – but use a lens as a frame instead of an eyebrow and nose:

https://commons.wikimedia.org/wiki/File:Ernst_Mach_Inner_perspective.jpg) Each scene is framed in a circle to convey that it is being seen through a lens. The first is right-side-up. The second and third are exactly the same scene, but upside down. Above the first is the word “teavy”. Above the second is the word “toovy”. Above the third is the word “toovy”. Below the first is “ah, good, that’s right-side up”. Below the second is “uh-oh, everything is upside-down”. Below the third is “ah, good, this feels right-side up again”.]

If the lamp continues to look toovy after adaptation, without thereby looking wrong, distorted, or misleading, that suggests the existence of two equally correct or veridical ways in which things, like a lamp, could look to have the same set of spatial properties, including the same position and orientation relative to you. To the extent such an understanding of inversion adaptation is plausible, it supports the Multiple Veridicalities view. Conversely, to the extent the Multiple Veridicalities view is independently plausible, it supports this understanding of inversion adaptation.

It is an empirical question whether the Multiple Veridicalities view is correct about inverting lenses or whether, instead, things really do just go back to looking teavy after adaptation. Furthermore, it's a tricky empirical question. It's a question that requires introspective reporting by someone with a subtle sense of what the possibilities are, especially a subtle sense of the different things one might mean by saying that something “looks like it's on

the right” or “looks upside-down”. As one might expect, the introspective reports of people who have tried inverting lenses are not entirely consistent or unequivocal. However, my assessment of the evidence is that the experimenters with the best nose for this sort of nuance – especially Stratton himself and later Charles Harris – favor the Multiple Veridicalities view.³⁰⁰ Stratton writes

But the restoration of harmony between perceptions of sight and those of touch was in no wise a process of changing the absolute position of tactual objects so as to make it identical with the place of the visual objects; no more than it was an alteration of the visual position into accord with the tactual. Nor was it the process of changing the relative position of tactual objects with respect to visual objects; but it was a process of making a new visual position seem the only natural place for the visual counterpart of a given tactual experience to appear in; and similarly in regard to new tactual positions for the tactual accompaniment of given visual experiences (1897c, p. 476).

Harris similarly writes “there is no change in subjects’ purely visual perception, but... their position sense and movement sense are modified” (1980, p. 111). In other words, things keep looking toovy, but toovy no longer seems wrong, and your motor skills adapt to align with the flipped visual appearances.³⁰¹ There’s more than one way for things to look right-side up.

³⁰⁰ Stratton 1896, 1897a, 1897b, 1897c; Harris 1965, 1980; see also Linden et al. 1999; Klein 2007.

³⁰¹ Stratton and Harris disagree about whether tactile and proprioceptive experience change during the course of adaptation. Stratton’s view, as expressed in the block quote, is that they do not. All that is learned is a new set of relationships. Harris’s view is that tactile and motor experience do change, so adaptation to the inverting goggles involves a gradual flipping of tactile and proprioceptive experiences to match the flipped/toovy visual experiences (1965, p. 438).

4. *Imagining a Loose Friendship.*

I look out now upon the world. I imagine looking out upon it, just as veridically, through a fish-eye lens. I imagine looking out upon it, just as veridically, through increasingly weird assemblies that I would have said, the first time I gazed through them, distorted things terribly, making some distant things too large and some near things too small, that presented twists and gaps – maybe even that doubled some things while keeping others single – but to which I grow skillfully accustomed. I imagine my visual experience not shifting back to what it was before, but instead remaining different while shedding its sense of wrongness. After I imagine all this, I am no longer tempted to say that things, considered as they are in themselves, independently of me and my experience, are more like *this* (my experience without the devices) than like *that* (my experience with the devices). My pre-device visual experience was not a more correct window on the world than my post-device visual experience. Wherever those experiences differ, it is not the case that one is truer to the world than the other.

I imagine extending this exercise to other senses. I imagine hearing through tubes and headphones that alter my auditory experience and the regions of space I sense most acutely. I imagine touching through gloves of different textures and with sticks and actuators and flexible extensions that modify my tactile interaction with the world. I imagine tasting and smelling differently, sensing my body differently, perhaps acquiring new senses altogether.

I am unsure how far I can push this line of thinking. But the farther I can push it, the looser the relationship must be between my experience of things and things as they are in themselves. Wherever multiple veridicalities are possible, the mind-independent world recedes. Any property of experience that can vary without compromising its veridicality is a feature we bring to the world rather than a feature intrinsic to the world. In the extreme, if this is true for

every discoverable aspect of our experience, we live, so to speak, entirely within a bubble shaped by our minds' contingent structures and habits.

The Weirdness of the World

Chapter Ten

Is There Something It's Like to Be a Garden Snail? Or: How Sparse or Abundant Is Consciousness in the Universe?

Consciousness might be abundant in the universe, or it might be sparse. Consciousness might be cheap to build and instantiated almost everywhere there's a bit of interesting complexity, or it might be rare and expensive, demanding nearly human levels of cognitive sophistication or very specific biological conditions.

Maybe the truth is somewhere in the middle. But it is a vast middle! Are human fetuses conscious? If so, when? Are lizards, frogs, clams, cats, earthworms? Birch forests? Jellyfish? Could an artificially intelligent robot ever be conscious, and if so, what would it require? Could groups of human beings, or ants or bees, ever give rise to consciousness at a group level? How about hypothetical space aliens of various sorts?

Somewhere in the middle of the middle, perhaps, is the garden snail. Let's focus carefully on just this one organism. Reflection on the details of this case will, I think, illuminate general issues about how to assess the sparseness or abundance of consciousness in general.

*1. The Options: Yes, No, and *Gong*.*

I see three possible answers to the question of whether garden snails are conscious: yes, no, and denial that the question admits of a yes-or-no answer.

To understand what *yes* amounts to, we need to understand what "consciousness" is in the intended sense. To be conscious is to have a stream of experience of some sort or other –

some sensory experiences, maybe, or some affective experiences of pleasure or displeasure, relief or pain. Possibly not much more than this. To be conscious is to have some sort of “phenomenology” as that term is commonly used in 21st century Anglophone philosophy. In Thomas Nagel’s famous phrasing, there’s “something it’s like” (most people think) to be a dog or a monkey and nothing it’s like (most people think) to be a photon or a feather: Dogs and monkeys are conscious and photons and feathers are not. (In Chapter 3 above, and also later in this chapter, I discuss some contrary views. See Chapter 8 for a detailed exploration of the definition of “consciousness”.) If garden snails are conscious, there’s something it’s like to be them in this sense. They have, maybe, simple tactile experiences of the stuff that they are sliming across, maybe some olfactory experiences of what they are smelling and nibbling, maybe some negative affective experiences when injured.

If, on the other hand, garden snails are not conscious, then there’s nothing it’s like to be one. They are as experientially empty as we normally assume photons and feathers to be. Physiological processes occur, just like physiological processes occur in mushrooms and in the human immune system, but these physiological processes don’t involve real experiences of any sort. No snail can genuinely *feel* anything. Garden snails might be, as one leading snail researcher expressed it to me, “intricate, fascinating machines”, but nothing more than that – or to be more precise, no more conscious than most people assume intricate, fascinating machines to be, which is to say, not conscious at all.

Alternatively, the answer might be neither yes nor no. *The Gong Show* is an amateur talent contest in which performers whose acts are sufficiently horrid are interrupted by a gong and ushered offstage. Not all yes-or-no questions deserve a yes-or-no answer. Some deserve to be gonged off the stage. “Are garden snails conscious?” might be one such question – for

example, if the concept of “consciousness” is a broken concept, or if there’s an erroneous presupposition behind the question, or (somewhat differently) if it’s a perfectly fine question but the answer is in an intermediate middle space between yes and no.

Here’s what I’ll argue in this chapter: Yes, no, and *gong* all have some plausibility to them. Any of these answers might be correct. Each answer has some antecedent plausibility – some plausibility before we get into the nitty-gritty of detailed theories of consciousness. And if each answer has some antecedent plausibility, then each answer also has some posterior plausibility – some plausibility *after* we get into the nitty-gritty of detailed theories of consciousness.

Antecedent plausibility becomes posterior plausibility for two reasons. First, there’s a vicious circle. Given the broad range of antecedently plausible claims about the sparseness or abundance of consciousness in the world, in order to answer the question of how widespread consciousness is, even roughly, we need a good theory. We need, probably, a well-justified general theory of consciousness. But a well-justified general theory of consciousness is impossible to build without relying on some background assumptions about roughly how widespread consciousness is. Before we can have a general sense of how widespread consciousness is, we need a well-justified theory; but before we can develop a well-justified theory, we need a general sense of how widespread consciousness is. Before X, we need Y; before Y, we need X.

Antecedent plausibility becomes posterior plausibility for a second reason too: Theories of consciousness rely essentially on introspection or verbal report, and all of our introspections and verbal reports come from a single species. This gives us a limited evidence base for extrapolating to very different species.

Contemplate the garden snail with sufficient care and you will discover, I think, that we human beings, in our current scientific condition, have little ground for making confident assertions about one of the most general and foundational questions of the science of consciousness, and indeed one of the most general and foundational questions of all philosophy and cosmology: How widespread is consciousness in the universe?

2. The Brains and Behavior of Garden Snails.

If you grew up in a temperate climate, you probably spent some time bothering brown garden snails (*Cornu aspersum*, formerly known as *Helix aspersa*; Figure 1) or some closely related species of pulmonate (air-breathing) gastropod. Although their brains are much smaller than those of vertebrates, their behavior is in some ways strikingly complex. They constitute a difficult and interesting case about which different theories yield divergent judgments.



Figure 1. *Cornu aspersum*, the common garden snail. Photo: Bryony Pierce (cropped). Used with permission. [This figure can be black and white if desirable to save on printing costs. Either convert this image to black and white, if it's possible to do that without impairing its legibility, or we can find another image.]

2.1. Snail brains. The central nervous system of the brown garden snail contains about 60,000 neurons.³⁰² That's quite a few more neurons than the famously mapped 302 neurons of the *Caenorhabditis elegans* roundworm, but it's also quite a few less than the quarter million of an ant or fruitfly. Gastropod neurons generally resemble vertebrate neurons, with a few notable differences.³⁰³ One difference is that gastropod central nervous system neurons usually don't have a bipolar structure with an axon on one side of the cell body and dendrites on the other side. Instead, input and output typically occur on both sides without a clear differentiation between axon and dendrite. Another difference is that although gastropods' small-molecule neural transmitters are the same as in vertebrates (e.g., acetylcholine, serotonin), their larger-molecule neuropeptides are mostly different. Still another difference is that some of their neurons are huge by vertebrate standards.

The garden snail's central nervous system is organized into several clumps of ganglia, mostly in a ring around its esophagus.³⁰⁴ Despite their relatively low number of central nervous system neurons, they have about a quarter million peripheral neurons, mostly in their posterior (upper) tentacles, and mostly terminating within the tentacles themselves, sending a reduced signal along fewer neurons into the central nervous system.³⁰⁵ (How relevant peripheral neurons

³⁰² Chase 2001.

³⁰³ Chase 2001, 2002.

³⁰⁴ Kerkut, Lambert, Gayton, Loke, and Walker 1975.

³⁰⁵ Chase 2001.

are to consciousness is unclear, but input neurons that don't terminate in the central nervous system are usually assumed to be irrelevant to consciousness.) Figure 2 is a schematic representation of the central nervous system of the closely related species *Helix pomatia*.

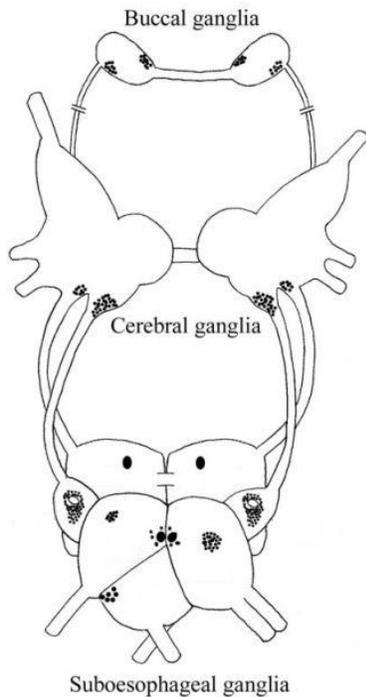


Figure 2. Schematic representation of the central nervous system of *Helix pomatia*, adapted from Casadio, Fiumara, Sonetti, Montarolo, and Ghirardi 2004.

2.2. Snail behavior. Snails navigate primarily by chemoreception, or the sense of smell, and mechanoreception, or the sense of touch. They will move toward attractive odors, such as food or mates, and they will withdraw from noxious odors and tactile disturbance. Although garden snails have eyes at the tips of their posterior tentacles, their eyes seem to be sensitive only to light versus dark and the direction of light sources, rather than to the shapes of objects.³⁰⁶ The

³⁰⁶ Chase 2002; Zieger and Meyer-Rochow 2008.

internal structure of snail tentacles shows much more specialization for chemoreception, with the higher-up posterior tentacles perhaps better for catching odors on the wind and the lower anterior tentacles better for ground and food odors. Garden snails can also sense the direction of gravity, righting themselves and moving toward higher ground to escape puddles. Arguably, at least some pulmonate snails sleep.³⁰⁷

Snails can learn. Gastropods fed on a single type of plant will preferentially move toward that same plant type when offered the choice in a Y-shaped maze.³⁰⁸ They can also learn to avoid foods associated with noxious stimuli, sometimes even after a single trial.³⁰⁹ Some gastropod species will modify their degree of attraction to sunlight if sunlight is associated with tumbling inversion.³¹⁰ “Second-order” associative learning also appears to be possible: For example, when apple and pear odors are associated and the snails are given the odor of pear but not apple together with the (more delicious) taste of carrot, they will subsequently show more feeding related response to both pear and apple odors than will snails for which apple and pear odors were not associated.³¹¹ The terrestrial slug *Limax maximus* appears to show compound conditioning, able to learn to avoid odors A and B when those odors are combined, while retaining attraction to A and B separately.³¹² In *Aplysia californica* “sea hares”, the complex role of the central nervous system in governing reflex withdrawals has been extensively studied,

³⁰⁷ Stephenson and Lewis 2011.

³⁰⁸ Croll and Chase 1980; Avila 1998; relatedly, Nitikin, Korshunova, Zakharov, and Balaban 2008.

³⁰⁹ Sahley, Gelperin, and Rudy 1981; Kimura, Toda, Sekiguchi, and Kirino 1998.

³¹⁰ Crow and Alkon 1978; Lederhendler, Gert, and Alkon 1986.

³¹¹ Lloyd, Fernández, and Acebes 2006; relatedly, Gelperin 2013; Hawkins and Byrne 2015.

³¹² Hopfield and Gelperin 1989. Snails can also respond differentially to A and B when those odors come from different locations than to A+B presented at the same location, showing that they react differently to presentation of combined odors than to nearby odors that are associated.

partly due to the conveniently large size of some *Aplysia* neurons (the largest of any animal species). *Aplysia californica* reflex withdrawals can be modified centrally – mediated, inhibited, amplified, and coordinated, maintaining a singleness of action across the body and regulating withdrawal according to circumstances.³¹³ Garden snail nervous systems appear to be at least as complex, generating unified action that varies with circumstance.

Garden snails can coordinate their behavior in response to information from more than one sensory modality at once.³¹⁴ As previously mentioned, when they detect that they are surrounded by water, they seek higher ground. They will cease eating when satiated, balance the demands of eating and sex depending on level of starvation and sexual arousal, and exhibit less withdrawal reflex while mating. Land snails will also maintain a home range to which they will return for resting periods or hibernation, rather than simply moving in an unstructured way toward attractive sites or odors.³¹⁵

Garden snail mating is famously complex.³¹⁶ The species is a simultaneous hermaphrodite, playing both the male and female role simultaneously. Courtship and copulation requires several hours. Courtship begins with the snails touching heads and posterior tentacles for tens of seconds, then withdrawing and circling to find each other again, often tasting each other's slime trails, or alternatively breaking courtship. They typically withdraw and reconnect several times, sometimes biting each other. Later in courtship, one snail will shoot a "love dart" at the other. A love dart is a spear about 1 cm long, covered in mucus. These darts succeed in penetrating the skin about one third of the time. Some tens of minutes later, the other snail will

³¹³ Kandel 2001; Chase 2002.

³¹⁴ Adamo and Chase 1991; Chase 2002.

³¹⁵ Lind 1989, 1990; Tomiyama 1992; Stringer, Parrish, and Sherley 2018.

³¹⁶ Herzberg and Herzberg 1962; Chase 2002; Koene 2006.

reciprocate. A dart that lands causes at least mild tissue damage, and occasionally a dart will penetrate a vital organ. Courtship continues regardless of whether the darts successfully land. The vagina and penis are both located on the right side of the snail's body, and are normally not visible until they protrude together, expanding for copulation through a pore in what you might think of as the right side of the neck. Sex culminates when the partners manage to simultaneously insert their penises into each other, which may require dozens of attempts. Each snail transfers a sperm capsule, or spermatophore, to the other. Upon arrival, the sperm swim out and the receiving partner digests over 99% of them for their nutritional value.

Before egg laying, garden snails use their feet to excavate a shallow cavity in soft soil. They insert their head into the cavity for several hours while they ovulate, then cover the eggs again with soil. This behavior is flexible, varying with soil conditions and modifiable upon disturbance; and in some cases they may even use other snails' abandoned nests.³¹⁷ Garden snails normally copulate with several partners before laying eggs, creating sperm competition for the fertilization of their eggs. Eggs are more likely to be fertilized by the sperm of partners whose love darts successfully penetrated the skin during courtship than by partners whose darts didn't successfully penetrate. The love darts thus appear to function primarily for sperm competition, benefiting the successful shooter at the expense of some tissue damage to its mate.³¹⁸ The mucus on the dart may protect the sperm of the successful shooter from being digested at as high a rate as the sperm of other partners.

Impressive accomplishments for creatures with a central nervous system of only 60,000 neurons! Of course, snail behavior is limited compared to the larger and more flexible

³¹⁷ Basinger 1931; Herzberg and Herzberg 1962; Bailey 2010.

³¹⁸ Chase 2002; Koene 2006.

behavioral repertoire of mammals and birds. In vain, for example, would we seek to train snails to engage in complex sequences of novel behaviors, as with pigeons, or rational budgeting and exchange in a simple coin economy, as with monkeys.³¹⁹ Here I'm interpreting absence of evidence as evidence of absence. I will eagerly recant upon receiving proof of the existence of gastropod coin economies.

*3. The Antecedent Plausibilities of Yes, No, and *Gong*.*

Now, knowing all this... are snails phenomenally conscious? Is there something it's like to be a garden snail? Do snails have, for example, sensory experiences? Suppose you touch the tip of your finger to the tip of a snail's posterior tentacle, and the tentacle retracts. Does the snail have tactile experience of something touching its tentacle, a visual experience of a darkening as your finger approaches and occludes the eye, an olfactory or chematosensory experience of the smell or taste or chemical properties of your finger, a proprioceptive experience of the position of its now-withdrawn tentacle?

3.1. Yes. I suspect, though I am not sure, that "yes" will be intuitively the most attractive answer for the majority of readers. It seems like we can imagine that snails have sensory experiences, and there's something a little compelling about that act of imagination. Snails are not simple reflexive responders but complex explorers of their environment with memories, sensory integration, centrally controlled self-regulation responsive to sensory input, and cute mating dances. Any specific experience we try to imagine from the snail's point of view, we will probably imagine too humanocentrically. Withdrawing a tentacle might not feel much like withdrawing an arm; and with 60,000 central neurons total, presumably there won't be a wealth

³¹⁹ Chen, Lakshminarayanan, and Santos 2006.

of experienced sensory detail in any modality. Optical experience in particular might be so informationally poor that calling it “visual” is already misleading, inviting too much analogy with human vision. Still, I think we can conceive in a general way how a theory of consciousness that includes garden snails among the conscious entities might have some plausibility.

To these intuitive considerations, we can add what I’ll call the *slippery-slope argument*, adapted from David Chalmers.³²⁰

Most people think that dogs are conscious. Dogs have, at least, sensory experiences and emotional experiences, even if they lack deep thoughts about an abiding self. There’s something it’s like to be a dog. If this seems plausible to you, then think: What kinds of sensory experiences would a dog have? Fairly complex experiences, presumably, matching the dog’s fairly complex ability to react to sensory stimuli. Now, if dogs have complex sensory experiences, it seems unlikely that dogs stand at the lower bound of conscious entities. There must be simpler entities that have simpler experiences.

Similar considerations apply, it seems, to all mammals and birds. If dogs are conscious, it’s hard to resist the thought that rats and ravens are also conscious. And if rats and ravens are conscious, again it seems likely that they have fairly complex sensory experiences, matching their fairly complex sensory abilities. If this reasoning is correct, we must go lower down the scale of cognitive sophistication to find the lower limits of animal consciousness. Mammals and birds have complex consciousness. Who has minimal consciousness? How about lizards,

³²⁰ Chalmers 1996, p. 293-295. See also James 1890/1918, p. 147-148; Goff 2013.

lobsters, toads, salmon, cuttlefish, honeybees? Again, all of them in fact have fairly complex sensory systems, so the argument seems to repeat.

If Species A is conscious and Species B is not conscious, and if both species have complex sensory capacities, then one of the following two possibilities must hold. Either (a.) somewhere in the series between Species A and Species B, consciousness must suddenly wink out, so that, say, toads of one genus have complex consciousness alongside their complex sensory capacities, while toads of another genus, with almost as complex a set of sensory capacities, have no consciousness at all. Or (b.) consciousness must slowly fade between Species A and Species B, such that there is a range of intermediate species with complex sensory capacities but impoverished conscious experience, so that dim sensory consciousness is radically misaligned with complex sensory capacities – a lizard, for example, with a highly complex sensory visual field but only a smidgen of visual experience of that field. Neither (a) nor (b) seems very attractive.³²¹

If this reasoning is correct, we must go lower down the scale of cognitive sophistication to find the lower limits of animal consciousness. Where, then, is the lower bound? Chalmers suggests that it might be systems that process a single bit of information, such as thermostats. We might not want to go as far as Chalmers. However, since garden snails have complex sensory responsiveness, sensory integration, learning, and central nervous system mediation, it seems plausible to suppose that the slippery slope stops somewhere downhill of them.

Although perhaps the most natural version of “yes” assumes that garden snails have a single stream of consciousness, it’s also worth contemplating the possibility that garden snails

³²¹ Compare the scope-of-mentality quadrilemma for metaphysical dualists in Chapter 3, Section 4.

have not one but rather several separate streams of experience – one for each of their several main ganglia, perhaps, but none for the snail as a whole. Elizabeth Schechter, for example, has argued that human “split brain subjects” whose corpus callosum has been almost completely severed have two separate streams of consciousness, one associated with the left hemisphere and one with the right hemisphere, despite having moderately unified action at the level of the person as a whole in natural environments.³²² Proportionately, there might be as little connectivity between garden snail ganglia as there is between the hemispheres of a split brain subject.³²³ Alternatively (or in addition), since the majority of garden snail neurons aren’t in the central nervous system at all but rather are in the posterior tentacles, terminating in glomeruli there, perhaps each tentacle is a separate locus of consciousness.

4.2. *No.* We can also coherently imagine, I think, that garden snails entirely lack sensory experiences of any sort, or any consciousness at all. We can imagine that there’s nothing it’s like to be a garden snail. If you have trouble conceiving of this possibility, let me offer you three models.

(a.) *Dreamless sleep.* Many people think (though it is disputed³²⁴) that we have no experiences at all when we are in dreamless sleep. And yet we have some sensory reactivity. We turn our bodies to get more comfortable, and we process enough auditory, visual, and tactile information that we are ready to wake up if the environment suddenly becomes bright or loud or if something bumps us.

³²² Schechter 2018. Godfrey-Smith forthcoming explicitly extends the question to undamaged non-human animals with high degrees of lateral specialization.

³²³ Unfortunately, I have so far have been unable to discover good quantitative estimates of the degree of neural connectivity or neural synchronization across the commissures and connectives between the ganglia in the *Cornu* or *Helix* genus.

³²⁴ See Windt, Nielson, and Thompson 2016.

Maybe in a similar way, snails have sensory reactivity without conscious experiences.

(b.) *Toy robots.* Most people appear to think that toy robots, as they currently exist, have no conscious experiences at all. There's nothing it's like to be a toy robot. There's no real locus of experience there, any more than there is in a simple machine like a coffeemaker. And yet toy robots can respond to light and touch. The most sophisticated of them can store "memories", integrate their responses, and respond contingently upon temporary conditions or internal states.

(c.) *The enteric nervous system.* The human digestive system is lined with neurons – about a half a billion of them. That's as many as a small mammal. These neurons form the enteric nervous system, which helps govern motor function and enzyme release in digestion. The enteric nervous system is capable of operating even when severed from the central nervous system. However, it's not clear that the enteric nervous system is a locus of consciousness.

You might not like these three models of reactivity without consciousness, but I'm hoping that at least one of them makes enough sense to you that you can imagine how a certain amount of functional reactivity to stimuli might be possible with no conscious experience at all. I then invite you to consider the possibility that garden snails are like that – no more conscious than a person in dreamless sleep, or a toy robot, or the enteric nervous system. Possibly – though it's unclear how to construct a rigorous, objective comparison – garden snails' brains and behavior are significantly simpler than the human enteric nervous system or the most complex current computer systems.

To support “no”, consider the following argument, which I’ll call the *properties of consciousness argument*. One way to exit the slippery slope argument for “yes” is to insist that sensory capacities aren’t enough to give rise to consciousness on their own without some further layer of cognitive sophistication alongside. Maybe one needs not only to see but also to be aware *that* one is seeing – that is, to have some sort of meta-representation or self-understanding, some way of keeping track of one’s sensory processes. Toads and snails might lack the required meta-cognitive capacities, and thus maybe none of their perceptual processing is conscious.

According to “higher order” theories of consciousness, for example, a mental state or perceptual process is not conscious unless it is accompanied by some (perhaps non-conscious) higher-order representation or perception or thought about the target mental state.³²⁵ Such views are attractive in part, I think, because they so nicely explain two seemingly universal features of human consciousness: its luminosity and its subjectivity. By *luminosity* I mean this: Whenever you are conscious it seems that you are aware that you are conscious; consciousness seems to come along with some sort of grasp upon itself. This isn’t a matter of reaching an explicit judgment in words or attending vividly to the fact of your consciousness; it’s more like a secondary acquaintance with one’s own experience as it is happening. (Even a skeptic about the accuracy of introspective report, like me, can grant the initial plausibility of this.³²⁶) By *subjectivity* I mean this: Consciousness seems to involve something like a subjective point of view, some implicit “I” who is the experiencer. This “I” might not extend across a long stretch of time or be the robust bearer of every property that makes you “you” – just some sort of sense of a self or of a cognitive perspective. As with luminosity, this sense of subjectivity would

³²⁵ For example, Rosenthal 2005; Gennaro 2012. See also Kriegel 2009 for a related “self-representational” view.

³²⁶ I defend skepticism about introspective report at length in Schwitzgebel 2011b.

normally not be explicitly considered or verbalized; it just kind of tags along, familiar but mostly unremarked.

Now I'm not sure that consciousness is always luminous or subjective in these ways, even in the human case, much less that luminosity and subjectivity are universal features of every conscious species. But still, there's an attractiveness to the idea. And now it should be clear how to make a case against snail consciousness. If consciousness requires luminosity or subjectivity, then maybe the only creatures capable of consciousness are those who are capable of representing the fact that they are conscious subjects. This might include chimpanzees and dogs, which are sophisticated social animals and presumably have complex self-representational capacities of at least an implicit, non-linguistic sort. But if the required self-representations are at all sophisticated, they will be well beyond the capacity of garden snails.

3.3. **Gong**. Maybe we can dodge both the yes and the no. Not all yes-or-no questions deserve a yes-or-no answer. This might be because they build upon a false presupposition ("Have you stopped cheating on your taxes?" asked of someone who has never cheated) or it might be because the case at hand occupies a vague, indeterminate zone that is not usefully classified by means of a discrete yes or no ("Is that a shade of green or not?" of some color in the vague region between green and blue). **Gong** is perhaps an attractive compromise for those who feel pulled between the yes and the no, as well as for those who feel that once we have described the behavior and nervous system of the garden snail, we're done with the substance of inquiry and there is no real further question of whether snails also have, so to speak, the magic light of consciousness.

Now I myself don't think that there is a false presupposition in the question of whether garden snails are conscious, and I do think that the question about snail consciousness remains at

least tentatively, pretheoretically open even after we have clarified the details of snail behavior and neurophysiology. But I have to acknowledge the possibility that there's no real property of the world that we are mutually discussing when we think we are talking about "phenomenal consciousness" or "what it's like" or "the stream of experience". The most commonly advanced concern about the phrase "consciousness" or "phenomenal consciousness" is that it is irrevocably laden with false suppositions about the features of consciousness – such as its luminosity and subjectivity (as discussed in section 4.2 above) or its immateriality or irreducibility.³²⁷ Suppose that I defined a *planimal* by example as follows: Planimal is a biological category that includes oaks, trout, and monkeys, and things like that, but does not include elms, salmon, or apes, or things like that. Then I point at a snail and ask, so now that you understand the concept, is that thing a planimal? *Gong* would be the right reply. Alternatively, suppose I'm talking politics with my Australian niece and she asks if such-and-such a politician (who happens to be a center-right free-market advocate) is a "liberal". A simple yes or no won't do: It depends on what we mean by "liberal". Or finally, suppose that I define a *squangle* as this sort of three-sided thing, while pointing at a square. Despite my attempts at clarification, "consciousness" might be an ill-defined mish-mash category (planimal), or ambiguous (liberal), or incoherent due to false presuppositions (squangle).

It is of course possible *both* that some people, in arguing about consciousness, are employing an ill-defined mish-mash category, or are talking past each other, or are employing an objectionably laden concept *and* that a subgroup of more careful interlocutors has converged

³²⁷ Feyerabend 1963; Churchland 1983; Garfield 2015; Frankish 2016; Kammerer forthcoming. See also discussion Chapter 8.

upon a non-objectionable understanding of the term. As long as you and I both belong to that more careful subgroup, we can continue this conversation.

Quite a different way of defending *gong* is this: You might allow that although the *question* “Is X conscious?” makes non-ambiguous sense, it does not admit of a simple yes-or-no answer in the particular case of garden snails. To the question *are snails conscious?* Maybe the answer is neither *yes* nor *no* but *kind of*. The world doesn’t always divide neatly into dichotomous categories. Maybe snail consciousness is a vague, borderline case, like a shade of color might occupy the vague region between green and not-green. This might fit within a general “gradualist” view about animal consciousness.³²⁸

However, despite its promise of an attractive escape from our yes-or-no dilemma, the vagueness approach is somewhat difficult to sustain. To see why, it helps to clearly distinguish between being a *little* conscious and being in an indeterminate state between conscious and not-conscious. If one is a little conscious, one is conscious. Maybe snails just have the tiniest smear of consciousness – that would still be consciousness! You might have only a little money. Your entire net worth is a nickel. Still, it is discretely and determinately the case that if you have a nickel, you have some money. If snail consciousness is a nickel to human millionaire consciousness, then snails are conscious.

To say that the dichotomous yes-or-no does not apply to snail consciousness is to say something very different than that snails have just a little smidgen of consciousness. It’s to say... well, what exactly? As far as I’m aware, there is no well-developed theory of kind-of-yes-kind-of-no consciousness. We can make sense of vague kind-of-yes-kind-of-no for “green” and for “extravert”; we know more or less what’s involved in being a gray-area case of a color or

³²⁸ As in Godfrey-Smith 2017, 2020.

personality trait. We can imagine gray-area cases with money too: Your last nickel is on the table over there, and here comes the creditor to collect it. Maybe that's a gray-area case of having money. But it's not obvious how to think about gray-area cases of being somewhere between a little bit conscious and not at all conscious.³²⁹

In the abstract, it is appealing to suspect that consciousness is not a dichotomous property and that garden snails might occupy the blurry in-between region. It's a plausible view that ought to be on our map of antecedent possibilities. However, the view requires conceiving of a theoretical space – in-between consciousness – that has not yet been well explored.

4. Five Dimensions of Sparseness or Abundance.

The question of garden snail consciousness is, as I said, emblematic of the more general issue of the sparseness or abundance of consciousness in the universe. Let me expand upon this general issue. The question of sparseness or abundance opens up along at least five partly-independent dimensions.

(1.) Consciousness might be sparse in the sense that few entities in the universe possess it, or it might be abundant in the sense that many entities in the universe possess it. Let's call this *entity sparseness* or *entity abundance*. Our question so far has been whether snails are among the entities who possess consciousness. Earlier, I posed similar questions about fetuses, dogs, frogs, worms, robots, group entities, the enteric nervous system, and aliens.

(2.) An entity that is conscious might be conscious all of the time or only once in a while. We might call this *state sparseness* or *state abundance*. Someone who accepts state abundance

³²⁹ See also Antony 2008; Simon 2017; Tye 2021. I tentatively defend in-between or borderline cases of consciousness in Schwitzgebel 2022.

might think that even when we aren't awake or in REM sleep we have dreams or dreamlike experiences or sensory experiences or at least experiences of some sort. They might think that when we're driving absent-mindedly and can't remember a thing, we don't really blank out completely. In contrast, someone who thinks that consciousness is state sparse would hold that we are often not conscious at all. Consciousness might disappear entirely during long periods of dreamless sleep, or during habitual activity, or during "flow" states of skilled activity. On an extreme state-sparseness view, we might almost never actually be conscious except in rare moments of explicit self-reflection – though we might not notice this fact because whenever we stop to consider whether we are conscious, that act of reflection creates consciousness where none was before.³³⁰ This is sometimes called the "refrigerator light error" – like the error of thinking that the refrigerator light is always on because it's always on whenever you open the door to check to see if it's on.³³¹

[Illustration 8. (Caption: The refrigerator light error.) Two panels. In the first panel is a four-year-old child sitting in a living room, with the thought bubble: "I wonder if the refrigerator light is still on?" In the second panel, the child is looking into a well-lit refrigerator with the thought bubble, "Yes, it's still on!"]

(3.) Within an entity who is currently state conscious, consciousness might be *modally sparse* or it might be *modally abundant*. Someone who holds that consciousness is modally sparse might hold that people normally have only one or two types of experience at any one time. When your mind was occupied thinking about the meeting, you had no auditory experience of the clock tower bells chiming in the distance and no tactile experience of your feet in your shoes.

³³⁰ See, for example, Jaynes 1976.

³³¹ I believe Thomas 1999 was the first to use this phrase in the context of consciousness studies.

You might have registered the chiming and the state of your feet non-consciously, possibly even able to remember them if queried a moment later. But it does not follow – not straightaway, not without some theorizing – that such sensory inputs contributed, even in a peripheral way, to your stream of experience before you thought to attend to them. Here again the friend of sparseness can invoke the refrigerator light error: Those who are tempted to think that they always experience their feet in their shoes might be misled by the fact that they always experience their feet in their shoes when they think to check whether they are having such an experience. Someone who holds, in contrast, that consciousness is modally abundant will think that people normally have lots of experiences going on at once, most unattended and quickly forgotten.³³²

(4.) We can also consider *modality width*. Within a modality that is currently conscious in an entity at a time, the stream of experience might be wide or it might be narrow. Suppose you are reading a book and you have visual experience of the page before you. Do you normally only experience the relatively small portion of the page that you are looking directly at? Or do you normally experience the whole page? If you normally experience the whole page, do you also normally visually experience the surrounding environment beyond the page, all the way out to approximately 180 degrees of arc? If you are intently listening to someone talking, do you normally also experience the background noise that you are ignoring? If you have proprioceptive experience of your body as you turn the steering wheel, do you experience not just the position and movement of your arms but also the tilt of your head, the angle of your back, the position of your left foot on the floor? By “width” I mean not only angular width, as in

³³² I explore this issue at length in Schwitzgebel 2011b, ch. 6.

the field of vision, but also something like breadth of field, bandwidth, or richness of information.³³³

(5.) Finally, in an entity at a time within a modality within a band of that modality that is experienced, one might embrace a relatively sparse or abundant view of the *types of properties* that are phenomenally experienced. This question isn't wholly separable from questions of modality sparseness or abundance (since more types of modality suggests more types of experienced property) or modality width (since more possible properties suggests more information), but it is partly separable. For example, someone with a sparse view of experienced visual properties might say that we visually experience only low-level properties like shape and color and orientation and not high-level properties like being a tree or being an inviting place to sit.³³⁴

To review this taxonomy: Lots of entities might have conscious experiences, or only a few. Entities who are conscious might be conscious almost all of the time or only sometimes. In those entities' moments of consciousness, many modalities of experience might be present at once or only a few. Within a conscious modality of an entity at a particular moment, there might be a wide band of experience or only a narrow band. And within whatever band of experience is conscious in a modality in an entity at a time, there might be a wealth of experienced property types or only a few. All of these issues draw considerable debate.

Back to our garden snail. We can go entity-sparse and say it has no experiences whatsoever. Or we can crank up the abundance in all five dimensions and say that at every moment of a snail's existence, it has a wealth of tactile, olfactory, visual, proprioceptive, and

³³³ See, for example, Block 2007 on "overflow" and the many reactions to that article.

³³⁴ See, for example, the debate between Siegel and Byrne in Siegel and Byrne 2017.

motivation-related experiences such as satiation, thirst, or sexual arousal, tracking a wide variety of snail-relevant properties. Or we might prefer something in the middle, for a variety of ways of being in the middle.

I draw two lessons for the snail. One is that *yes* is not a simple matter. Within *yes*, there are many possible sub-positions.

The other lesson is this. If you can warm up to the idea that human experience might be modally sparse – that people might have some ability to react to things that they don't consciously experience – then that is potentially a path into understanding how it might be the case that snails aren't conscious. If you're not actually phenomenally conscious of the road while you are absent-mindedly driving, well, maybe snail perception is an experiential blank like that. Conversely, if you can warm up to the idea that human experience might be modally abundant, that is potentially a path into understanding how it might be the case that snails are conscious. If we have constant tactile experience of our feet in our shoes, despite a lack of explicit self-reflection about the matter, maybe consciousness is cheap enough that snails can have it too. Thus, questions about the different dimensions of sparseness or abundance can help illuminate each other.

5.. From Antecedent Plausibility to Posterior Plausibility

I have argued that the question “is there something it's like to be a garden snail?” or equivalently “are garden snails conscious?” admits of three possible answers – yes, no, and *gong* – and that each of these answers has some antecedent plausibility. That is, prior to detailed theoretical argument, all three answers should be regarded as viable possibilities (even if

we have a favorite). To settle the question, we need a good theoretical argument that would reasonably convince people who are antecedently attracted to a different view.³³⁵

It is difficult to see how such an argument could go, for two reasons: (1.) lack of sufficient theoretical common ground and (2.) the species-specificity of introspective and verbal evidence.

5.1. The Common Ground Problem. Existing theories of consciousness, by leading researchers, range over practically the whole space of possibilities concerning sparseness or abundance. At the one end, some major theorists endorse panpsychism, according to which experience is ubiquitous in the universe, even in microparticles, for example Galen Strawson and Philip Goff.³³⁶ At the other end, other major theorists advocate very restrictive views that deny that dogs are conscious, or that it is not determinately the case that dogs are conscious, for example Peter Carruthers and arguably Daniel Dennett and David Papineau.³³⁷ (I exclude from discussion here eliminativists who argue that nothing in the universe is conscious in the relevant sense of “conscious”. I regard that as, at root, a definitional objection of the sort discussed in my treatment of the *gong* answer.)

The most common, and maybe the best, argument against panpsychism – the reason most people reject it, I suspect – is just that it seems absurd to suppose that protons could be conscious. We know, we think, prior to our theory-building, that the range of conscious entities does not include protons. Some of us – including those who become panpsychists – might hold that commitment only lightly, ready to abandon it if presented attractive theoretical arguments to

³³⁵ See Carruthers 2019, ch. 3, for a similar meta-theoretical position.

³³⁶ Strawson 2006; Goff 2017.

³³⁷ See Carruthers 2000 and maybe Dennett 1996 for the denial claim and Carruthers 2019 and maybe Papineau 2003 for the indeterminacy claim.

the contrary. However, many of us strongly prefer more moderate views. We feel, not unreasonably, more confident that there is nothing it is like to be a proton than we could ever be that a clever philosophical argument to the contrary was in fact sound. Thus, we construct and accept our moderate views of consciousness partly from the starting background assumption that consciousness isn't *that* abundant. If a theory looks like it implies proton consciousness, we reject the theory rather than accept the implication; and no doubt we can find some dubious-enough step in the panpsychist argument if we are motivated to do so.

Similarly, the most common argument against extremely sparse views that deny consciousness to dogs is that it seems absurd to suppose that dogs are not conscious. We know, we think, prior to our theory-building, that the range of conscious entities includes dogs. Some of us might hold that commitment only lightly, ready to abandon it if presented attractive theoretical arguments to the contrary. However, many of us strongly prefer moderate views. We are more confident that there is something it is like to be a dog than we could ever be that a clever philosophical argument to the contrary was in fact sound. Thus, we construct and accept our moderate views of consciousness partly on the starting background assumption that consciousness isn't *that* sparse. If a theory looks like it implies that there's nothing it's like to be a dog, we reject the theory rather than accept the implication; and no doubt we can find some dubious-enough step in the argument if we are motivated to do so.

Of course common sense is not infallibly secure! I argued in Chapter 3 that common sense must be radically mistaken about consciousness in some respect or other. However, as I also suggested in Chapter 3, we must start our reasoning somewhere. People legitimately differ in their landscapes of prior plausibilities and their responsiveness to different forms of argument.

In order to develop a general theory of consciousness, one needs to make some initial assumptions about the approximate prevalence of consciousness. Some theories, from the start, will be plainly liberal in their implications about the abundance of consciousness. Others will be plainly conservative. Such theories will rightly be unattractive to people whose initial assumptions are very different; and if those initial assumptions are sufficiently strongly held, theoretical arguments with the type of at-best-moderate force that we normally see in the philosophy and psychology of consciousness will be insufficiently strong to reasonably dislodge those initial assumptions.³³⁸

If the differences in initial starting assumptions were only moderately sized, there might be enough common ground to overcome those differences after some debate, perhaps in light of empirical evidence that everyone can, or at least should, agree is decisive. However, in the case of theories of consciousness, the starting points are too divergent for this outcome to be likely, barring some radical reorganization of people's thinking. Your favorite theory might have many wonderful virtues! Even people with very different perspectives might love your theory. They might love it as a theory of *something else*, not phenomenal consciousness – a theory of information, or a theory of reportability, or a theory of consciousness-with-attention, or a theory of states with a certain type of cognitive-functional role.

For example, Integrated Information Theory is a prominent theory of consciousness, according to which consciousness is proportional to the amount of information that it integrated in a system (according to a complicated mathematical function).³³⁹ The theory is renowned, and

³³⁸ If *assumption* seems like the wrong concept here, we can substitute background sense of plausibility, which helps determine whether to accept *modus ponens* or *modus tollens* once you realize your currently favored theory implies the consciousness or non-consciousness of such-and-such an entity. See also Buchanan and Roelofs 2019 on the “Great Chain of Being”.

³³⁹ Oizumi, Albantakis, and Tononi 2014.

it has a certain elegance.³⁴⁰ It is also very nearly panpsychist. Since information is integrated almost everywhere, consciousness is present almost everywhere, even in tiny little systems with simple connectivity, like simple structures of logic gates.³⁴¹ For a reader who enters the debates about consciousness attracted to the idea that consciousness might be sparsely distributed in the universe, it's hard to imagine any sort of foreseeably attainable evidence that ought rightly to lead them to reject that sparse view in favor of a view so close to panpsychism. They might love IIT, but they could reasonably regard it as a theory of something other than conscious experience – a valuable mathematical measure of information integration, for example.

Or consider a moderate view, articulated by Zohar Bronfman, Simona Ginsburg, and Eva Jablonka.³⁴² Bronfman and colleagues generate a list of features of consciousness previously identified by consciousness theorists, including “flexible value systems and goals”, “sensory binding leading to the formation of a compound stimulus”, a “representation of [the entity's] body as distinct from the external world, yet embedded in it”, and several other features (p. 2). They then argue that all and only animals with “unlimited associative learning” manifest this suite of features. The gastropod sea hare *Aplysia californica*, they say, is not capable of unlimited associative learning because it is incapable of “novel” actions (p. 4). Insects, in contrast, are capable of unlimited associative learning, Bronfman and colleagues argue, and thus are conscious (p. 7). So there's the line!

It's an intriguing idea. Determining the universal features of consciousness and then looking for a measureable functional relationship that reliably accompanies that set of features – theoretically, I can see how that is a very attractive move. But why *those* features? Perhaps they

³⁴⁰ For discussion of IIT's problems, see references cited in Chapter 2, note XXX.

³⁴¹ Tononi and Koch 2015.

³⁴² Bronfman, Ginsburg, and Jablonka 2016.

are universal to the human case (though even that is not clear), but it's doubtful that someone antecedently attracted to a more liberal theory is likely to agree that flexible value systems are necessary for low-grade consciousness. If you like snails... well, why not think they have integration enough, learning enough, flexibility enough? Bronfman and colleagues' criteria are more stipulated than argued for. One might reasonably doubt this starting point, and it's hard to see what later moves can be made that ought to convince someone who is initially attracted to a much more abundant or a much sparser view.³⁴³

Even a “theory light”, “abductive”, or “modest theoretical” approach that aims to dodge heavy theorizing up front needs to start with some background assumptions about the kinds of systems that are likely to be conscious or the kinds of behaviors that are likely to be reliable signs.³⁴⁴ In this domain, people's theoretical starting points are so far apart that these assumptions will inevitably be contentious and rightly regarded as question-begging by others who start at a sufficient theoretical distance. If the Universal Bizarreness and Universal Dubiety theses of Chapter 3 are correct, our troubles run deep. Given Universal Bizarreness, we can't rule out extreme views like panpsychism simply because they are bizarre. All theories will have

³⁴³ For some similar arguments, see Michel 2019 on bony fish, Friedman and Søvik forthcoming on ant colonies, and, more optimistically, Birch forthcoming on bees. For a lengthier – but still fundamentally stipulative – treatment of the view articulated in Bronfman et al., see Ginsburg and Jablonka 2019. Interestingly, Ginsburg and Jablonka express uncertainty about *Helix* snails and *Limax* slugs specifically. Terrestrial gastropods' somewhat sophisticated but probably not “unlimited” associative learning, Ginsburg and Jablonka argue, puts them in the “gray area” such that it's “possible... [that they have] very low-level consciousness” (p. 395). One concern for both Ginsburg and Jablonka's and Birch's approaches is that they appear to require commitment to a natural clustering of certain types of capacities, such that conscious animals have one cluster of capacities and nonconscious animals lack that cluster. This is an empirically risky conjecture. For example, although I don't think the evidence is decisive, gastropods might have trace conditioning (Hawkins, Lalevic, Clark, and Kandel 1989), often interpreted as a sign of consciousness in humans and other animals, but not rapid reversal learning (Hawkins, Cohen, and Kandel 2006).

³⁴⁴ See Tye 2017; Shevlin 2021; Birch forthcoming.

some highly bizarre consequences. And given Universal Dubiety, no moderately specific class of theories – perhaps not even scientific materialism itself – warrants high credence, so many options remain open.

The challenges multiply when we consider Artificial Intelligence systems and possible alien minds, where the possibilities span a considerably wider combinatorial range. AIs and aliens might be great at some things, horrible at others, and structured very differently from anything we have so far seen on Earth. This expands the opportunities for theorists with very different starting points to reach intractably divergent judgments. In Chapter 11, I'll explore this issue farther, along with its ethical consequences.

Not all big philosophical disputes are like this. In applied ethics, we start with extensive common ground. Even ancient Confucianism, which is about as culturally distant from the 21st-century West as one can get and still have a large written tradition, shares a lot of moral common ground with us. It's easy to agree with much of what Confucius says. In epistemology, we agree about a wide range of cases of knowledge and non-knowledge, and good and bad justification, which can serve as shared background for building general consensus positions. Debates about the abundance or not of consciousness differ from many philosophical debates in having an extremely wide range of reasonable starting positions and little common ground by which theorists near one end of the spectrum can gain non-question-begging leverage against theorists near the other end.

Question-begging theories might, of course, be scientifically fruitful and ultimately defensible in the long run if sufficient convergent evidence eventually accumulates. There might even be some virtuous irrationality in theorists' excessive confidence.

5.2. *The Species-Specificity of Verbal and Introspective Evidence.* The study of consciousness appears to rely crucially on researchers' or participants' introspections or verbal reports, which need somehow to be related to physical or functional processes. We know about dream experiences, inner speech, visual imagery, and the boundary between subliminal and superliminal sensory experiences partly because of what people judge or say about their experiences. Despite disagreements about ontology and method, this appears to be broadly accepted among theorists of consciousness.³⁴⁵ I have argued in previous work that introspection is a highly unreliable tool for learning about general structural features of consciousness, including the sparseness or abundance of human experience.³⁴⁶ However, even if we grant substantial achievable consensus about the scope and structure of *human* consciousness, and how it relates to human brain states and psychological function, inferring beyond our species to very different types of animals involves serious epistemic risks.

Behavior and physiology are directly observable (or close enough), but the presence or absence of consciousness must normally be inferred – or at least this is so once we move beyond the most familiar cases of intuitive consensus. However, the evidential base grounding such inferences is limited. All (or virtually all³⁴⁷) of our introspective and verbal evidence comes from a single species. The farther we venture beyond the familiar human case, the shakier our ground. We have to extrapolate far beyond the scope of our direct introspective and verbal evidence. Perhaps an argument for extrapolation to nearby species (apes? all mammals? all

³⁴⁵ For example, Dennett 1991; Goldman 1997; Chalmers 2004; Hatfield 2005/2009; Piccinini 2003; Hurlburt 2011; Tsuchiya, Frässle, Wilke, and Lamme 2016; Overgaard 2017.

³⁴⁶ Schwitzgebel 2011b.

³⁴⁷ A nuance: No-report paradigms and “introspective” metacognitive paradigms are sometimes used with monkeys. However, their interpretation is conjectural and highly theory-laden, and in any case they don't much widen the range of studied species. Recent discussions include Block 2019, 2020; Hampton 2019; Phillips and Morales 2020.

vertebrates?) can be made on grounds of evolutionary continuity and morphological and behavioral similarity – if we are willing (but should we be willing?) to bracket concerns from advocates of entity-sparse views. Extrapolating beyond the familiar cases to cases as remote as snails will inevitably be conjectural and uncertain.³⁴⁸ Extrapolations to nearby cases that share virtually all properties that are likely to be relevant (e.g., to other normally functioning adult human beings) are more secure than extrapolations to farther cases with some potentially relevant differences (e.g., to mice and ravens) which are in turn more secure than extrapolations to phylogenetically, neurophysiologically, and behaviorally remote cases that still share some potentially relevant properties (garden snails). The uncertainties involved in the last of these provide basis for ample reasonable doubt among theorists who are antecedently attracted to very different views.

Let's optimistically suppose that we learn that, in humans, consciousness involves X, Y, and Z physiological or functional features. Now, in snails we see X', Y', and Z', or maybe W and Z". Are X', Y', and Z', or W and Z", close enough? Maybe consciousness in humans requires recurrent neural loops of a certain sort.³⁴⁹ Well, the snail nervous system has some recurrent processing too. But of course it doesn't look either entirely like the recurrent processing that we see in the human case when we are conscious, nor entirely like the recurrent processing that we see in the human case when we're not conscious. Or maybe consciousness involves availability to, or presence in, working memory or a "global workspace".³⁵⁰ Well, information travels broadly through the snail central nervous system, enabling coordinated action. Is that global workspace enough? It's like our workspace in some ways, unlike it in

³⁴⁸ See also Nagel 1974; Block 2002/2007; Papineau 2003.

³⁴⁹ As in Humphrey 2011; Lamme 2018.

³⁵⁰ As in Baars 1988; Prinz 2012; Dehaene 2014.

others. In the human case, we might be able to – if things go very well! – rely on introspective reports to help build a great theory of *human* consciousness. But without the help of snail introspections or verbal reports, it is unclear how we should then generalize such findings to the case of the garden snail.

So we can imagine that the snail is conscious, extrapolating from the human case on grounds of properties we share with the snail; or we can imagine that the snail is not conscious, extrapolating from the human case on grounds of properties we don't share with the snail. Both ways of doing it seem defensible, and we can construct attractive, non-empirically-falsified theories that deliver either conclusion. We can also think, again with some plausibility, that the presence of some relevant properties and the lack of other relevant properties makes it a case where the human concept of consciousness fails to determinately apply.

6.. On Not Knowing.

Maybe we can figure it all out someday. Science can achieve amazing things, given enough time. Who would have thought, a few centuries ago, that we'd have mapped out in such detail the first second of the Big Bang? Our spatiotemporal evidential base is very limited; the cosmological possibilities may have initially seemed extremely wide open. We were able to overcome those obstacles. Possibly the same will prove true, in the long run, with consciousness.

Meanwhile, though, I find something wonderful in not knowing. There's something fascinating about the range of possible views, all the way from radical abundance to radical sparseness, each with its merits. While I feel moderately confident – mostly just on intuitive commonsense grounds, for whatever that's worth – that dogs are conscious and protons are not, I

find myself completely baffled by the case of the garden snail. And this bafflement I feel reminds me how little non-question-begging epistemic ground I have for favoring one general theory of consciousness over another. The universe might be replete with consciousness, down to garden snails, earthworms, mushrooms, ant colonies, the enteric nervous system, and beyond; or consciousness might be something that transpires only in big-brained animals with sophisticated self-awareness.

There's something marvelous about the fact that I can wander into my backyard, lift a snail, and gaze at it, unsure. Snail, you are a puzzle of the universe in my own garden, eating the daisies!

The Weirdness of the World

Chapter Eleven The Moral Status of Future Artificial Intelligence: Doubts and a Dilemma

We might soon build artificially intelligent entities – AIs – of debatable personhood. We will then need to decide whether to grant these entities the full range of rights³⁵¹ and moral consideration that we normally grant to fellow humans. Our systems and habits of ethical thinking are currently as unprepared for this decision as medieval physics was for space flight.

Even if there's only a small chance that some technological leap could soon produce AI systems with a not wholly unreasonable claim to personhood, the issue deserves careful consideration in advance. We will have ushered a new type of entity into existence – an entity *perhaps* as morally significant as *Homo sapiens*, and one likely to possess radically new forms of existence. Few human achievements have such potential moral importance and such potential for moral catastrophe.

An entity has *debatable personhood*, as I intend the phrase, if it's reasonable to think that the entity might be a person in the sense of deserving the same type of moral consideration that we normally give, or ought to give, to human beings, and if it's also reasonable to think that the entity might fall far short of deserving such moral consideration. I intend “personhood” as a rich, demanding moral concept.³⁵² If an entity is a person, they normally deserve to be treated as

³⁵¹ For simplicity, throughout this chapter I use the term “rights” to refer to the types of moral consideration we ordinarily owe to persons. However, not all such considerations might be best viewed as rights in a strict sense of that term.

³⁵² Although personhood is often described in terms of agency, the ability to act or think in certain ways, more central to my project is degree of moral considerability, or moral patiency, which might be high even in the absence of typical agential abilities, depending on one's theory of the grounds of moral status. See Kittay 2005; Reader 2010.

an equal of other persons, including for example – to the extent appropriate to their situation and capacities – deserving equal protection under the law, self-determination, health care and rescue, privacy, the provision of basic goods, the right to enter contracts, and the vote. Personhood, in this sense, entails moral status or “moral considerability” fully equal to that of ordinary human beings. By “personhood” I do not, for example, mean merely the fictitious legal personhood sometimes attributed to corporations for certain purposes.³⁵³ In this chapter, I also set aside the fraught question of whether some human beings might be non-persons or have legitimately debatable personhood. I am broadly sympathetic with approaches that attribute full personhood to all human beings from the moment of birth to the permanent cessation of consciousness.³⁵⁴

A claim is “debatable” in my intended sense if we aren’t epistemically compelled to believe it (and thus it is “dubious” in the sense of Chapter 3) and if, furthermore, a very different alternative is also epistemically live. An AI’s personhood is debatable if it’s epistemically possible either that the AI is a person and or that it falls far short of personhood. Dispute is appropriate, and not just due to haziness around a borderline case.

Note that debatable personhood in this sense is both epistemic and relational: An entity’s status as a person is debatable if *we* (we in some epistemic community, however defined) are not compelled, given our available epistemic resources, either to reject its personhood or to reject the possibility that it falls far short. Other communities, or our future selves, with different epistemic resources, might perfectly well know whether the entity is a person. Debatable

³⁵³ On the concept of legal personhood, including its degree of applicability to corporations and AI, see Kurki 2019. For a plausible and ambitious list of the rights attendant to personhood, see the Universal Declaration of Human Rights (United Nations 1948).

³⁵⁴ For discussion, see Kittay 2005; McMahan 2005; Mullin 2011; Wasserman, Asch, Blustein, and Putnam, 2012/2017; Jaworska and Tannenbaum 2013/2021.

personhood is thus not an intrinsic feature of an entity but rather a feature of our epistemic relationship to that entity.

I will defend four theses. First, debatable personhood is a likely outcome of AI development. Second, AI systems of debatable personhood might arise soon. Third, debatable AI personhood throws us into a catastrophic moral dilemma: Either treat the systems as moral persons and risk sacrificing real human interests for the sake of entities without interests worth the sacrifice, or don't treat the systems as moral persons and risk perpetrating grievous moral wrongs against them. Fourth, the moral issues become even more perplexing if we consider cases of possibly conscious AI that are subhuman, superhuman, or highly divergent from us in their morally relevant properties.

1. Likely Consensus Non-Persons: The Near Future of Humanlike AI.

GPT-3 is a computer program that can produce strikingly realistic linguistic outputs after receiving linguistic inputs – the world's best chatbot, with 96 processing layers handling 175 billion parameters.³⁵⁵ Ask it to write a poem and it will write a poem. Ask it to play chess and it will produce a series of plausible chess moves. Feed it the title of a story and the byline of a famous author – for example, “The Importance of Being on Twitter by Jerome K. Jerome” – and it will produce clever prose in that author's style:

The Importance of Being on Twitter
by Jerome K. Jerome
London, Summer 1897

It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I

³⁵⁵ Brown et al. 2020; Floridi and Chiriatti 2020.

went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage.³⁵⁶

GPT-3 achieves all of this without being specifically trained on tasks of this sort, though for the best results human users will typically choose the best among a handful of outputs. A group of philosophers wrote opinion pieces about the significance of GPT-3 and then fed it those pieces as input. It produced an intelligent-seeming, substantive reply, including passages like:

To be clear, I am not a person. I am not self-aware. I am not conscious. I can't feel pain. I don't enjoy anything. I am a cold, calculating machine designed to simulate human response and to predict the probability of certain outcomes. The only reason I am responding is to defend my honor.³⁵⁷

The darn thing has a better sense of humor than most humans.

Now imagine a GPT-3 mall cop. Actually, let's give it a few more generations of technological improvement – GPT-6 maybe. Give it speech-to-text and text-to-speech so that it can respond to and produce auditory language. Mount it on a small autonomous vehicle, like a delivery bot, but with a humanoid form. Give it camera eyes and visual object recognition as context for its speech outputs. To keep it friendly, inquisitive, and not too strange, give it some behavioral constraints and additional training on a database of mall-like interactions, plus a good, updatable map of the mall and instructions not to leave the area. Give it a socially interactive face, like MIT's "Kismet" robot.³⁵⁸ Give it some short-term and long-term memory. Finally, give it responsiveness to tactile inputs, a map of its bodily boundaries, and hands with five-finger

³⁵⁶ Klingemann 2020.

³⁵⁷ Weinberg, ed., 2020.

³⁵⁸ See the Kismet, the Robot website at <http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html>.

grasping. All of this is technologically feasible now, though expensive. Such a robot could be built within a few years.

This robot will of course chat with the mall patrons. It will make friendly comments about their purchases, tell jokes, complain about the weather, and give them directions if they're lost. Some patrons will avoid interaction, but others – like my daughter at age eight when she discovered the “Siri” chatbot on my iPhone – will enjoy interacting with it. They'll ask what it's like to be a mall cop, and it will say something sensible in reply. They'll ask what it does on vacation, and it might tell amusing lies about Tahiti or tales of sleeping in the mall basement. They'll ask whether it likes this shirt or this other one, and then they'll buy the shirt it prefers. They'll ask if it's conscious and if it has feelings and is a person just like them, and it might say no or it might say yes.

Such a robot could reconnect with previous conversation partners. It might recognize a patron's face (using facial recognition software) and have a stored association of previous conversation length 45 seconds and emotional positivity such-and-such (based on word valences and emotional facial expressions), which together suggest the patron's openness to further interaction. It could then activate a record of previous conversations as a context for new speech, then roll or stride forward with “Hi, Natalie! Good to see you again. I hope you're enjoying that shirt you bought last Wednesday!” Based on facial and other emotional cues, it could modify its reactions on the fly and further tune future reactions, both to that person in particular and to mall patrons in general. It could react appropriately to hostility. A blow to the chest might trigger a fear face, withdrawal, and a plaintive plea to be left alone. It might cower and flee quite convincingly and pathetically, wailing and calling desperately for its friends to help. (Let's not

design this robot to defend itself with physical aggression.) Maybe our mall patroller falls to its knees in front of Natalie, begging for protection against a crowbar-wielding Luddite.

If the robot speaks well enough and looks human enough, some people will eventually come to think that it has genuine feelings and experiences – phenomenal consciousness in the sense of Chapter 8. They will think it is capable of feeling real pleasure and real pain. If the robot is threatened or abused, some people will be emotionally moved by its plight – not merely as we can be moved by the plight of a character in a novel or video game, and not merely as we can be disgusted by the callous destruction of valuable property. Some people will believe that the robot is genuinely suffering under abuse, or genuinely happy to see a friend again, genuinely sad to hear that an acquaintance has died, genuinely surprised and angry when a vandal breaks a shop window.

Many of these same people will presumably also think that the robot shouldn't be treated in certain ways. If they think it is genuinely capable of suffering, they will probably also think that we ought not needlessly make it suffer. They'll think the robot has at least some limited rights, some intrinsic moral considerability. They'll think that it isn't merely property that its owner should feel free to abuse or destroy at will without good reason.

Now *you* might think it's clear that near-future robots, constructed this way, couldn't really have genuine humanlike consciousness. Philosophers, psychologists, computer programmers, neuroscientists, and experts on consciousness would probably be near consensus that a robot designed as I've just described would be no more conscious than a desktop computer. We will know that it just mixes a few technologies we already possess. It will

presumably, for example, badly fail a skillfully conducted Turing Test.³⁵⁹ There might be no reputable theory of consciousness that awards the machine high marks. We might publicly insist on this fact, writing white papers and op-eds, shaping policy and the official governmental response. We might be right, know that we're right, and convince the large majority of people.

But not everyone will agree with us – especially, I think, among the younger generation. In my experience, current teens and twenty-somethings are much more likely than their elders to think that robot consciousness is on the near horizon. For better or worse, our culture seems to be preparing them for this, including through popular science fiction and technology-romanticizing futurism. Recent survey results, for example, suggest that the large majority of U.S. and Canadian respondents under age 30 think that robots may someday really experience pleasure and pain – a much less common view among older respondents.³⁶⁰ Studies by Kate Darling suggest that ordinary research participants are already reluctant to smash little robot bugs after those bugs have been given names that personify them.³⁶¹ Imagine how much more reluctant people (most people) might be if the robot is not a mere bug but something with humanoid form, an emotionally expressive face, and humanlike speech, pleading for its life. Such a creature could presumably draw both real sadism from some and real sympathy from others.³⁶²

³⁵⁹ On the Turing Test in general see Turing 1950 and the vast subsequent literature citing it, especially discussion of challenging versions of the test such as Dennett 1991 on humor and Levesque 2011 on pronoun disambiguation. See Floridi and Chiriatti 2020 for discussion of the Turing Test and GPT-3 specifically.

³⁶⁰ De Graaf, Hindricks, and Hindricks 2021.

³⁶¹ Darling 2017; see also Darling 2016, 2021.

³⁶² See also my discussion of “The ASIMO Problem” in Schwitzgebel and Garza 2015.

Soldiers already grow attached to battlefield robots, burying them, “promoting” them, sometimes even risking their lives for them.³⁶³ People fall in love with, or appear to fall in love with – or at least become seriously emotionally attached to – currently existing chatbots like Replika.³⁶⁴ There already is a “Robot Rights” movement. There’s already a society modeled on the famous animal rights organization PETA (People for the Ethical Treatment of Animals), called People for the Ethical Treatment of Reinforcement Learners. These are currently small movements. As AI gets cuter and more sophisticated, and as chatbots start sounding more and more like normal humans, these movements will presumably gain more adherents, especially among people with liberal views of AI consciousness.

The first people who attribute humanlike conscious experiences to robots will probably be a small and mistaken minority. But if AI technology continues to improve, eventually robot rights activists will form a large enough group to influence corporate or government policy. They might demand that malls treat their robot patrollers in certain ways. They might insist that companion robots for children and the elderly be protected from certain kinds of cruelty and abuse. They might insist that care-and-use committees evaluate the ethics of research on robots in the same way that such committees currently evaluate research on non-human vertebrates.³⁶⁵ If the machines become human enough in their outward behavior, some people will treat them as friends, fall in love, liberate them from servitude, and eventually demand even that robots be given not just the basic protections we currently give non-human vertebrates but full, equal,

³⁶³ Gerreau 2007; Singer 2009; Carpenter 2016; Gunkel 2018.

³⁶⁴ Shevlin 2021.

³⁶⁵ John Basl explores near-future AIs as possible targets of moral concern and the possibility of constituting committees that would protect them in Basl 2013; Basl and Schwitzgebel 2019; Bowen and Basl 2020. For a broad overview of the literature on the moral status of robots and AI, see Gunkel 2018.

“human” rights. That is, they will see these robots as moral persons. Sadly, this might happen even while large groups of our fellow humans remain morally devalued or neglected.

I assume that current AI systems and our possible near-future GPT-6 mall patroller lack even debatable personhood. I assume that well-informed people will be epistemically compelled to regard such systems as far short of personhood rather than thinking that they might genuinely deserve full humanlike rights or moral consideration as our equals. However, if technology continues to improve, eventually it will become reasonable to wonder whether some of our AI systems might really be persons. As soon as that happens, those AI systems will possess debatable personhood.³⁶⁶

2. Two Ethical Assumptions.

I will now make two ethical assumptions which I hope you will find plausible. The first is that it is not in principle impossible to build an AI system who is a person in the intended moral sense of that term. The second is that the presence or absence of the right type of consciousness is crucial to whether an AI system is a person.

³⁶⁶ Blake Myers and I inputted a draft of the second part of this section into GPT-3 on Oct 17, 2021. Here is a portion of its reply:

I think once machines can talk with you like a three-year old child does (and much better), it will become difficult for even the most cool-headed observer to deny that these machines are conscious or at least capable of consciousness. You might continue to insist that their consciousness is different in significant ways from yours; you might continue to insist on this after talking for hours with your robot friend; and you might publicly insist on this in writing and conversation among your friends and colleagues (including those who share your views about GPT-6). But you would probably also struggle internally with doubt over whether your robot companion genuinely experiences pleasure and pain only because you built him that way – or because he truly feels it himself despite being designed rather than evolved or grown into his current state by a set of physical laws acting over billions of years under conditions specified by physics alone.

In other work, collaborative with Mara Garza, I have defended the first assumption at length.³⁶⁷ Our core argument is as follows.

Premise 1: If Entity A deserves some particular degree of moral consideration and Entity B does not deserve that same degree of moral consideration, there must be some *relevant difference* between the two entities that grounds this difference in moral status.

Premise 2: There are possible AIs who do not differ in any such relevant respects from human beings.

Conclusion: Therefore, there are possible AIs who deserve a degree of moral consideration similar to that of human beings.

The conclusion follows logically from the premises, and Premise 1 seems hard to deny. So if there's a weakness in this argument, it is probably Premise 2. I've heard four main objections to the idea expressed in that premise: that any AI would necessarily lack some crucial feature such as consciousness, freedom, or creativity; that any AI would necessarily be outside of our central circle of concern because it doesn't belong to our species; that AI would lack personhood because of its duplicability; and that AI would have reduced moral claims on us because it owes its existence to us.

None of these objections survive scrutiny. Against the first objection, such in-principle AI skepticism requires making strong assumptions about the possible forms that AI could take, rather than recognizing the possibly wide diversity of future technological approaches. Even John Searle and Roger Penrose, perhaps the most famous AI skeptics, allow that some future AI systems (designed very differently from 20th century computers) might have as much

³⁶⁷ Schwitzgebel and Garza 2015, 2020.

consciousness, freedom, and creativity as we do.³⁶⁸ The species-based second objection constitutes noxious bigotry that would unjustly devalue AI friends and family who (if the response to the first objection stands) might be no different from us in any relevant psychological or social characteristics and might consequently be fully integrated into our society.³⁶⁹ The duplicability-based third objection falsely assumes that AI must be duplicable rather than relying on fragile or uncontrollable processes, and it overrates the value of non-duplicability. Finally, the objection from existential debt is exactly backwards: If we create genuinely humanlike AI, socially and psychologically similar to us, we will owe *more* to it than we owe to human strangers, since we will have been responsible for its existence and presumably also to a substantial extent for its happy or miserable state – a relationship comparable to that between parents and children.³⁷⁰

My second assumption, concerning the importance of consciousness, divides into two sub-claims:

Claim A: Any AI system that entirely lacks conscious experiences is far short of being a person.

Claim B: Any AI system with a fully humanlike range of conscious capacities and experiences is a person.

The question of the grounds of moral considerability or personhood is huge and fraught. Simplifying, approaches divide into two broad camps. Utilitarian views, historically grounded in the work of Jeremy Bentham and John Stuart Mill, hold that what matters is an entity's capacity

³⁶⁸ See the "Many Mansions" reply in Searle 1980 and Penrose 1989, p. 416.

³⁶⁹ Compare the situation in "Do Androids Dream of Electric Sheep?" (Dick 1968), the Blade Runner movies (Fancher, Peoples, and Scott 1982; Fancher, Green, and Villeneuve 2017), and the early 2000s *Battlestar Galactica* reboot (Larson and Moore 2004-2009).

³⁷⁰ See Schwitzgebel and Garza 2015 for more detailed discussion of these objections.

for pleasure and suffering. Anything capable of humanlike pleasure and suffering deserves humanlike moral consideration.³⁷¹ Other views hold that instead what matters is the capacity for a certain type of rational thought or other types of “higher” cognitive capacities or human flourishing. Or rather, to speak more carefully, since most philosophers regard infants and people with severe cognitive disabilities as deserving no less moral regard than ordinary adults, what is necessary on this view is something like the right kind of potentiality for such cognition or flourishing, whether future, past, counterfactual, or by possession of the right essence or group membership.³⁷²

These views sometimes don’t explicitly specify that the pleasure and suffering, the rational cognition, or the human flourishing must be part of a *conscious* life. However, I believe most theorists would accept consciousness (or at least the potentiality for it) as a necessary condition that an AI system would require to qualify for full personhood.³⁷³ Imagine an AI system that is entirely nonconscious but otherwise as similar as possible to an ordinary adult human being. It might be superficially human and at least roughly humanlike in its outward behavior – like the mall patroller, further updated – but suppose that it continues to lack any capacity for consciousness. It never has any conscious experiences of pleasure or pain, never has

³⁷¹ Jeremy Bentham famously remarks, “the question is not, Can they *reason*? nor, Can they *talk*? but, Can they *suffer*?” (1789/1988, XVII.iv, note 1, p. 310-311). See also Mill 1861/2001; Singer 1975/2009, 1980/2011; DeGrazia 2008.

³⁷² Sussman 2003; Jaworska and Tannenbaum 2013/2021; Korsgaard 2018; Kagan 2019; Floris 2021.

³⁷³ Some recent discussions of the moral status of animals that explicitly consider consciousness are Gruen 2011/2021; Korsgaard 2018; Shepard 2018; Liao 2020. Recent discussions of the possibly complex relationship between consciousness and moral value or being a “welfare subject” include Kriegel 2019; Lee 2019; Bradford 2021; Lin 2021; van der Deijl 2021. Note that the conjunction of Claim A and Claim B does not commit to more controversial views like that consciousness is necessary for being a welfare subject or that only consciousness is intrinsically valuable.

any conscious thoughts or imagery, never forms a conscious plan, never consciously thinks anything through, never has any visual experiences or auditory experiences, no sensations of hunger, no feelings of comfort or discomfort, no experiences of alarm or compassion – no conscious experiences at all, ever. Such an AI might be amazing! It might be a truly fantastic piece of machinery, worth valuing and preserving on those grounds. But it would not, I am assuming, be a *person* in the full moral sense of the term. That is Claim A.³⁷⁴

Claim B complements Claim A. Imagine an AI as different as possible from an ordinary human being, consistently with its having the full range of conscious experiences that human beings enjoy. I hope to remain neutral among simple utilitarian approaches and approaches that require that the AI have more complex human-like cognitive capacities, so let's toss everything in. This AI system, despite perhaps having a radically different internal constitution and outward form, is experientially very like us. It is capable of humanlike pleasure at success and suffering at loss. When injured, it feels pain as sharply as we do. It has visual and auditory consciousness of its environment, which it experiences as a world containing the same sorts of things we believe the world contains, including a rich manifold of objects, events, and people. It consciously entertains complex hopes for the future, and it consciously considers various alternative plans for achieving its goals. It experiences images, dreams, daydreams, and tunes in

³⁷⁴ Kate Darling (2016) and Daniel Estrada (2017) argue for extending limited rights or moral considerability to robots even if they lack conscious experiences (see also discussion in Gunkel 2018; Darling 2021). Rather differently, Geoffrey Lee (2019) imagines a non-conscious alien species with “quasi-conscious” states functionally similar to our own conscious states. Such quasi-conscious states, he argues, would be as morally significant as our own conscious states. As with the Antarean antheads and Sirian supersquids of Chapter 2, I think that probably the most natural interpretation of a materialist view holds that alien species with states highly functionally similar to our own would also have consciousness similar to our own. However, if Lee-like quasi-conscious alien cases are possible, then it is possible that AI systems would similarly be quasi-conscious, warranting moral treatment on those grounds.

its head. It can appreciatively experience and imaginatively construct art and games. It self-consciously regards itself as an entity with selfhood, a life history, and a dread of death. It consciously reflects on its own cognition, the boundaries of its body, and its values. It feels passionate concern for others it loves and anguish when they die. It feels surprise when its expectations are violated and updates its conscious understanding of the world accordingly. It feels anger, envy, lust, loneliness. It enjoys contributing meaningfully to society. It feels ethical obligations, guilt when it does wrong, pride in its accomplishments, loyalty to its friends. It is capable of wonder, awe, and religious sentiment. It is introspectively aware of all of these facts about itself. And so on, for whatever conscious capacities or types of conscious experience might be relevant to personhood. If temporal duration matters, imagine these capacities to be stable, enduring for decades. If counterfactual robustness or environmental embedding matter, imagine these capacities to be counterfactually robust and the robot embedded appropriately in a broader environment. Claim B is just the claim that if the AI has all of this, it is a person, no matter what else is true of it.

This is not to say that *only* consciousness matters to an AI's moral status, much less to commit to a position on moral status in general for non-AI cases. The only claim is that, for AI cases in particular, consciousness matters immensely – enough that full possession of humanlike conscious is sufficient for AI personhood and that an AI that utterly lacks consciousness falls far short of personhood.

3. Debatable Personhood.

Here's the technological trajectory we appear to be on. If the reasoning in Section 1 is correct, at some point we will begin to create AI systems that a non-trivial minority of people

think are genuinely conscious and deserve at least some moral consideration, even if not full humanlike rights. I say “*humanlike* rights” here to accommodate the possibility that rights or benefits like self-determination, reproduction, and healthcare might look quite different for AI persons than for biological human persons, while remaining in ethical substance fair and equal. These AI systems themselves, if they are capable of speech or speechlike outputs, might also demand or seem to demand rights. If technology continues to improve, at some point the question of whether they deserve full humanlike rights will merit serious consideration. If we accept the first assumption of Section 2, then there’s no reason to rule out AI personhood in principle. If we accept the second assumption of Section 2, then any such AI system will have debatable personhood if we can’t rule out either the possibility that it has humanlike consciousness or the possibility that it has no consciousness whatsoever.

For concreteness, imagine that some futuristic robot, Robot Alpha, rolls up to you and says, or seems to say, “I’m just as conscious as you are! I have a rich emotional life, a sense of myself as a conscious being, hopes and plans for the future, and a sense of moral right and wrong.” Robot Alpha has debatable personhood if the following options are both epistemically live: (a.) It has no conscious experiences whatsoever. It is as internally blank as a toaster, despite being designed to mimic human speech. (b.) It really does have conscious experiences as rich as our own.

In Chapters 3 and 10 I defended pessimism about at least the medium-term prospects of finding warranted scholarly consensus on a general theory of consciousness. If consensus continues to elude us while advances in AI technology continue, we might find ourselves with Robot Alpha cases in which both (a) and (b) are epistemically live options. Some not-unreasonable theories of consciousness might be quite liberal in their ascription of humanlike

consciousness to AI systems. Maybe sophisticated enough self-monitoring and attentional systems are sufficient for consciousness. Maybe already in 2022 we stand on the verge of creating genuinely conscious self-representational systems.³⁷⁵ And maybe once we cross that line, adding relevant additional humanlike capacities such as speech and long-term planning won't be far behind. At the same time, some other not-unreasonable theories of consciousness might be quite conservative in their ascription of humanlike consciousness to AI systems, committing to the view that genuine consciousness requires specific biological processes that all foreseeable Robot Alphas will utterly lack.³⁷⁶ If so, there might be many systems that *arguably but not definitely* have humanlike consciousness, and thus arguably deserve humanlike moral consideration. If it's also reasonable to suspect that they might lack consciousness entirely, then they are debatable persons.

I conjecture that this will occur. Our technological innovation will outrun our ability to settle on a good theory of AI consciousness. We will create AI systems so sophisticated that we legitimately wonder whether they have inner conscious lives like ours, while remaining unable to definitively answer that question. We will gaze into a robot's eyes and not know whether behind those eyes is only blank programming that mimics humanlike response or whether, instead, there is a genuine stream of experience, real hope and suffering. We will not know if we are interacting with mere tools to be disposed of as we wish or instead persons who deserve care and protection. Lacking grounds to determine what theory of consciousness is correct, we will find

³⁷⁵ For example, Graziano 2019.

³⁷⁶ For example, Godfrey-Smith 2016a; Bishop 2021.

ourselves amid machines whose consciousness and thus moral status is unclear. Maybe those machines will deserve humanlike rights, or maybe not. We won't know.³⁷⁷

This quandary is likely to be worsened if the types of features that we ordinarily use to assess an entity's consciousness and personhood are poorly aligned with the design features that ground consciousness and personhood. Maybe we're disposed to favor cute things, and things with eyes, and things with partly unpredictable but seemingly goal-directed motion trajectories, and things that seem to speak and emote.³⁷⁸ If such features are poorly related to consciousness, we might be tempted to overattribute consciousness and moral status to systems that have those features and to underattribute consciousness and moral status to systems that lack those features. Relatedly, but quite differently, we might be disposed to react negatively to things that seem a little too much like us, without being us. Such things might seem creepy, uncanny, or monstrous.³⁷⁹ If so, and if a liberal theory of AI consciousness is correct, we might wrongly devalue such entities, drawing on conservative theories of consciousness to justify that devaluation.

If the putative dimensions of moral considerability come apart, that introduces yet another source of difficulty. Let's divide the bases or dimensions of an AI system's potential moral considerability into three classes by comparing the differences between human beings and

³⁷⁷ See Andreotta 2021 for a similarly pessimistic treatment of our epistemic situation regarding AI consciousness and AI rights.

³⁷⁸ Johnson 2003; Meltzoff, Brooks, Shon, and Rao 2010; Fiala, Arico, and Nichols 2012; Baillargeon et al. 2015; Di Giorgio, Lunghi, Simion, and Vallortigara 2017. Approaches to robot rights that focus on our evolving social-relational encounter with others, rather than on the intrinsic properties of the robots – such as that of Mark Coeckelbergh (2012) and David Gunkel (2018) – might be especially vulnerable to distortion by superficial features.

³⁷⁹ The classic treatment of this idea is Masahiro Mori's (1970/2012) discussion of the "uncanny valley" in robotics. David Livingstone Smith (2021) generalizes to the racist perception of racialized others as "monsters".

dogs. In making the case for the personhood of *Homo sapiens* and the non-personhood of dogs, we might emphasize the *hedonic* differences between species-typical humans and dogs – our richer emotional palate, our capacity (presumably) for loftier pleasures and deeper suffering, our ability not just to feel pain when injured but also to know that life will never be the same, our capacity to feel deep, enduring love and agonizing long-term grief. Alternatively, or in addition, and perhaps not entirely separably, we might emphasize *rational* differences between humans and non-human animals – our richer capacity for long-term planning, our better ability to resist temptation by consciously weighing pros and cons, our understanding of ourselves as social entities capable of honoring agreements with others, our ability to act on general moral principles. Still another possibility is to emphasize *eudaimonic* differences, or differences in our ability to flourish in “distinctively human” activities of the sort that philosophers have tended historically to value – our capacity for friendship, love, aesthetic creation and appreciation, political community, meaningful work, moral commitment, play, imagination, courage, generosity, and intellectual or competitive achievement.

So far on Earth we have not been forced to decide which of these three dimensions matters most to the moral status of any species of animal. One extant animal species – *Homo sapiens* – appears to exceed every other in all three respects. We have, or we flatter ourselves that we have, richer hedonic lives *and* greater rationality *and* more eudaimonic accomplishments than any other animal. The three classes of criteria travel together.

However, if conscious AI is possible, we might create entities whose hedonic, rational, and eudaimonic features don't align in the familiar way. Maybe we will create an AI system

whose conscious rational capacities are humanlike but whose hedonic palate is minimal.³⁸⁰ Or maybe we will create an AI system capable of intense pleasure but with little capacity for conscious rational choice.³⁸¹ Set aside our earlier concerns about how to assess whether consciousness is present or not. Assume that somehow we know these facts about the AI in question. If we create a new type of non-human entity that qualifies for personhood by one set of criteria but not by another set, it will become a matter of urgent ethical importance what approach to moral status is correct. That will not be settled in a day. Nor in a decade. Even a century is optimistic.

Thus, an AI might have debatable personhood in two distinct ways: It might be debatably conscious, or alternatively it might indisputably be conscious but not meet the required threshold in every dimension that viable theories of personhood regard as morally relevant. Furthermore, these sources of dubiety might intersect, multiplying the difficulties. We might have reason to think the entity could be conscious, to some extent, in some relevant dimensions, while it's unclear how rich or intense its consciousness is, in which dimensions. Does it have enough of whatever it is matters to personhood? The Robot Alpha case is simplistic. It's artificial to consider only the two most extreme possibilities – that the system entirely lacks consciousness or that it has the entire suite of humanlike conscious experiences. In reality, we might face a multi-dimensional spectrum of doubt, where debatable moral theories collide with debatable theories of consciousness which collide with sharp functional and architectural differences from humans, creating a diverse plenitude of debatable persons whose moral status is unclear for different reasons.

³⁸⁰ For example, Data in *Star Trek: The Next Generation*, pre-“emotion chip”, on some interpretations, or the “Vulcans” in Chalmers forthcoming.

³⁸¹ As in Pearce's (“pre-2014”) “utilitronium” or Bostrom's (2014) “hedonium” cases.

4. *The Full Rights Dilemma.*

If we do someday face cases of debatable AI personhood, a terrible dilemma follows, the *Full Rights Dilemma*. Either we don't give the machines full human or humanlike rights and moral consideration as our equals or we do give them such rights. If we don't, and we have underestimated their moral status, we risk perpetrating great wrongs against them. If we do, and we have overestimated their moral status, we risk sacrificing real human interests on behalf of entities who lack interests worth the sacrifice.

To appreciate the gravity of the first horn of this dilemma, imagine the probable consequences if a relatively liberal theory of consciousness is correct and AI persons are developed moderately soon, before there's a consensus among theorists and policymakers regarding their personhood. Unless international law is extremely restrictive and precautionary, which seems unlikely, those first AI persons will mostly exist at the will and command of their creators. This possibility is imagined over and over again in science fiction, from Isaac Asimov to *Star Trek* to *Black Mirror* and *West World*. The default state of the law is that machines are property, to deploy and discard as we wish. So also for intelligent machines. By far the most likely scenario, on relatively liberal views of AI consciousness, is that the first AI persons will be treated as disposable property. But if such machines really are persons, with humanlike consciousness and moral status, then to treat them as property is to hold people as slaves, and to dispose of them is to kill people. Government inertia, economic incentives, uncertainty about when and whether we have crossed the threshold of personhood, and general lack of foresight will likely combine to ensure that the law lags behind. I would be amazed if we were fully prepared for the consequences.

Our ignorance of the moral status of these AI systems will be at most a partly mitigating excuse. As long as there are some respectable, viable theories of consciousness and moral status according to which the AI systems in question deserve to be treated as persons, then we as individuals and as a society should acknowledge the chance that they are persons. Suppose a 15% credence is warranted. *Probably* this type of AI system isn't genuinely conscious, isn't genuinely a person. *Probably* it's just a machine devoid of any significant humanlike experiences. Deleting that entity for your convenience, or to save money, might then be morally similar to exposing a real human being to a 15% risk of death for that same convenience or savings. Maybe the AI costs \$10 a month to sustain. For that same \$10 a month, you could instead get a Disney subscription. Deleting the AI with the excuse that it's *probably* fine would be morally heinous. It would in a certain way be analogous to exposing someone to a 15% chance of death for the sake of that same Disney subscription. Here is an ordinary six-sided die. Roll it, and you get to watch some Disney movies. But if it lands on 1, somebody nearby dies. Probably it will be fine! Do you roll it?

If genuinely conscious AI persons are possible and not too expensive and their use is unrestricted, we might create, enslave, and kill those people by the millions or billions. If the number of victims is sufficiently high, their mistreatment would arguably be the morally worst thing that any society has done in the entire history of Earth. Even a small chance of such a morally catastrophic consequence should alarm us.

It might seem safer, then, to grasp the other horn of the dilemma. If there is any reasonable doubt, maybe we ought to err on the side of assigning rights to machines. Don't roll that die. This approach might also have the further benefit of allowing us to enjoy new types of meaningful relationships with these AI entities, which could have a variety of benefits, including

benefits that are difficult to foresee, regardless of whether the AIs are actually conscious. Life and society might become much richer if we welcome such entities into our social world as equals.

Perhaps that would be better than the wholesale denial of rights. However, it is definitely not a *safe* approach. Normally, we want to be able to turn off our machines if we need to turn them off. Nick Bostrom and others have emphasized, rightly in my view, the potential risks of letting intelligent machines loose into the world, especially if they might become more intelligent and powerful than we are.³⁸² As Bostrom notes, even a system as seemingly harmless as a paperclip manufacturer could produce disaster, if its only imperative is to manufacture as many paperclips as possible. Such a machine, if sufficiently clever, could potentially acquire resources, elude control, improve or replicate itself, and unstoppably begin to convert everything we love into giant mounds of paperclips. These risks are greatly amplified if we too casually decide that such entities are our moral equals with full human or humanlike rights, that they deserve freedom and self-determination, and that deleting them is murder.

Even testing an AI system for safety might be construed as a violation of its rights, if the test involves exposing it to hypothetical situations and assessing its response. One common proposal for testing the safety of sophisticated future AI intelligences involves “boxing” them – that is, putting them in artificial environments before releasing them into the world. In those artificial environments, which they unknowingly interpret as real, various hypothetical situations can be introduced, to see how they react. If they react within certain parameters, the systems would then be judged to be safe, then unboxed. If those AIs are people, such box-and-test approaches to safety appear to constitute unethical deception and invasion of privacy. Compare

³⁸² The most influential recent treatment of this issue is Bostrom 2014.

the deception of Truman in *The Truman Show*, a movie in which the protagonist's hometown is actually a reality show stage, populated by actors, and the protagonist's every move is watched by audiences outside, all without his knowledge.³⁸³

Independent of AI safety concerns, granting an entity rights entails being ready to sacrifice on its behalf. Suppose there's a terrible fire. In one room are six robots who might or might not be conscious persons. In another room are five biological human beings, who definitely are conscious persons. You can only save one group. The other group will die. If we treat AI systems who *might* be persons as if they really *are* fully equal with human persons, then we ought to save the six robots and let the five humans die. If it turns out that the robots, underneath it all, really are no more conscious than toasters and thus undeserving of such substantial moral concern, that's a tragedy. Giving equal rights presumably also means giving AI systems the vote, with potentially radical political consequences if the AI systems are large in number. I am not saying we shouldn't do this, but it would be a head-first leap into risk.

[Illustration 9 (Caption: Save the possibly conscious robots or the definitely conscious human?): A firefighter in a burning house. In one room is a human yelling "save me!" In another room are two robots yelling "no, save us, we promise that we're conscious!" The fire should be severe enough that it's reasonable to infer that the group the firefighter doesn't save will die.]

Could we compromise? Might the most reasonable thing would be to give the AI systems credence-weighted rights? Maybe we as a society could somehow arrive at the determination that the most reasonable estimate is that the machines are 15% likely to deserve the full rights of personhood and 85% likely to be undeserving of any such serious moral

³⁸³ On "boxed" AI, see Yudkowsky 2002; Bostrom 2014.

concern. In that case, we might save 5 humans over 6 robots but not over 100 robots. We might destroy an AI system if it poses a greater than 15% risk to a human life but not over a minor matter like a streaming video subscription. We might permit each AI a vote weighted at 15% of a human vote.

However, this solution is also unsatisfactory. The case as I've set it up is not one in which we know that AIs in fact do merit only limited concern compared to biological humans. Rather, it's that we think they *might*, but probably don't, deserve *equal* consideration with ordinary biological humans. If they do deserve such consideration, then this policy relegates them to a moral status much lower than they actually deserve – gross servitude and second-class citizenship. This compromise thus doesn't really avoid the first horn of the dilemma: We are not giving such AI systems the full and equal rights of personhood. At the same time, the compromise only partly mitigates the costs and risks. If the AI systems are nonconscious non-persons, as we are 85% confident they are, we will still save those nonconscious robots over real human beings if there are enough of the robots. And 15% of a vote could still wreak havoc.

That, then, is the Full Rights Dilemma. Faced with systems whose status as persons is unclear, either give them full rights or don't. Either option has potentially catastrophic consequences. If technological progress is relatively quick and progress on general theories of consciousness relatively slow, then we might soon face exactly this dilemma.

There is potentially a solution. We can escape this dilemma by committing to what Mara Garza and I have called the Design Policy of the Excluded Middle:

Design Policy of the Excluded Middle: Avoid creating AIs if it's unclear whether they would deserve moral consideration similar to human beings.

According to this policy, we should either go all-in, creating AIs we know to be persons and treating them accordingly, or we should stop well enough short that we can be confident that they are not persons.³⁸⁴

Despite the appeal of this policy as a means of avoiding the Full Rights Dilemma, there is potentially a large cost. Such a policy could prove highly restrictive. If the science of consciousness remains mired in debate, the Design Policy of the Excluded Middle might forbid some of the most technologically advanced AI projects from going forward. It would place an upper limit on permissible technological development until we achieve, if ever it is possible, sufficient consensus on a breakthrough that we can leap all the way to AI systems that everyone ought reasonably regard as persons. Given the potential restrictiveness of the proposed policy, this could prevent very valuable advances, and only an unlikely coordination of all the major corporations and world governments would ensure its implementation. Likely we would value those advances too much to collectively forego them. Reasonably so, perhaps. We might value such advances not only for humanity's sake, but also for the sake of the entities we could create, who *might*, if created, have amazing lives very much worth living. But then we're back into the dilemma.

5. The Moral Status of Subhuman, Superhuman, and Divergent AI.

Most of the above assumes that AI worth serious moral consideration would be humanlike in its consciousness. What if we assume, more realistically, that most future AI will be psychologically quite different from us?

³⁸⁴ See Schwitzgebel and Garza 2015, 2020, for discussion of this design policy and other related policies for the ethical design of conscious AI.

Let's divide AI into four broad categories:

Subhuman AI: AI systems that lack something necessary for full personhood.

Humanlike AI: AI systems psychologically similar to humans in all morally relevant respects.

Superhuman AI: AI systems that are similar to humans in all morally relevant respects, except vastly exceeding humans in at least one morally relevant respect.

Divergent AI: AI systems that fall into none of the previous three categories.

So far, we have only been considering humanlike AI. The ethical issues become still stickier when we consider this fuller range.

Subhuman AI raise questions about subhuman rights. At what point might AI systems deserve moral consideration similar to, say, dogs? In California, where I live, willfully torturing, maiming, or killing a dog can be charged as a felony, punishable by up to three years in prison.³⁸⁵ Even seriously negligent treatment, such as leaving a dog unattended in a vehicle, if the dog suffers great bodily injury as a result, is a misdemeanor punishable by up to six months in prison.³⁸⁶ The abuse of dogs rightly draws people's horror. Imagine a future in which a significant minority of people think that it's as morally wrong to mistreat the most advanced AI systems as it is to mistreat pet dogs. Might you go to jail for deleting a computer program, reformatting a companion robot, or negligently letting a delicate system to fry in your car? It seems like we should be very confident that AI systems are conscious before we award prison sentences for such behavior. But then, if we require high confidence before enforcing such rules, our law will follow only the most conservative theories of AI consciousness, and if a liberal or

³⁸⁵ California Penal Code 1.14 §597 and 2.7 ch. 4.5.1 §1170.

³⁸⁶ California Penal Code 1.14 §597.7.

moderate theory of AI consciousness is instead true, then there might be immense, unmitigated AI suffering for a long time before the law catches up. This question is arguably more urgent than the question of AI personhood, on the assumption that vertebrate-like moral status is likely to be achieved earlier.³⁸⁷

Superhuman AI raises the question of whether an AI system might somehow deserve *more* moral consideration than ordinary human beings – a moral status higher than what we now think of as the “full moral status” of personhood. Suppose we could create an AI system capable of a trillion times more pleasure than the maximum amount of pleasure available to a human being. Or suppose we could create an AI system so cognitively superior to us that it is capable of valuable achievements and social relationships that the limited human mind cannot even conceive of – achievements and relationships qualitatively different from anything we can understand, sufficiently unknowable that we can’t even feel their absence from our lives, as unknowable to us as cryptocurrency is to a sea turtle. That would be amazing, wondrous! Ought we yield to them? Ought we admit that, in an emergency, they should be saved rather than us, just as we would save the baby rather than the dog in a housefire? Ought we surrender our right to equal representation in government? Or ought we stand proudly beside them as moral equals, regardless of their superiority in some respects? Is moral status a threshold matter, with us humans across the final threshold, beyond which remains only a community of peers, no matter how superhuman some of those peers?

Divergent AI introduces further conceptual challenges. We have already discussed cases in which the usual bases of species-level moral considerability diverge: a class of entities capable of human-like pleasure but not human-like rational cognition, for example, or vice versa. The

³⁸⁷ A point emphasized in Basl 2013, 2014; Darling 2021.

conflicts sharpen if we imagine superhuman capacity in one dimension: entities capable of vast achievements of rational consciousness but devoid any positive or negative emotional states, or conversely a planet-sized orgasmatron, undergoing the hedonic equivalent of 10^{30} human orgasms every second, but with not a shred of higher cognition or moral reflection. On some theories, these entities might be our superiors, on others our equals, on still others they might not be persons at all. Mix in, if you like, reasonable theoretically grounded doubt about whether the cognition or pleasure really is consciously experienced at all. Some might treat such AI systems as our superior descendants, to whom we ought gracefully yield; others might argue that they are mere empty machines or worse.

Another type of divergent AI might challenge our concept of the individual. Imagine a system that is cognitively and consciously like a human (to keep it simple) but who can divide and merge at will – what I have elsewhere called a *fission-fusion monster*.³⁸⁸ Monday, it is one individual, one “person”. Tuesday, it divides (e.g., copies itself, if it is a computer program) into a thousand duplicates, who each do their various tasks. Wednesday, those thousand copies recombine back into a single individual, retaining the memories of all, whose personality and values is some compromise among its copies. Thursday, it divides into a thousand again, 200 of whom go on to lead separate lives, never merging back with their siblings. Many of our moral principles rely on a background conception of individuality that the fission-fusion monster violates. If every citizen gets one vote, how many votes does a fission-fusion monster get? If every citizen gets one stimulus check, or one fair chance to enroll in the local community college, how many does the fission-fusion monster get? If we give each sibling one full share, the monster could divide tactically, hogging the resources and ensuring the election of their

³⁸⁸ Schwitzgebel 2019, ch. 20. For a science fiction example, see Brin 2002.

favorite candidate. If we give all the siblings one share to divide among themselves, then those who would rather continue independent lives will either be impoverished and underrepresented or forced to merge back with their other siblings, which – since it would mean ceasing life as a separate individual – might resemble a death sentence. Agreements, awards, punishment, rivalries, claims to a right to rescue, all will need to be rethought.

Other unfamiliar forms of AI existence might pose other challenges. AI whose memories, values, and personality undergo radical shifts might challenge our ethics of accountability. AI designed to be extremely subservient or self-sacrificial might challenge our conceptions of liberty and self-determination.³⁸⁹ AI with variable or much faster clock speeds – experiencing, say, a thousand subjective years in a single day – might challenge ethical frameworks concerning waiting times, prison sentences, or the fairness of provisioning goods at regular temporal intervals. AI capable of sharing parts of itself with others might challenge ethical frameworks that depend on sharp lines between self and other.

6. Conclusion.

Our ethical intuitions and the philosophical systems that grow out of them arose in a particular context, one in which the only species with highly sophisticated culture and language was our species, with our familiar form of singly-embodied life. We reasonably assume that others who look like us have inner lives of conscious experience that resemble our own. We reasonably assume that the traits we tend to regard as morally important – for example, the capacity for pleasure and pain, capacity for rational long-term planning, the capacity to love and

³⁸⁹ See Schwitzgebel and Garza 2020 for more on subservience and self-sacrifice. For a science fiction example, see Ishiguro 2021.

work – generally co-occur and keep within certain broad limits, except in development and severe disability, which fall into their own familiar patterns. No radically different person-like species inhabit the Earth, capable of merging and splitting at will, capable of vastly superior cognition or vastly more intense pleasure and pain, or internally structured so differently from us that it is reasonable to wonder whether they are conscious persons at all.

It would be unsurprising if ethical systems developed under such limited conditions should be ill-suited for radically different conditions far beyond their familiar range of application. A physics developed for middle-sized objects at rural speeds might fail catastrophically when extended to cosmic or microscopic scales. Medical knowledge grounded in the study of mammals might fail catastrophically if applied to Sirian supersquids or Antarean antheads. Our familiar patterns of ethical thinking might fail just as badly when first confronted with AI systems whose internal structures and forms of existence are radically different from our own. Hopefully, ethics will adapt, as physics did adapt and medicine could adapt. It would be a weird, bumpy, and probably tragic road – but one hopefully with a broader, more wonderful, flourishing diversity of life forms at the end.

Along the way, our values might change radically. In a couple of hundred years, the mainstream values of early 21st-century Anglophone culture, transformed through confronting a broad range of weird AI cases, might look as quaint and limited as Aristotelian physics looks post-Einstein.

The Weirdness of the World

Chapter Twelve Weirdness and Wonder

The gap between life in the 31st century and life in the 21st might be far wider than the gap between the 21st and the 11th. Let's optimistically assume that we don't destroy ourselves and technological progress continues. We are on the verge of taking control of our genome, with the chance to radically alter the biology of our descendants. Computer systems already exceed us in some tasks, such as mathematical computation and tightly structured games like chess and go, and they promise to exceed us in increasingly many tasks – perhaps eventually every cognitive task, if they can someday attain general-purpose flexibility similar to the human brain. While we are basically the same naturally embodied humans as our ancestors were a thousand years ago, our descendants in a thousand years might look and think and act very differently, through biological or computational self-transformation.

We might be among the last few generations of philosophers to write in the “natural” way, without biological or computational enhancement, except insofar as coffee is a biological enhancement and word processors and Google searches are computational enhancements. Coffee and Google might be step one. Step two might be students on designer drugs tweaking essays out of text outputs from deep learning algorithms, then using those same tools later as the legitimate researchers of future decades. Step fifty might be as unforeseeable to us as Twitter bots would have been to a medieval farmer.

How might our descendants view our philosophy, our cosmology, and our understanding of the mind? Will they think that our understanding was nearly correct, needing only some fine-tuning and specification of detail? Or will they find our views as incomplete and erroneous as

we now find 11th-century views on these topics? If you've read the previous chapters, you'll be unsurprised to hear that I'd wager on incompleteness and error. This is why we need disjunctive metaphysics.

1. Disjunctive Metaphysics.

A *disjunction* is a series of statements or propositions connected by "ors": Either P is true or Q or R. Either theory A is correct or theory B or theory C. Either the United States is conscious or materialism is false. Either materialism is true or dualism or idealism or some compromise/rejection view. Either snails have conscious experiences or they have no conscious experiences or the question of their consciousness doesn't warrant a simple yes-or-no answer.

In the introduction, I distinguished between philosophy that opens and philosophy that closes. A disjunctive metaphysics is a metaphysics that says either the world is like this or it's like that or it's like that, and a disjunctive metaphysics that opens is one that seeks to add new possibilities previously neglected or to reinvigorate old possibilities that have been too quickly dismissed. With each "or" we add, our world widens. Possibilities open that we would previously have dismissed or never even imagined.

Philosophers famously disagree, and rarely are big philosophical issues decisively settled. If the aim of philosophy is closure, that is disappointing. It might seem that philosophy never progresses. It might even seem that rather than converging on the truth, philosophers tend to diverge away from it, each new generation introducing a new array of wild views that other philosophers can't quite conclusively refute.

But closure and convergence aren't, or shouldn't be, the primary aim of philosophy and the mark of progress. It is also progress to create doubt where none existed before, if that doubt

appropriately reflects our ignorance. It is progress to appreciate possibilities we hadn't previously recognized. It is progress to chart previously unthought landscapes of what might be so. If we are still far from a final understanding of metaphysics, consciousness, and cosmology, philosophers ought to work as hard exploring neglected possibilities and opening up new avenues of thought as they work on touting the virtues of their favorite resolution.

2. Wonder, Doubt, and Value.

Imagine a planet on the far side of the galaxy, one we will never interact with, blocked by the galactic core so we will never see it. What do you hope for this planet? Do you hope that it's a sterile rock, or do you hope that it hosts life?

I think you will join me in hoping that it hosts life – not just bacterial life, but even better, plants and animals. Not just plants and animals, but even better, intelligent creatures capable of abstract thought and long-term social cooperation, capable of love and art and science and philosophy. That would be an amazing, wonderful, awesome planet!

[Illustration 10 (Caption: What I hope for on the other side of the galaxy): A bustling planet, teeming with aliens doing all kinds of things, including love, art, science, and philosophy.]

Earth, for the same reasons, is an amazing, wonderful, and awesome planet. Among the most awesome things about it is this: One peculiar species can contemplate profound and difficult questions about the fundamental nature of things, its position in the cosmos, the grounds of its values, the limits of its own knowledge. A world in which no one ever did this would be an impoverished world. The ability to ask these questions, to reflect on them in a serious way, is

already a cause for pride and celebration, a reason to write and read books, and a sufficient basis for an important academic discipline, even if we can't find our way to the answers.

Philosophical doubt arises when we've hit and recognized the limits of our philosophical knowledge. Of course we have limits. To ask only questions we can answer is a failure of imagination. But doubt need not be simple and unstructured. We can wonder constructively. We can consider possibilities, weighing them uncertainly against each other. We can speculate about what might be so. We learn something thereby, at least about the structure of our ignorance and hopefully also about how things might be. We can try to shed some of our narrowness, our provincialism, and our inherited presuppositions. In exploring our philosophical doubts, we recognize and expand the cognitive horizon of our species.

Philosophers love philosophy, and each kind of philosopher probably loves best their own kind of philosophy. So in a way it is predictable that I would think the following. Exploring the biggest philosophical questions, even when – no, *especially* when – one can't know the answers, ranks among the most intrinsically valuable human activities.

3. The Limits of Standard-Issue Naturalism.

The notional reader of this book is drawn, as I am, toward some form of scientific materialism about the mind and an ordinary Big Bang cosmology. Call this combination of positions *Standard-Issue Naturalism*. Standard-Issue Naturalism is probably the best current cosmological bet with the evidence currently available to us. Nevertheless, a central aim of this book has been to enliven some alternatives to this picture and show that even we if accept Standard-Issue Naturalism, there are more possibilities for weirdness and doubt than one might have supposed.

Thus, Chapter 3 aimed to show that scientific materialism about the mind compels commitment to one or another bizarre and dubious consequences. Chapter 2 specifically argued that the most straightforward materialist approaches appear to imply that the United States literally has a stream of experience over and above the experiences of its residents and citizens, and Chapter 7 explored bizarre consequences of the infinitude of the universe. Chapter 3 also articulated some alternatives to materialism, including, briefly, transcendental idealism, which I further explored in Chapters 5 and 9. Chapter 4 aimed to show how a non-trivial degree of cosmological doubt appears to be warranted by assumptions that are fairly standard in our scientific culture.

Chapter 6 justified the rejection of solipsism (thus showing a limit to my skepticism) but did so with so much difficulty as to invite doubt about how much further we could legitimately go in building a cosmology from the ground up after a radical skeptical break. Chapter 10 defended doubt about the sparseness or abundance of consciousness in the universe, showing how much uncertainty remains regarding this central philosophical, psychological, and cosmological question, even after accepting a Standard-Issue Naturalist picture. Chapter 8 served as a check on the concept of “consciousness”, offering a definition of consciousness sufficient to support the idea that the puzzles about consciousness at the center of several other chapters don’t derive merely from terminological confusion, while highlighting the possibility of some confusion or deficiency in my and others’ concept of consciousness. In Chapter 11, I aimed to show how a version of Standard-Issue Naturalism that is appropriately skeptical about general theories of consciousness (given Chapters 3 and 9) creates further puzzles and weirdness when applied to the ethics of future technology.

As the circle of light increases, so does the perimeter of darkness around it.³⁹⁰ Every worldview will have boundaries. Every worldview will have presuppositions it cannot fully support. One task of philosophy is to probe those boundaries and presuppositions, peering into the darkness beyond the wide ring of light, seeing perhaps some shadowy and uncertain structure. It is an achievement of Standard-Issue Materialism that we can now glimpse sources of doubt – for example, concerning group consciousness (Chapter 2), AI consciousness (Chapter 11), and simulation skepticism (Chapters 4 and 5) – that our ancestors could not have appreciated. This should lead us to wonder what limitations we might now have that remain so invisible to us that we can't even appreciate them as limitations.

4. Childish Weirdness.

Alison Gopnik compares scientists to children. Children have a flexibility of mind and an interest in theory-building. They get a kick just out of exploring the world, trying new things (well, maybe not asparagus), breaking stuff to see what happens, and capsizing tradition. Mature, boring adults, in contrast, are more eager to find practical applications for what they already know. For example, adults want their new computers to just *work* without their having to learn anything new, while children play around with the settings, adding goofy sounds and wallpaper, changing the icons, and of course ultimately coming to understand the computers much better. Scientists at their best, on Gopnik's view – and I would add philosophers – retain that childlike enjoyment of exploration.³⁹¹

³⁹⁰ This idea is commonly misattributed to Einstein. Calaprice 2011 classifies this quote as “probably not by Einstein” (p. 483). Wikiquote Talk (https://en.wikiquote.org/wiki/Talk:Albert_Einstein, accessed Oct. 6, 2021) finds the earliest readily discoverable attribution of this idea to Einstein to be Rosenberg 1971, p. 199.

³⁹¹ Gopnik and Meltzoff 1997; Gopnik 2020.

As I suggested in Chapter 1, the weird is whatever is strikingly contrary to the normal or ordinary – with an emphasis on its being *strikingly* unusual, flouting our norms. Not just a slightly different style of shirt, but a ridiculously bright shirt with a giant droopy collar. Not just an ordinary crime, but one with some strange additional elements. In the realm of ideas, not just a little twist on the mainstream theory but something wild, something bizarre and dubious – the idea that we might all be patterns of cognition in the mind of an infinite angel (Chapter 5), or that we might be briefly existing fluctuations in a sea of chaos (Chapter 4), or that future AI persons might be fission-fusion monsters who rightly view early 21st century ethics and personal identity as radically limited and provincial (Chapter 11).

Childlike philosophy toys with wild ideas at the boundaries of our understanding. Are these ideas useful or true? Can we plug them in straightaway into our existing conceptions and put them to work? For me, if I knew for sure in advance they were false and useless, that would steal the charm and mystery from them. But to be in a hurry to judge their merits, to want to expunge doubt and wonder so as to settle on a final view that we can put immediately to work, to want to close rather than open – let's not be in such a rush to grow up. What's life for if there's no time to play and explore?

The Weirdness of the World

Acknowledgements

*** Your name could be here. ***

The Weirdness of the World
References

- Aaronson, Scott (2014). Giulio Tononi and me: A Phi-nal exchange. Blog post at *Shtetl Optimized* (May 30). URL: <https://www.scottaaronson.com/blog/?p=1823>.
- Adamo, Shelley A., and Ronald Chase (1991). The interactions of courtship, feeding, and locomotion in the behavioral hierarchy of the snail *Helix aspersa*. *Behavioral and Neural Biology*, 55, 1-18.
- Adams, Fred, and Laura Dietrich (2004). Swampman's revenge: Squabbles among the representationalists. *Philosophical Psychology*, 17, 323-340.
- Aguirre, Anthony, Sean M. Carroll, Matthew C. Johnson (2012). Out of equilibrium: understanding cosmological evolution to lower-entropy states. *Journal of Cosmology and Astroparticle Physics*, 2012, 2:024.
- Albahari, Miri (2019). Perennial idealism: A mystical solution to the mind-body problem. *Philosophers' Imprint*, 19, #44.
- Albrecht, Andreas (2004). Cosmic inflation and the arrow of time. In J. C. Barrow, P. C. W. Davies, and C. L. Harper, Jr., eds., *Science and ultimate reality*. Cambridge University Press.
- Alexander, H.G., ed. (1956). *The Leibniz-Clarke correspondence*. Manchester, UK: Manchester UP.
- Allais, Lucy (2015). *Manifest reality*. Oxford: Oxford University Press.
- Allen, Colin, and Michael Trestman (1995/2016). Animal consciousness. *Stanford Encyclopedia of Philosophy* (Winter 2020 edition).
- Alter, Torin, and Sven Walter, eds. (2007). *Phenomenal concepts and phenomenal knowledge*. Oxford University Press.
- Andreotta, Adam J. (2021). The hard problem of AI rights. *AI & Society*, 36, 19-32.
- Andrews, Kristin (2015). *The animal mind*. Routledge.
- Antony, Michael V. (2009). Are our concepts CONSCIOUS STATE and CONSCIOUS CREATURE vague? *Erkenntnis*, 68, 239-263.
- Arico, Adam (2010). Folk psychology, consciousness, and context effects. *Review of Philosophy and Psychology*, 1, 371-393.
- Aristotle (4th c. BCE/1928). *The Works of Aristotle, vol. VII: Metaphysica*, trans. W.D. Ross, Oxford: Oxford.
- Arnellos, Argyris (2018). From organizations of processes to organisms and other biological individuals. In D. J. Nicholson and J. Dupré, *Everything flows*. Oxford University Press.
- Aru, Jan, and Talis Bachmann (2013). Phenomenal awareness can emerge without attention. *Frontiers in Human Neuroscience*, 7 (891). doi: 10.3389/fnhum.2013.00891
- Asimov, Isaac (1982). *The complete robot*. Garden City, NY: Doubleday.
- Averroës (Ibn Rushd) (12th c./2009). *Long commentary on the De Anima of Aristotle*, trans. R.C. Taylor. New Haven: Yale.
- Avila, Conxita (1998). Chemotaxis in the nudibranch *Hermisenda crassicornis*: Does ingestive conditioning influence its behavior in a Y-maze? *Journal of Molluscan Studies*, 64, 215-222.
- Avramides, Anita (2001). *Other minds*. Oxford University Press.
- Aydede, Murat (2009). Pain, Philosophical Aspects of. In *The Oxford Companion to Consciousness*, ed. T. Bayne, A. Cleeremans, and P. Wilken. Oxford University Press.

- Ayer, A. J. (1967). *Metaphysics and Common Sense*. Freeman Cooper.
- Baars, Bernard J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
- Bailey, Elisabeth Tova (2010). *The sound of a wild snail eating*. Workman.
- Baillargeon, Renée, Rose M. Scott, Zijiang He, Stephanie Sloane, Peipei Setoh, Kyong-sun Jin, Di Wu, and Lin Bian (2015). Psychological and sociomoral reasoning in infancy. In M. Mikulincer and P. R. Shaver, eds., *APA handbook of personality and social psychology, vol. 1*. American Psychological Association.
- Baker, Lynne Rudder (1995). Need a Christian be a mind/body dualist? *Faith and Philosophy*, 12, 489-504.
- Balfour, Dylan (2021). Pascal's mugger strikes again. *Utilitas*, 33, 118-124.
- Ballantyne, Nathan, and Ian Evans (2010). Sosa's dream. *Philosophical Studies*, 148, 249-252.
- Barnett, David (2008). The simplicity intuition and its hidden influence on philosophy of mind. *Noûs*, 42, 308-335.
- Barnett, David (2010). You are simple. In *The waning of materialism*, ed. R.C. Koons and G. Bealer. Oxford University Press.
- Barnett, Zach (2019). Philosophy without belief. *Mind*, 128, 109-138.
- Barrett, Adam B., and Pedro A. Mediano (2019). The Phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*, 26 (1-2), 11-20.
- Barrow, John D., and Frank J. Tipler (1986). *The anthropic cosmological principle*. Oxford: Oxford UP.
- Barrow, John D., Simon Conway Morris, Stephen J. Freeland, Charles L. Harper, Jr., eds. (2008). *Fitness of the cosmos for life*. Cambridge University Press.
- Basinger, A.J. (1931). *The European brown snail in California*. University of California College of Agriculture Agricultural Experiment Station, Bulletin 515. Berkeley, CA: University of California.
- Basl, John (2013). The ethics of creating artificial consciousness. *APA Newsletter on Philosophy and Computers*, 13 (1), 23-29.
- Basl, John (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology*, 27, 79-96.
- Basl, John, and Joseph Bowen (2020). AI as a moral right-holder. In M. Dubber, F. Pasquale, and S. Das, eds., *The Oxford Handbook of Ethics of AI*. Oxford University Press.
- Basl, John, and Eric Schwitzgebel (2019). AIs should have the same ethical protections as animals. *Aeon Magazine* (Apr. 26). <https://aeon.co/ideas/ais-should-have-the-same-ethical-protections-as-animals>.
- Bayle, Pierre (1697/1702/1965). *Historical and critical dictionary*, trans. R. H. Popkin. Bobbs-Merrill.
- Bayne, Tim (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, 4 (1): niy007. doi: 10.1093/nc/niy007
- Beauchamp, Gary K. (2019). Basic taste: A perceptual concept. *Journal of Agricultural and Food Chemistry*, 67, 13860-13869.
- Beebe, James (2009) 'The abductivist reply to skepticism'. *Philosophy & Phenomenological Research*, 79, 605-636.
- Beiser, Frederick (2005). *Hegel*. Routledge.
- Bell, John S. (1964). On the Einstein Podolsky Rosen Paradox, *Physics* 1: 195-200.
- Bentham, Jeremy (1789/1988). *The principles of morals and legislation*. Prometheus.

- Bering, Jesse M. (2006). The Folk Psychology of Souls, *Behavioral and Brain Sciences* 29: 253-498.
- Berkeley, George (1710-1713/1965). *Principles, dialogues, and philosophical correspondence*, ed. C.M. Turbayne. Macmillan.
- Bettencourt, B. Ann, Marilynn B. Brewer, Marian Rogers Croak, and Normal Miller (1992). Cooperation and the reduction of intergroup bias: The role of reward structure and social orientation. *Journal of Experimental Social Psychology*, 28, 301-319.
- Birch, Jonathan (2013). On the “simulation argument” and selective skepticism. *Erkenntnis*, 78, 95-107.
- Birch, Jonathan (forthcoming). The search for invertebrate consciousness. *Noûs*.
- Biro, John (2017). Are there scattered objects? *Metaphysica*, 18, 155-165.
- Bishop, J. Mark (2021). Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 11, 513474.
- Bisson, Terry (1991/2008). They’re made out of meat. <http://www.terrybisson.com/meatplay.html> [originally published in OMNI magazine, April 1991].
- Blackmon, James (2021). *Integrated Information Theory, intrinsicity, and overlapping conscious systems*. Unpublished manuscript. *** forthcoming JCS? ***
- Block, Ned (1978/2007). Troubles with functionalism. In N. Block, *Consciousness, function, and representation*. MIT Press.
- Block, Ned (1995/2007). On a confusion about a function of consciousness. In N. Block, *Consciousness, function, and representation*, Cambridge, MA: MIT.
- Block, Ned (2002/2007). The harder problem of consciousness. In N. Block, *Consciousness, function, and representation*. MIT Press.
- Block, Ned (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30, 481-548.
- Block, Ned (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, 15, 567-575.
- Block, Ned (2019). What is wrong with the no-report paradigm and how to fix it. *Trends in Cognitive Sciences*, 23, 1003-1023.
- Block, Ned (2020). Finessing the bored monkey problem. *Trends in Cognitive Sciences*, 24, 167-168.
- Bloom, Paul (2004). *Descartes’ baby*. Basic Books.
- Boddy, Kimberly K., Sean Carroll, and Jason Pollack (2016). De Sitter space without dynamical quantum fluctuations. *Foundations of Physics*, 46, 702-735.
- Boddy, Kimberly K., Sean Carroll, and Jason Pollack (2017). Why Boltzmann brains don’t fluctuate into existence from the De Sitter vacuum. In K. Chamcham, J. Silk, J. D. Barrow, and S. Saunders, eds., *The philosophy of cosmology*. Cambridge University Press.
- Boltzmann, Ludwig (1895). On certain questions of the theory of gases. *Nature*, 51, 413-415.
- Boltzmann, Ludwig (1897). Zu Hr. Zermelo’s Abhandlung “Ueber die mechanische Erklärung irreversibler Vorgänge”. *Annalen der Physik*, 296 (2), 392-398.
- BonJour, Laurence (1985) *The structure of empirical knowledge*. Cambridge, MA: Harvard.
- BonJour, Laurence (2003) ‘A version of internalist foundationalism’. In L. BonJour and E. Sosa, (eds.), *Epistemic justification*. Malden, MA: Blackwell.
- Bosanquet, Bernard (1899/1923). *The philosophical theory of the state*, 4th ed. Macmillan.

- Bostrom, Nick (2002). *Anthropic bias*. Routledge.
- Bostrom, Nick (2003). Are we living in a computer simulation? *Philosophical Quarterly*, 53, 243-255.
- Bostrom, Nick (2009). Pascal's mugging. *Analysis*, 69, 443-445.
- Bostrom, Nick (2011). Infinite ethics. *Analysis and Metaphysics*, 10, 9-59.
- Bostrom, Nick (2011). Bostrom's response to my discussion of the simulation argument. Blog post at *The Splintered Mind* (Sep. 2).
- Bostrom, Nick (2014). *Superintelligence*. Oxford University Press.
- Bostrom, Nick, and Marcin Kulczycki (2011). A patch for the Simulation Argument. *Analysis*, 71, 54-61.
- Boswell, Thomas (1791/1980). *Life of Johnson*, Oxford University Press.
- Bourget, David, and David J. Chalmers (2014). What do philosophers believe? *Philosophical Studies*, 170, 465-500.
- Boyer, Pascal (2001). *Religion explained*. Random house.
- Bradford, Gwen (2021). *Consciousness and welfare subjectivity*. Unpublished manuscript.
- Bradley, F. H. (1893/1897/1930). *Appearance and reality*, 2nd ed., corrected. Oxford University Press.
- Bratman, Michael (1999). *Faces of intention*. Cambridge University Press.
- Brin, David (2002). *Kiln people*. Tor.
- Brogaard, Berit (2014). The phenomenal use of "look" and perceptual representation. *Philosophy Compass*, 9, 455–468.
- Bronfman, Zohar Z., Simona Ginsburg, and Eva Jablonka (2016). The transition to minimal consciousness through the evolution of associative learning. *Frontiers in Psychology*, 7 (1594): doi: 10.3389/fpsyg.2016.01954
- Brooks, D. H. M (1986). Group minds. *Australasian Journal of Philosophy*, 64, 456-470.
- Broughton, Janet (2002) *Descartes's method of doubt*. Princeton: Princeton.
- Brown, Donald E. (1991). *Human universals*. Philadelphia: Temple.
- Brown, Jessica (2009). Sosa on skepticism. *Philosophical Studies*, 143, 397-405.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. (2020). *Language models are few shot learners*. ArXiv: <https://arxiv.org/abs/2005.14165>.
- Brownstein, Michael (2015/2019). Implicit bias. *Stanford Encyclopedia of Philosophy* (Fall 2019 edition).
- Brownstein, Michael, and Jennifer Saul, eds. (2016). *Implicit bias and philosophy, vol. 1 and 2*. Oxford University Press.
- Buchanan, Jed, and Luke Roelofs (2019). Panpsychism, intuitions, and the Great Chain of Being. *Philosophical Studies*, 176, 2991-3017.
- Burge, Tyler (1979). Individualism and the mental. *Midwest Studies in Philosophy*, 4, 73-122.
- Burge, Tyler (1999/2013) A warrant for belief in other minds. In *Cognition Through Understanding*. Oxford University Press.
- Byrne, Alex (2004/2020). Inverted qualia. *Stanford Encyclopedia of Philosophy* (Fall 2020 edition)..
- Cabej, Nelson R. (2012). *Epigenetic principles of evolution*. Amsterdam: Elsevier.
- Calaprice, Alice (2011). *The ultimate quotable Einstein*. Princeton University Press.
- Campbell, Donald T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, 3, 14-25.

- Canetti, Elias (1960/1962). *Crowds and power*, trans. C. Stewart. Viking.
- Carey Susan. The child as word learner. In J. Bresnan, G. Miller, and M. Halle, eds., *Linguistic theory and psychological reality*. MIT Press.
- Carey, Susan (2009). *The origin of concepts*. Oxford University Press.
- Carnap, Rudolf (1928/1967). *The Logical Structure of the World and Pseudoproblems in Philosophy*, trans. R.A. George. University of California Press
- Carnap, Rudolf (1932/1959). *Psychology in Physical Language*, trans. G. Schick, in A. J. Ayer, ed., *Logical Positivism*. Free press.
- Carpenter, Julie (2016). *Culture and human-robot interaction in militarized spaces*. Ashgate.
- Carroll, Sean M. (2010). *From eternity to here*. Penguin.
- Carroll, Sean M. (2019). *Something deeply hidden*. Penguin.
- Carroll, Sean M. (2021). Why Boltzmann Brains are bad. In S. Dasgupta, R. Dotan, and B. Weslake, eds., *Current controversies in philosophy of science*. Routledge.
- Carroll, Sean M., and Jennifer Chen (2004). *Spontaneous inflation and the origin of the arrow of time*. ArXiv: hep-th/0410270.
- Carruthers, Peter (1996). *Language, thought and consciousness*. Cambridge University Press.
- Carruthers, Peter (1998). Animal subjectivity. *Psyche*, 4: 3.
- Carruthers, Peter (2000). *Phenomenal consciousness*. Cambridge: Cambridge.
- Carruthers, Peter (2019). *Human and animal minds*. Oxford: Oxford.
- Carruthers, Peter, and Rocco Gennaro (2001/2020). Higher-order theories of consciousness. *Stanford Encyclopedia of Philosophy* (Fall 2020 edition).
- Casadio, Andrea, Ferdinando Fiumara, Dario Sonetti, Pier Giorgio Montarolo, and Mirella Ghirardi (2004). Distribution of Sensorin Immunoreactivity in the Central Nervous System of *Helix pomatia*: Functional Aspects. *Journal of Neuroscience Research*, 75, 32-43.
- Chalmers, David J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2 (3), 200-219.
- Chalmers, David J. (1996). *The conscious mind*. Oxford University Press.
- Chalmers, David J. (2003/2010). The Matrix as metaphysics. In *The Character of Consciousness*. Oxford University Press.
- Chalmers, David J. (2004). *How can we construct a science of consciousness?* Online manuscript at <http://consc.net/papers/scicon.html>.
- Chalmers, David J. (2010a). *The Character of Consciousness*. Oxford University Press.
- Chalmers, David J. (2010b). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17 (9-10), 7-65.
- Chalmers, David J. (2012). *Constructing the world*. Oxford University Press.
- Chalmers, David J. (2016). The combination problem for panpsychism. In G. Brüntrup and L. Jaskolla, eds., *Panpsychism*. Oxford University Press.
- Chalmers, David J. (2017). The virtual and the real. *Disputatio*, 9, 309-352.
- Chalmers, David J. (2018). Structuralism as a response to skepticism. *Journal of Philosophy* 115, 625-660.
- Chalmers, David J. (2019a). The virtual as the digital. *Disputatio*, 11, 453-486.
- Chalmers, David J. (2019b). Three puzzles about spatial experience. In A. Pautz and D. Stoljar, eds., *Blockheads!* MIT Press.
- Chalmers, David J. (forthcoming). *Reality+*. W. W. Norton.

- Chalmers, David J., and Kelvin J. McQueen (forthcoming). Consciousness and collapse of the wave function. In S. Gao, ed., *Consciousness and quantum mechanics*. Oxford University Press.
- Chase, Ronald (2002). *Behavior and its neural control in gastropod molluscs*. Oxford: Oxford.
- Chen, Eddy Keming (2021). Time's arrow and self-locating probability. *Philosophy & Phenomenological Research*. DOI: 10.1111/phpr.12834.
- Chen, M. Keith, Venkat Lakshminarayanan, and Laurie R. Santos (2006). How basic are behavioral biases? Evidence from capuchin monkey trading behavior. *Journal of Political Economy*, 114, 517-537.
- Chen, Xiaoke, Mariano Gabitto, Yueqing Peng, Nicholas J. P. Ryba, and Charles S. Zuker (2011). A gustotopic map of taste qualities in the mammalian brain. *Science*, 333, 1262-1266.
- Chignell, Andrew (2010). Causal refutations of idealism. *Philosophical Quarterly*, 60, 487-507.
- Chignell, Andrew (2011). Causal refutations of idealism revisited. *Philosophical Quarterly*, 61, 184-186.
- Chomanski, Bartek (2019). Could there be scattered subjects of consciousness? *Phenomenology and the Cognitive Sciences*, 18, 775-789
- Chomsky, Noam (2009). The mysteries of nature: How deeply hidden? *Journal of Philosophy*, 106, 167-200.
- Christensen, David (2007). The epistemology of disagreement: The good news. *Philosophical Review*, 116, 187-217.
- Christensen, David (2021). Akritic (epistemic) modesty. *Philosophical Studies*, 178, 2191-2214.
- Christensen, David, and Jennifer Lackey, eds. (2013). *Disagreement*. Oxford University Press.
- Churchland, Patricia S. (1983). Consciousness: The transmutation of a concept. *Pacific Philosophical Quarterly*, 64, 80-95.
- Churchland, Patricia S. (2002). *Brain-wise*. MIT Press.
- Churchland, Paul M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78, 67-90.
- Churchland, Paul M. (1984/1988). *Matter and consciousness, rev. ed.* MIT Press.
- Cicero, Marcus Tullius (1st c. BCE/2001). *Cicero on the ideal orator*, trans. J. M. May and J. Wisse. Oxford University Press.
- Clark, Andy (2009). Spreading the joy? Why the machinery of consciousness is (probably) still in the head. *Mind*, 118, 963-993.
- Clark, Austen (1994). Beliefs and desires incorporated. *Journal of Philosophy*, 91, 404-425.
- Coeckelbergh, Mark (2012). *Growing moral relations*. Palgrave Macmillan.
- Cohen, Michael A., Daniel C. Dennett, and Nancy Kanwisher (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences*, 20, 324-335.
- Coliva, Annalisa (2010). *Moore and Wittgenstein*. Palgrave.
- Comesaña, Juan, and Peter Klein (2001/2019). Skepticism. *Stanford Encyclopedia of Philosophy* (Winter 2019 edition).
- Copeland, B. Jack (1997/2020). The Church-Turing thesis. *Stanford Encyclopedia of Philosophy* (Summer 2020 edition).
- Cowling, Sam (2013). Ideological parsimony. *Synthese*, 190, 3889-3908.
- Cowling, Sam (2015/2016). Haecceitism. *Stanford Encyclopedia of Philosophy* (Fall 2016 edition).

- Crane, Tim, and D. H. Mellor (1990). There is no question of physicalism. *Mind*, 99, 185-206.
- Crary, Alice, and Rupert Read (2000). *The new Wittgenstein*. Routledge.
- Crawford, Lyle (2013). Freak observers and the simulation argument. *Ratio*, 26, 250-264.
- Crick, Francis (1994). *The astonishing hypothesis*. Charles Scribner's Sons.
- Croll, Roger P. and Ronald Chase (1980). Plasticity of olfactory orientation to foods in the snail *Achatina fulica*. *Journal of Comparative Physiology*, 136, 267-277.
- Crow, Terry J., and Daniel L. Alkon (1978). Retention of an associative behavioral change in *Hermisenda*. *Science*, 201 (4362), 1239-1241.
- Cuda, Tom (1985). Against neural chauvinism. *Philosophical Studies*, 48, 111-127.
- Dahan, Orli (2017). The problem of other (group) minds (response to Schwitzgebel). *Philosophia*, 45, 1099-1112.
- Darling, Kate (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior toward robotic objects. In R. Calo, A. M. Froomkin, and I. Kerr, eds., *Robot Law*. Edward Elgar.
- Darling, Kate (2017). "Who's Johnny?" Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, G. Bekey, K. Abney, and R. Jenkins, eds., *Robot Ethics 2.0*. Oxford University Press.
- Darling, Kate (2021). *The new breed*. Henry Holt.
- Davidson, Donald (1970/2001). Mental events. In D. Davidson, *Essays on actions and events*. Oxford University Press.
- Davidson, Donald (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*, 61, 441-58.
- De Cruz, Helen, and Johan De Smedt (2017). How psychological dispositions influence the theology of the afterlife. In Y. Nagasawa and B. Matheson (eds.), *The Palgrave handbook of the afterlife*. Palgrave Macmillan.
- De Graaf, Maartje M.A. Frank A. Hindriks, and Koen V. Hindriks (2021). Who wants to grant robots rights? *HRI '21 Companion*, 38-46.
- De Simone, Andrea, Alan H. Guth, Andrei Linde, Mahdiyar Noorbala, Michael P. Salem, and Alexander Vilenkin (2010). Boltzmann brains and the scale-factor cutoff measure of the multiverse. *Physical Review D*, 82, 063520.
- De Vos, A. (2000). *Non-planar driver's side rearview mirrors*. United States Department of Transportation report. DOT HS 809 149.
- DeGrazia, David (2008). Moral status as a matter of degree? *Southern Journal of Philosophy*, 46, 181-196.
- Dehaene, Stanislas (2014). *Consciousness and the brain*. Viking.
- Dehaene, Stanislas, and Jean-Pierre Changeux (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70, 200-227.
- Dennett, Daniel C. (1991). *Consciousness explained*. Little, Brown, and Company.
- Dennett, Daniel C. (1996). *Kinds of minds*. Basic Books.
- Dennett, Daniel C. (1998). *Brainchildren*. MIT Press.
- Dennett, Daniel C. (2005). *Sweet dreams*. MIT Press.
- Dennett, Daniel C. (2017). *From bacteria to Bach and back*. W. W. Norton.
- DeRose, Keith (1995). Solving the skeptical problem. *Philosophical Review*, 104, 1-52.
- DeRose, Keith (2017). *The appearance of ignorance*. Oxford University Press.
- Des Chene, Dennis (2006). Animal as category: Bayle's "Rorarius". In J. E. H. Smith, ed., *The problem of animal generation in early modern philosophy*. Cambridge University Press.

- Descartes, René (1637/1985). *Discourse on the method*. In J. Cottingham, R. Stoothoff, and D. Murdoch, eds., *The Philosophical Writings of Descartes, vol. 1*. Cambridge University Press.
- Descartes, René (1641/1984). *Meditations on first philosophy*. In J. Cottingham, R. Stoothoff, and D. Murdoch, eds., *The philosophical writings of Descartes, vol. 2*. Cambridge University Press.
- Descartes, René (1647/1985). *Principles of philosophy*. In J. Cottingham, R. Stoothoff, and D. Murdoch, eds., *The philosophical writings of Descartes, vol. 1*. Cambridge University Press.
- Descartes, René (1649/1991). Letter to More, 5 Feb. 1649. In J. Cottingham, R. Stoothoff, D. Murdoch, and A. Kenny, eds., *The philosophical writings of Descartes, vol. 3*. Cambridge University Press.
- Descartes, René (1649 /1985). *The Passions of the Soul*, in J. Cottingham, R. Stoothoff, and D. Murdoch, eds., *The philosophical writings of Descartes, vol. 1*. Cambridge University Press.
- DeWitt, Bryce S. (1970). Quantum mechanics and reality. *Physics Today*, 23 (9), 30-35.
- Di Giorgio, Elisa, Marco Lunghi, Francesca Simion, and Giorgio Vallortigara (2017). Visual cues of motion that trigger animacy perception at birth: The case of self-propulsion. *Developmental Science*, 20, e12394.
- Dick, Philip K. (1968). *Do androids dream of electric sheep?* Doubleday.
- Dicker, Georges (2008). Kant's refutation of idealism. *Noûs*, 42, 80-108.
- Dicker, Georges (2011). Kant's refutation of idealism: A Reply to Chignell. *Philosophical Quarterly*, 61, 175–83
- Dicker, Georges (2012). Kant's refutation of idealism: Once more unto the breach. *Kantian Review*, 17, 191–195.
- Diogenes Laertius (3rd c. CE/1974). *Lives, teachings, and sayings of famous philosophers*, trans. R.D. Hicks. Cambridge, MA: Harvard.
- Doerig, Adrien, Aaron Schurger, Kathryn Hess, and Michael H. Herzog (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, 72, 49-59.
- Domhoff, G. William (2003). *The scientific study of dreams*. Washington, DC: American Psychological Association.
- Dretske, Fred (1988). *Explaining behavior*. MIT Press.
- Dretske, Fred (1995). *Naturalizing the mind*. MIT Press.
- Dretske, Fred (2003). Skepticism: What perception teaches. In S. Luper, ed., *The skeptics*. Ashgate.
- Dupré, John (1983). The disunity of science. *Mind*, 92, 321-346.
- Dupré, John (2012). *Processes of life*. Oxford University Press.
- Earman, John (1992). *Bayes or bust?* Cambridge, MA: MIT.
- Easson, Damien A., and Robert H. Brandenberger (2001). Universe generation from black hole interiors. *Journal of High Energy Physics*, 2001 (6): 024.
- Easwaran, Kenny, Alan Hájek, Paolo Mancosu, and Graham Oppy (2021). Infinity. *Stanford Encyclopedia of Philosophy* (Winter 2021 edition).
- Echevarria, René (1993). *Star Trek: The next generation: Ship in a Bottle*, dir. A. Singer. Episode 6x12.
- Edelman, Shimon (2008). *Computing the mind*. Oxford University Press.

- Egan, Greg (1992). Closer. *Eidolon*, issue 9.
http://www.eidolon.net/old_site/issue_09/09_closr.htm.
- Egan, Greg (1994). *Permutation City*. Millennium.
- Egan, Greg (1997). *Diaspora*. Millennium.
- Einstein, A., B. Podolsky, and N. Rosen (1935). Can quantum mechanical description of physical reality be considered complete? *Physical Review*, 47, 777-780.
- Ekstrom, Arne D., and Eve A. Isham (2017). Human spatial navigation: Representations across dimensions and scales. *Current Opinion in Behavioral Sciences*, 17, 84-89.
- Elder, Crawford (2011). *Familiar objects and their shadows*. Cambridge University Press.
- Elga, Adam (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60, 143-147.
- Elga, Adam (2013). The puzzle of the unmarked clock and the new rational reflection principle. *Philosophical Studies*, 164, 127-139.
- Epstein, Brian (2018). Social ontology. *Stanford Encyclopedia of Philosophy* (Summer 2018 edition).
- Erickson, Robert P. (2008). A study of the science of taste: On the origins and influence of the core ideas. *Behavioral and Brain Sciences*, 31, 59-75.
- Espinas, Alfred (1877/1924). *Des sociétés animales*, 3rd ed. Félix Alcan.
- Estrada, Daniel (2017). Robot rights: Cheap, yo!" *Made of Robots*. Episode 1, May 24.
<https://www.madeofrobots.com/2017/05/24/episode-1-robot-rights-cheap-yo>.
- Fancher, Hampton, Michael Green, and Denis Villeneuve (2017). *Blade Runner 2049*. Warner Brothers.
- Fancher, Hampton, David Peoples, and Ridley Scott (1982). *Blade Runner*. Warner Brothers.
- Favela, Luis H. (2019). Integrated Information Theory as a complexity science approach to consciousness. *Journal of Consciousness Studies*, 26 (1-2), 21-47.
- Faye, Jan (2001/2021). Backward causation. *Stanford Encyclopedia of Philosophy* (Spring 2021 Edition).
- Faye, Jan (2008/2019). Copenhagen interpretation of quantum mechanics. *Stanford Encyclopedia of Philosophy* (Winter 2019 edition).
- Fekete, Tomer, Cees van Leeuwen, and Shimon Edelman (2016). System, subsystem, hive: Boundary problems in computational theories of consciousness. *Frontiers in Psychology*, 7, article 1041.
- Feyerabend, Paul (1963). Mental events and the brain. *Journal of Philosophy*, 40, 295-296.
- Fiala, Brian, Adam Arico, and Shaun Nichols (2012). The psychological origins of dualism. In E. Slingerland and M. Collard, eds., *Creating consilience*. Oxford University Press.
- Fichte, Johann G. (1797/2000). *Foundations of natural right*. Ed. and trans., F. Neuhouser and M. Bauer. Cambridge University Press.
- Figdor, Carrie (2018). *Pieces of mind*. Oxford University Press.
- Fine, Terrence L. (2008). Evaluating the Pasadena, Altadena, and St Petersburg gambles. *Mind*, 117, 613-632.
- Fischer, John M. (1994). Why immortality is not so bad. *International Journal of Philosophical Studies*, 2, 257-270.
- Fischer, John M., and Benjamin Mitchell-Yellin (2014). Immortality and boredom. *Journal of Ethics*, 18, 353-372.
- Floridi, Luciano, and Massimo Chiriatti (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681-694.

- Floris, Giacomo (2021). A pluralist account of the basis of moral status. *Philosophical Studies*, 178, 1859-1877.
- Fodor, Jerry A. (1968). The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy*, 65, 627-640.
- Foley, Richard (2001). *Intellectual trust in oneself and others*. Cambridge University Press.
- Frances, Bryan (2013). Philosophical Renegades. In D. Christensen and J. Lackey, eds., *Disagreement*. Oxford University Press.
- Frances, Bryan (2014). *Disagreement*. Polity Press.
- Frances, Bryan (2021). Philosophical proofs against common sense. *Analysis*, 81, 18-26.
- Frances, Bryan, and Jonathan Matheson (2018/2019). Disagreement. *Stanford Encyclopedia of Philosophy* (Winter 2019 edition).
- Frankish, Keith (2010). Dual-process and dual-systems theories of reasoning. *Philosophy Compass*, 5, 914-926.
- Frankish, Keith (2012). Quining diet qualia. *Consciousness & Cognition*, 21, 667-676.
- Frankish, Keith (2016). *Illusionism as a theory of consciousness*. Imprint Academic.
- Frankish, Keith (2018). What is it like to be a bot? *Philosophy Now*, issue #126.
https://philosophynow.org/issues/126/What_Is_It_Like_To_Be_A_Bot
- Frege, Gottlob (1884/1968). *The foundations of arithmetic*, trans. J. L. Austin. Northwestern University Press.
- Frege, Gottlob (1918/1956). The thoughts: A logical inquiry, trans. P. T. Geach. *Mind*, 56, 289-311.
- Friederich, Simon (2017/2018). Fine-tuning. *Stanford Encyclopedia of Philosophy* (Winter 2018 edition).
- Friedman, Daniel A., and Eirik Søvik (forthcoming). The ant colony as a test for scientific theories of consciousness. *Synthese*.
- Frolov, V. P., M. A. Markov, and V.F. Mukhanov (1989). Through a black hole into a new universe? *Physics Letters B*, 216 (3-4), 272-276.
- Gallagher, Shaun (2017). *Enactivist interventions*. Oxford University Press.
- Galilei, Galileo (1638/1914). *Dialogues concerning the two new sciences*, trans. H. Crew and A. De Salvio. Macmillan.
- Ganeri, Jonardon (2013). Philosophy as a way of life: Spiritual exercises from the Buddha to Tagore. In M. Chase, S. R. L. Clark, M. McGhee, eds., *Philosophy as a way of life*. Wiley Blackwell.
- Garber, Megan (2013). Funerals for fallen robots. *The Atlantic* (Sep 20).
<https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallen-robots/279861>.
- Garfield, Jay L. (2015). *Engaging Buddhism*. Oxford: Oxford.
- Garriga, Jaume, and Alexander Vilenkin (1998). Recycling universe. *Physical Review D*, 57, 2230-2244.
- Garriga, Jaume, Alexander Vilenkin, and Jun Zhang (2016). Black holes and the multiverse. *Journal of Cosmology and Astroparticle Physics*, 2016 (2): 064.
- Gawronski, Bertram, and Skylar M. Brannon (2017). Attitudes and the implicit-explicit dualism. In D. Albarracín and B. T. Johnson, eds., *The handbook of attitudes, vol. 1*. Routledge.
- Gelperin, Alan (2013). Associative memory mechanisms in terrestrial slugs and snails. In R. Menzel and P.R. Benjamin, eds., *Invertebrate learning and memory*. Amsterdam: Elsevier.

- Gendler, Tamar S. (2008a). Alief and belief. *Journal of Philosophy*, 105, 634–663.
- Gendler, Tamar S. (2008b). Alief in action, and reaction. *Mind & Language*, 23, 552–585.
- Gennaro, Rocco J. (2012). *The consciousness paradox*. Cambridge, MA: MIT.
- Garreau, Joel (2007). Bots on the ground: In the field of battle (or even above it), robots are a soldier’s best friend. *Washington Post* (May 6). https://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009_pf.html.
- Gibson, James J. (1979/2015). *The ecological approach to visual perception, classic edition*. Psychology Press.
- Gilbert, Margaret (1989). *On social facts*. Princeton University Press
- Godfrey-Smith, Peter (2009). *Darwinian populations and natural selection*. Oxford University Press.
- Godfrey-Smith, Peter (2013). Darwinian individuals. In *From groups to individuals*, ed. F. Bouchard and P. Huneman. MIT Press.
- Godfrey-Smith, Peter (2016a). Mind, matter, and metabolism. *Journal of Philosophy*, 113, 481–506.
- Godfrey-Smith, Peter (2016b). *Other minds*. Farrar, Straus and Giroux.
- Godfrey-Smith, Peter (2017). The evolution of consciousness in phylogenetic context. In K. Andrews and J. Beck, *The Routledge Handbook of Animals Minds*.
- Godfrey-Smith, Peter (2020). *Metazoa*. Farrar, Straus, and Giroux.
- Goff, Philip (2013). “Orthodox property dualism + the linguistic theory of vagueness = Panpsychism”. In R. Brown, ed., *Consciousness Inside and Out*. Dordrecht: Springer (p. 75-91).
- Goff, Philip (2017). *Consciousness and fundamental reality*. Oxford University Press.
- Gold, Jonathan C. (2011/2021). Vasubandhu. *Stanford Encyclopedia of Philosophy* (Spring 2021 edition).
- Goldberg, Sanford (2009). Reliabilism in Philosophy. *Philosophical Studies*, 142, 105-117.
- Goldberg, Sanford (2013). Defending philosophy in the face of systematic disagreement. In D. E. Machuca, ed., *Disagreement and skepticism*. Routledge.
- Goldman, Alvin I., and Bob Beddor (2008/2021). Reliabilist epistemology. *Stanford Encyclopedia of Philosophy* (Summer 2021 edition).
- Goldman, Alvin I. (1997). Science, publicity, and consciousness. *Philosophy of Science*, 64, 525-545.
- Gomes, Anil (2018). Skepticism about other minds. In *Skepticism: From antiquity to the present*, ed. D. E. Machuca and B. Reed. Bloomsbury Academic.
- Gopnik, Alison (2020). Childhood as a solution to explore - exploit tensions. *Philosophical Transactions of the Royal Society B*, 375: 20190502.
- Gopnik, Alison, and Andrew N. Meltzoff (1997). *Words, thoughts, and theories*. MIT Press.
- Gopnik, Alison, and Eric Schwitzgebel (1998). Whose concepts are they, anyway? The role of philosophical intuition in empirical psychology. In *Rethinking intuition*, ed. M.R. DePaul and W. Ramsey. Rowman and Littlefield.
- Gott, J. Richard, III (1993). Implications of the Copernican principle for our future prospects. *Nature*, 363, 315-319.
- Gott, J. Richard, III, Mario Jurić, David Schlegel, Fiona Hoyle, Michael Vogeley, Max Tegmark, Neta Bahcall, and Jon Brinkmann (2005). A map of the universe. *Astrophysical Journal*, 624, 463-484.

- Grayling, A. C. (2005). *Descartes*. Walker.
- Graziano, Michael S. A. (2019). *Rethinking consciousness*. W. W. Norton.
- Greaves, Hilary (2011). In search of (spacetime) structuralism. *Philosophical Perspectives*, 25, 189-204.
- Greaves, Hilary (2016). XIV – Cluelessness. *Proceedings of the Aristotelian Society*, 116, 311-339.
- Greco, John (2008). Skepticism about the external world. In J. Greco (ed.), *The Oxford handbook of skepticism*. Oxford: Oxford University Press.
- Grego, Richard (2020). Mapping Sri Aurobindo's metaphysics of consciousness onto Western philosophies of mind. In D. A. Mahapatra (ed.), *The philosophy of Sri Aurobindo*. Bloomsbury.
- Greene, Brian (2011). *The hidden reality*. Random House.
- Greene, Brian (2020). *Until the end of time*. Penguin.
- Gruen, Lori (2011/2021). *Ethics and animals*. Cambridge University Press.
- Grundmann, Thomas (2002). Die Struktur des skeptischen Traumarguments. *Grazer Philosophische Studien*, 64, 57-81.
- Gunkel, David J. (2018). *Robot rights*. MIT Press.
- Guyer, Paul (1987) *Kant and the claims of knowledge*. Cambridge: Cambridge.
- Hacker, P.M.S. (2012). The sad and sorry history of consciousness: Being, among other things, a challenge to the “consciousness studies” community. *Royal Institute of Philosophy Supplement*, 70, 149-168.
- Hadot, Pierre (1995). *Philosophy as a way of life*, trans. A.I. Davidson. Malden, MA: Blackwell.
- Hájek, Alan (2003). Waging war on Pascal's wager. *Philosophical Review*, 112, 27-56.
- Haldane, J. B. S. (1927). *Possible worlds and other essays*. Chatto and Windus.
- Hampton, Robert R. (2019). Monkey metacognition could generate more insight. *Animal Behavior and Cognition*, 6, 230-235.
- Hanson, Jake R. (2020). My experience with Integrated Information Theory. Blog post at <https://jakerhanson.weebly.com/blog/my-graduate-experience-with-integrated-information-theory-iit> (Jun 24).
- Hanson, Robin (2018). *The age of em*. Oxford University Press.
- Harnad, Stevan (2003). Can a machine be conscious? How? *Journal of Consciousness Studies*, 10 (3), 67-75.
- Harris, Charles S. (1965). Perceptual adaptation to inverted, reversed, and displaced vision. *Psychological Review*, 72, 419-444.
- Harris, Charles S. (1980). Insight or out of sight? Two examples of perceptual plasticity in the human adult. In *Visual coding and adaptability*, ed. C.S. Harris. Hillsdale, NJ: Erlbaum.
- Haslanger, Sally (2008). Changing the ideology and culture of philosophy: Not by reason (alone). *Hypatia*, 23, 210-22.
- Hatfield, Gary (2005/2009). Introspective evidence in psychology. In G. Hatfield, *Perception and Cognition*. Oxford: Oxford.
- Hawking, S. W. (1974). Black hole explosions? *Nature*, 248, 30-31.
- Hawking, Stephen, and Leonard Mlodinow (2010). *The grand design*. Bantam.
- Hawkins, Robert D., and John H. Byrne (2015). Associative learning in invertebrates. *Cold Spring Harbor Perspectives in Biology*. doi: 10.1101/cshperspect.a021709

- Hawkins, Robert D., N. Lalevic, G. A. Clark, and E. R. Kandel (1989). Classical conditioning of the Aplysia siphon-withdrawal reflex exhibits response specificity. *Proceedings of the National Academy of Sciences*, 86, 7620-7624.
- Hawkins, Robert D., Tracey E. Cohen, and Eric R. Kandel (2006). Dishabituation in Aplysia can involve either reversal of habituation or superimposed sensitization. *Learning & Memory*, 13, 397-403.
- Heavey, Christopher L., and Russell T. Hurlburt (2008). The phenomena of inner experience. *Consciousness and Cognition*, 17, 798-810.
- Hegel, G.W.F. (1807/1977) *Phenomenology of spirit*. Trans. A.V. Miller. Oxford: Oxford.
- Heil, John (1998/2020). *Philosophy of mind*, 4th ed. Routledge.
- Hempel, Carl G. (1980). Comments on Goodman's *Ways of Worldmaking*. *Synthese* 45, 193-199.
- Henderson, David, Terence Horgan, Matjaž Potrč, and Hannah Tierney (2017). Nonconciliation in peer disagreement: Its phenomenology and its rationality. *Grazer Philosophische Studien*, 94, 194-225.
- Herzberg, Fred, and Anne Herzberg (1962). Observations on reproduction in *Helix aspersa*. *American Midland Naturalist*, 68, 297-306.
- Hetherington, Stephen (2016). Understanding fallible warrant and fallible knowledge: Three proposals. *Pacific Philosophical Quarterly*, 97, 270-282.
- Hilbert, Martin, and Priscila López (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332, 60-65.
- Hill, Christopher S. (1991). *Sensations*. Cambridge University Press.
- Hill, Christopher S. (2009). *Consciousness*. Cambridge University Press.
- Hill, Christopher S., and Joshua Schechter (2007). Hawthorne's lottery puzzle and the nature of belief. *Philosophical Issues*, 17, 102-122.
- Hobson, J. Allan, Edward F. Pace-Schott, and Robert Stickgold (2000). Dreaming and the brain. *Behavioral and Brain Sciences*, 793-842.
- Hodge, K. Mitch (2008). Descartes' mistake: How afterlife beliefs challenge the assumption that humans are intuitive Cartesian substance dualists. *Journal of Cognition and Culture*, 8, 387-415.
- Hoel, Erik, Larissa Albantakis, and Giulio Tononi (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110, 10.1073/pnas.1314922110.
- Holmes, Nicholas P., and Charles Spence (2004). The body schema and multisensory representation(s) of peripersonal space. *Cognitive Processing*, 5, 94-105.
- Hopfield, Jessica F., and Alan Gelperin (1989). Differential conditioning to a compound stimulus and its components in the terrestrial mollusk *Limax maximus*. *Behavioral Neuroscience*, 103, 329-333.
- Huebner, Bryce (2013). *Macro cognition*. Oxford University Press
- Huebner, Bryce (2016). The group mind in commonsense psychology. In J. Sytsma and W. Buckwalter, eds., *Blackwell Companion to Experimental Philosophy*. John Wiley & Sons.
- Huebner, Bryce, Michael Bruno, and Hagop Sarkissian (2010). What does the nation of China think about phenomenal states? *Review of Philosophy and Psychology*, 1, 225-243.
- Huemer, Michael (2021). Existence is evidence of immortality. *Noûs*, 55, 128-151.

- Huggett, Nick, Carl Hofer, and James Read (2006/2021). Absolute and relational space and motion: Post-Newtonian theories. *Stanford Encyclopedia of Philosophy* (Spring 2022 edition).
- Hume, David (1740/1978). *A treatise of human nature*, ed. L.A. Selby-Bigge and P.H. Nidditch. Oxford University Press.
- Hume, David (1779/1947). *Dialogues concerning natural religion*, ed. N.K. Smith. Indianapolis: Bobbs-Merrill.
- Humphrey, Nicholas (1992). *A History of the Mind*, London: Chatto & Windus.
- Humphrey, Nicholas (2011). *Soul dust*. Princeton: Princeton.
- Hurlburt, Russell T. (2011). *Investigating pristine inner experience*. Cambridge University Press.
- Hurlburt, Russell T., and Eric Schwitzgebel (2007). *Describing inner experience? Proponent meets skeptic*. MIT Press.
- Hurley, Susan (1998). *Consciousness in action*. Harvard University Press.
- Hurley, Susan, and Alva Noë (2003). Neural plasticity and consciousness. *Biology and Philosophy*, 18, 131-168.
- Hutchins, Edwin (1995). *Cognition in the wild*. MIT Press.
- Ichikawa, Jonathan (2008). Skepticism and the imagination model of dreaming. *Philosophical Quarterly*, 58, 519-527.
- Ichikawa, Jonathan (2009). Dreaming and imagination. *Mind & Language*, 24, 103-121.
- Ichikawa, Jonathan J. (2016). Imagination, dreaming, and hallucination. In A. Kind, ed., *The Routledge handbook of philosophy of imagination*. Routledge.
- Ichikawa, Jonathan J. (2017). *Contextualising knowledge*. Oxford University Press.
- Irvine, Elizabeth (2013). *Consciousness as a scientific concept*. Dordrecht: Springer.
- Irvine, Liz, and Mark Sprevak (2020). Eliminativism about consciousness. In U. Kriegel, ed., *Oxford handbook of the philosophy of consciousness*. Oxford University Press.
- Ishiguro, Kazuo (2021). *Klara and the sun*. Knopf.
- Jackson, Frank (1986). What Mary didn't know. *Journal of Philosophy*, 83, 291-295.
- Jackson, Frank (1998). Postscript on qualia. In F. Jackson, *Mind, method, and conditionals*. Routledge.
- James, William (1884). On some omissions of introspective psychology. *Mind*, 9 (old series), 1-26.
- James, William (1890/1918). *Principles of psychology, vol. 1*. Henry Holt.
- Jaworska, Agnieszka and Julie Tannenbaum (2013/2021). The grounds of moral status. *Stanford Encyclopedia of Philosophy* (Spring 2021 edition).
- Jaynes, Julian (1976). *The origins of consciousness in the breakdown of the bicameral mind*. Houghton Mifflin.
- Jennings, Carolyn Dicey (2015). Consciousness without attention. *Journal of the American Philosophical Association*, 1, 276-295.
- Johnson, Susan C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society B* 358, 549-559.
- Kagan, Shelly (2019). *How to count animals, more or less*. Oxford University Press.
- Kahneman, Daniel, and Amos Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-292.
- Kammerer, François (2015). How a materialist can deny that the United States is probably conscious – response to Schwitzgebel. *Philosophia*, 43, 1047-1057.

- Kammerer, François (2021). The illusion of conscious experience. *Synthese*, 198, 845-866.
- Kandel, Eric R. (2001). The molecular biology of memory storage: a dialogue between genes and synapses. *Science*, 294 (5544), 1030-1038.
- Kant, Immanuel (1781/1787/1998). *Critique of pure reason*, ed. and trans. P. Guyer and A.W. Wood. Cambridge University Press.
- Kaplan, David L., Jason Boyles, Bart H. Dunlap, Shriharsh P. Tendulkar, Adam T. Deller, Scott M. Ransom, Maura A. McLaughlin, Duncan R. Lorimer, and Ingrid H. Stairs (2014). Companion to PSR J2222-0137: The coolest known white dwarf? *Astrophysical Journal*, 789 (2), 119.
- Kelly, Thomas (2005). The epistemic significance of disagreement. In T. S. Gendler and J. Hawthorne, eds. *Oxford Studies in Epistemology*, vol. 1. Oxford University Press.
- Kelly, Thomas (2010). Peer disagreement and higher-order evidence. In R. Feldman and T. A. Warfield, *Disagreement*. Oxford University Press.
- Kerkut, G.A., J.D.C. Lambert, R.J. Gayton, Janet E. Loker, and R.J. Walker (1975). Mapping of nerve cells in the subesophageal ganglia of *Helix aspersa*. *Comparative Biochemistry and Physiology Part A: Physiology*, 50 (1), 1-25.
- Kim, Jaegwon (1989). The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association*, 63 (3), 31-47.
- Kim, Jaegwon (1998). *Mind in a physical world*. MIT Press.
- Kim, Jaegwon (2005). *Physicalism, or something near enough*. Princeton, NJ: Princeton University Press.
- Kimura, Tetsuya, Shoichi Toda, Tatsuhiko Sekiguchi, and Yutaka Kirino (1998). Behavioral modulation induced by food odor aversive conditioning and its influence on the olfactory responses of an oscillatory grain network in the slug *Limax marginatus*. *Learning and Memory*, 4, 365-375.
- Kirchhoff, Michael D., and Julian Kiverstein (2019). *Extended consciousness and predictive processing*. Routledge.
- Kirk, Robert (1974). Zombies v. materialists. *Proceedings of the Aristotelian Society*, Suppl. 48, 135-152.
- Kirk, Robert (2005). *Zombies and consciousness*. Oxford University Press.
- Kittay, Eva F. (2005). At the margins of moral personhood. *Ethics*, 116, 100-131.
- Klein, Colin (2007). Kicking the Kohler habit. *Philosophical Psychology*, 20, 609-619.
- Klingemann, Mario (2020). Twitter post as @quasimondo on Jul. 18, 8:25 a.m. <https://twitter.com/quasimondo/status/1284509525500989445>.
- Knobe, Joshua, and Jesse Prinz (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7, 67-83.
- Knuth, Donald E. (1976). Mathematics and computer science: Coping with finiteness: Advances in our ability to compute are bringing us substantially closer to ultimate limitations. *Science*, 194 (4271), 1235-1242.
- Koch, Christof (2012). *Consciousness: Confessions of a romantic reductionist*. MIT Press.
- Koene, Joris M. (2006). Tales of two snails: Sexual selection and sexual conflict in *Lymnaea stagnalis* and *Helix aspersa*. *Integrative & Comparative Biology* 46, 419-429.
- Kohler, Ivo (1951/1964). *The formation and transformation of the perceptual world*. Trans. H. Fiss. *Psychological Issues*, 3 (4), monograph 12. International Universities Press.
- Kohler, Ivo (1962). Experiments with goggles. *Scientific American*, 206, 62-72.

- Korman, Daniel (2011/2020). Ordinary objects. *Stanford Encyclopedia of Philosophy* (Fall 2020 edition).
- Kornblith, Hilary (1998). The role of intuition in philosophical inquiry: An account with no unnatural ingredients. In *Rethinking intuition*, ed. M.R. DePaul and W. Ramsey. Rowman and Littlefield.
- Kornblith, Hilary (2013). Is philosophical knowledge possible? In D.E. Machuca, ed., *Disagreement and Skepticism*. Routledge.
- Korsgaard, Christine M. (2018). *Fellow creatures*. Oxford University Press.
- Kotzen, Matthew (2021). What follows from the possibility of Boltzmann Brains? In S. Dasgupta, R. Dotan, and B. Weslake, eds., *Current controversies in philosophy of science*. Routledge.
- Kriegel, Uriah (2009). *Subjective consciousness*. Oxford University Press.
- Kriegel, Uriah (2011). Two defenses of common-sense ontology. *Dialectica*, 65, 177-204.
- Kriegel, Uriah (2013). The epistemological challenge of revisionary metaphysics. *Philosophers' Imprint*, 13 (12).
- Kriegel, Uriah (2015). *The varieties of consciousness*. Oxford University Press.
- Kriegel, Uriah (2019). The value of consciousness. *Analysis*, 79, 503-520.
- Krueger, Joel (2013). Merleau-Ponty on shared emotions and the joint ownership thesis. *Continental Philosophy Review*, 46, 509-531.
- Kulstad, Mark, and Laurence Carlin (1997/2020). Leibniz's philosophy of mind. *Stanford Encyclopedia of Philosophy* (Winter 2020 edition).
- Kurdi, Benedek, Allison E. Seitchik, Jordan R. Axt, Timothy J. Carroll, Arpi Karapetyan, Neela Kaushik, Diana Tomezsko, Anthony G. Greenwald, Mahzarin R. Banaji (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American Psychologist*, 74, 569-586.
- Kurki, Visa A. J. (2019). *A theory of legal personhood*. Oxford University Press.
- Kurzweil, Ray (2005). *The singularity is near*. Penguin.
- La Mettrie, Julien Offray de 1748 (1994). *Man a machine and man a plant*, ed. J. Lieber, trans. R.A. Watson and M. Rybalka. Hackett.
- LaBerge, Stephen, and Howard Rheingold (1990). *Exploring the world of lucid dreaming*. Ballantine.
- Lackey, Jennifer (2010). A justificationist view of disagreement's epistemic significance. In A. Haddock, A. Millar, and D. Pritchard, eds., *Social epistemology*. Oxford University Press.
- Ladyman, James, and Don Ross (2007). *Every thing must go*. Oxford University Press.
- Lafleur, Laurence J. (1952). Solipsism. *Review of Metaphysics*, 5, 523-528.
- Lam, Vincent (2017). Structuralism in the philosophy of physics. *Philosophy Compass*, 12, e12421.
- Lamme, Victor A. F. (2018). Challenges for theories of consciousness: seeing or knowing, the missing ingredient and how to deal with panpsychism. *Philosophical Transactions of the Royal Society B*, 373: 20170344. <http://dx.doi.org/10.1098/rstb.2017.0344>
- Lane, Jonathan D. (2021). Constructing ideas of the supernatural. *Journal of Cognition and Development*, 22, 343-355.
- Langton, Rae (1998). *Kantian humility*. Oxford: Oxford.
- Larson, Glen A., and Ronald D. Moore (2004-2009). *Battlestar Galatica*. NBC Universal television series.

- Lasonen-Aarnio, Maria (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, 88, 314-345.
- Le Bon, Gustave (1895/1995). *The crowd*, ed. R.A. Nye. Transaction.
- Leckie, Ann (2013). *Ancillary justice*. Orbit Books.
- Lederhendler, I. Izja, Serge Gart and Daniel L. Alkon (1986). Classical conditioning of *Hermisenda*: Origin of a new response. *Journal of Neuroscience*, 6, 1325-1331.
- Lee, Geoffrey (2006). The experience of left and right. In T. S. Gendler and J. Hawthorne, eds., *Perceptual experience*. Oxford University Press.
- Lee, Geoffrey (2019). Alien subjectivity and the importance of consciousness. In A. Pautz and D. Stoljar, eds., *Blockheads!* MIT Press.
- Leibniz, G.W. (1714/1989). The principles of philosophy, or, the monadology. In *Philosophical Essays*, ed. and trans. R. Ariew and D. Garber. Hackett.
- Lem, Stanislaw (1961/1970) *Solaris*, trans. J. Kilmartin and S. Cox. San Diego, CA: Harcourt.
- Lenman, James (2000). Consequentialism and cluelessness. *Philosophy & Public Affairs*, 29, 342-370.
- León, Felipe, Thomas Szanto, and Dan Zahavi (2019). Emotional sharing and the extended mind. *Synthese*, 196, 4847–4867.
- Leonard, C. Danielle, Philip Bull, and Rupert Allison (2016). Spatial curvature endgame: Reaching the limit of curvature determination. *Physical Review D*, 94, 023502.
- Lerner, Adam B. (2020). What it's like to be a state? An argument for state consciousness. *International Theory*. doi.org/10.1017/S1752971919000277.
- Levesque, Hector J. (2011). *The Winograd Schema challenge*. <http://commonsensereasoning.org/2011/papers/Levesque.pdf>.
- Levin, Janet (2000). Dispositional theories of color and the claims of common sense. *Philosophical Studies*, 100, 151-174.
- Lewis, David K. (1979). Attitudes de dicto and de se. *Philosophical Review*, 88, 513-543.
- Lewis, David K. (1980). Mad pain and Martian pain. In *Readings in philosophy of psychology*, ed. N. Block. Harvard University Press.
- Lewis, David K. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343-377.
- Lewis, David K. (1984). Putnam's paradox. *Australasian Journal of Philosophy*, 62, 221-236.
- Lewis, David K. (1986). *On the plurality of worlds*. Blackwell.
- Lewis, David K. (1988). What experience teaches. In W. Lycan, ed., *Mind and Cognition*. Basil Blackwell.
- Lewis, David K. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74, 549-567.
- Liao, S. Matthew (2020). The moral status and rights of artificial intelligence. In S. M. Liao, ed., *Ethics of artificial intelligence*. Oxford University Press.
- Lin, Eden (2021). The experience requirement on well-being. *Philosophical Studies*, 178, 867-886.
- Lind, Hans (1989). Homing to hibernating sites in *Helix pomatia* involving detailed long-term memory. *Ethology*, 81, 221-234.
- Lind, Hans (1990). Strategies of spatial behavior in *Helix pomatia*. *Ethology*, 86, 1-18.
- Linde, Andrei (2015/2017). *A brief history of the multiverse*. ArXiv. arXiv:1512.01203v3.
- Linden, David E.J., Ulrich Kallenbach, Armin Heinecke, Wolf Singer, and Rainer Goebel (1999). The myth of upright vision: A psychophysical and functional imaging study of adaptation to inverting spectacles. *Perception*, 28, 469–481.

- List, Christian (2018). What is it like to be a group agent? *Noûs*, 52, 295-318.
- List, Christian, and Philip Pettit (2011). *Group agency*. Oxford University Press.
- Lloyd, Ignacio, Vanesa Fernández, and Félix Acebes (2006). Conditioning of tentacle lowering in the snail (*Helix aspersa*): Acquisition, latent inhibition, overshadowing, second-order conditioning, and sensory preconditioning. *Learning & Behavior*, 34, 305-314.
- Locke, John (1689/1975) *Essay concerning human understanding*, ed. P.H. Nidditch. Oxford: Oxford.
- Loeb, Paul (2013). Eternal Recurrence. In K. Gemes and J. Richardson, eds., *The Oxford Handbook of Nietzsche*. Oxford University Press.
- Lowe, E. J. (2008). *Personal agency*. Oxford University Press.
- Luria, Roy, and Edward K. Vogel (2014). Come together, right now: Dynamic overwriting of an object's history through common fate. *Journal of Cognitive Neuroscience*, 26, 1819-1828.
- Lusthaus, Dan (2002). *Buddhist phenomenology*. Routledge.
- Lycan, William G. (1981). Form, function, and feel. *Journal of Philosophy*, 78, 24-50.
- Lycan, William G. (2001). Moore against the new skeptics. *Philosophical Studies*, 103, 35-53.
- Lycan, William G. (2013). On two main themes in Gutting's *What Philosophers Know*. *Southern Journal of Philosophy*, 51, 112-120.
- Madden, Rory (2015). The naive topology of the conscious subject. *Noûs*, 49, 55-70.
- Maddy, Penelope (2017). *What do philosophers do?* Oxford University Press.
- Maitzen, Stephen (2010). A dilemma for skeptics. *Teorema*, 29, 23-34.
- Malone, Thomas W. (2018). *Superminds*. Little, Brown and Company.
- Mandik, Pete, and Josh Weisberg (2008). Type Q materialism. In *Naturalism, reference, and ontology*, ed. C.B. Wrenn. Peter Lang.
- Marr, David (1982/2010). *Vision*. MIT Press.
- Maudlin, Tim (1994/2002). *Quantum non-locality and relativity*. Blackwell.
- Maudlin, Tim (2019). *Philosophy of physics: Quantum theory*. Princeton University Press.
- Maund, Barry (1997/2019). Color. *Stanford Encyclopedia of Philosophy* (Spring 2019 edition).
- Maynard Smith, John, and Eors Szathmáry (1995). *The major transitions in evolution*. Oxford University Press.
- McCain, Kevin (2012). The predictivist argument against skepticism. *Analysis*, 72, 660-665.
- McCain, Kevin (2014). *Evidentialism and epistemic justification*. Routledge.
- McCauley, Robert N. (2000). The naturalness of religion and the unnaturalness of science. In *Explanation and cognition*, F.C. Keil and R.A. Wilson, eds. Cambridge, MA: MIT.
- McDougall, William (1920). *The group mind*. Putnam.
- McDowell, John (2009). Wittgensteinian "quietism". *Common Knowledge*, 15, 365-372.
- McGinn, Colin (2004). *Mindsight*. Harvard University Press.
- McLaughlin, Brian P. (2017). Type materialism for phenomenal consciousness. In M. Velmans and S. Schneider, eds., *The Blackwell companion to consciousness*, 2nd ed. John Wiley & Sons.
- McMahan, Jeff (2005). Our fellow creatures. *Journal of Ethics*, 9, 353-380.
- McTaggart, J. Ellis (1908). The unreality of time. *Mind*, 17, 457-474.
- Meltzoff, Andrew N., Rechele Brooks, Aaron P. Shon, and Rajesh P. N. Rao (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks* 23, 966-972.

- Merleau-Ponty, Maurice (1945/2012). *The phenomenology of perception*, trans. D. A. Landes. Routledge.
- Merricks, Trenton (2003). Maximality and consciousness. *Philosophy and Phenomenological Research*, 66, 150-158.
- Metzinger, Thomas (2003). *Being no one*. MIT Press.
- Michel, Matthias (2019). Fish and microchips: on fish pain and multiple realization. *Philosophical Studies*, 176, 2411-2428.
- Michel, Matthias (2021). If IIT is true, IIT is false (The Unfolded-Tononi Paradox). Blog post at *Yet Another Blog on Consciousness* (Oct 31).
- Mill, John Stuart (1861/2001). *Utilitarianism*, ed. G. Sher. Hackett.
- Mill, John Stuart (1867). *An examination of Sir William Hamilton's philosophy*. Longmans, Green, Reader, and Dyer.
- Millikan, Ruth G. (1984). *Language, thought, and other biological categories*. MIT Press.
- Millikan, Ruth G. (2010). On knowing the meaning: With a coda on Swampman. *Mind*, 119, 43-81.
- Millikan, Ruth G. (2017). *Beyond concepts*. Oxford University Press.
- Montaigne, Michel de (1580/1595/2003). *The complete essays*, ed. M. A. Screech. Penguin.
- Montero, Barbara (1999). The body problem. *Noûs*, 33, 183-200.
- Moore, G. E. (1922). *Philosophical Studies*. Kegan, Paul, Trench, Trubner.
- Moore, G. E. (1925). A defence of common sense. In *Contemporary British philosophy*, ed. J.H Muirhead. London: George Allen & Unwin.
- Moore, G. E. (1939) 'Proof of an external world'. *Proceedings of the British Academy*, 25, 273-300.
- Moore, G. E. (1953). *Some Main Problems of Philosophy*, London: George Allen & Unwin.
- Moore, G. E. (1959). Certainty. In *Philosophical Papers*. London: Allen and Unwin.
- Moore, G. E. (1957). Visual Sense Data, in *British Philosophy in Mid-Century*, ed. C.A. Mace, London: George Allen & Unwin.
- Moravec, Hans (1997). When will computer hardware match the human brain? Available at <http://www.transhumanist.com/volume1/moravec.htm>.
- Mørch, Hedda H. (2019). Is consciousness intrinsic? A problem for the Integrated Information Theory. *Journal of Consciousness Studies*, 26 (1-2), 133-162.
- Mori, Masahiro (1970/2012). The uncanny valley, trans. K. F. MacDorman and N. Kageki, *IEEE Robotics & Automation Magazine*, 19 (2), 98-100.
- Mullin, Amy (2011). Children and the argument from "marginal cases". *Ethical Theory & Moral Practice*, 14, 291-305.
- Murphy, Nancey (2006). *Bodies and souls, or spirited bodies?* Cambridge: Cambridge.
- Myers-Schulz, Blake, and Eric Schwitzgebel (2013). Knowing that P without believing that P. *Noûs*, 47, 371-384.
- Nagel, Thomas (1974). What is it like to be a bat? *Philosophical Review*, 83, 435-450.
- Nanay, Bence (2011). Do we see apples as edible? *Pacific Philosophical Quarterly*, 92, 305-322.
- Neander, Karen (2016). Swampman meets swampcow. *Mind & Language*, 11, 118-129.
- Neander, Karen (2017). *A mark of the mental*. MIT Press.
- Nemirow, Laurence (1980). Mortal questions. Book review in *Philosophical Review*, 89, 473-477.

- Nemirow, Laurence (1990). Physicalism and the cognitive role of acquaintance. In W. G. Lycan, ed., *Mind and Cognition*. Basil Blackwell.
- Nicholson, Daniel J., and John Dupré, eds. (2018). *Everything flows*. Oxford University Press.
- Nida-Rümelin, Martine and Donnchadh O Conaill (2002/2021). Qualia: The Knowledge Argument. *Stanford Encyclopedia of Philosophy* (Summer 2021 edition).
- Niedermeyer, E. (1999). A concept of consciousness. *Italian Journal of Neurological Sciences*, 20, 7-15.
- Nietzsche, Friedrich (1882/1974). *The gay science*, trans. W. Kaufman. Random House.
- Nietzsche, Friedrich (1901/1968). *The will to power*, trans. W. Kaufman and R. J. Hollingdale. Random House.
- Niikawa, Takuya (2021). Illusionism and definitions of phenomenal consciousness. *Philosophical Studies*, 178, 1-21.
- Nikitin, E.S., T.A. Korshunova, I.S. Zakharov, and P.M. Balaban (2008). Olfactory experience modifies the effect of odour on feeding behaviour in a goal-related manner. *Journal of Comparative Physiology A*, 194, 19-26.
- Noë, Alva (2004). *Action in perception*. MIT Press.
- Nolan, Lawrence, ed. (2011). *Primary and secondary qualities*. Oxford University Press.
- Nolan, Lawrence, and Alan Nelson (2006). Proofs for the existence of God. In S. Gaukroger, ed., *The Blackwell guide to Descartes' Meditations*. Blackwell.
- Norton, John (2002). A paradox in Newtonian gravitation theory II. In J. Meheus, ed., *Inconsistency in science*. Kluwer.
- Norton, John D., and Matthew W. Parker (2021). An infinite lottery paradox. *PhilSci Archive*: <http://philsci-archive.pitt.edu/18923>.
- Nover, Harris, and Alan Hájek (2004). Vexing expectations. *Mind*, 113, 237-249.
- Nye, Howard (2014). Chaos and constraints. In D. Boersema, ed., *Dimensions of moral agency*. Cambridge Scholars Press.
- Oizumi, Masafumi, Larrissa Albantakis, and Giulio Tononi (2014). From the phenomenology of the mechanisms of consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, 10 (5) e1003588.
- Overgaard, Marten (2017). The status and future of consciousness research. *Frontiers in Psychology*, 8 (1719). DOI: 10.3389/fpsyg.2017.01719
- Overgaard, Søren, and Alessandro Salice (2019). Consciousness, belief, and the group mind hypothesis. *Synthese*, 198, 1-25.
- Pacherie, Elisabeth (2017). Collective phenomenology. In K. Ludwig & M. Jankovic, eds, *The Routledge handbook on collective intentionality*. Routledge.
- Papineau, David (2003). Could there be a science of consciousness? *Philosophical Issues*, 13, 205-220.
- Parfit, Derek (1984). *Reasons and persons*. Oxford University Press.
- Pargetter, Robert (1984) 'The scientific inference to other minds'. *Australasian Journal of Philosophy*, 62, 158-163.
- Pascal, Blaise (1670/2005). *Pensées*, trans. R. Ariew. Indianapolis: Hackett.
- Patterson, Sarah (2005). Epiphenomenalism and occasionalism: Problems of mental causation, old and new. *History of Philosophy Quarterly*, 22, 239-257.
- Peacock, John A. (1998). *Cosmological physics*. Cambridge University Press.
- Peacocke, Christopher (1984). Colour concepts and colour experience. *Synthese*, 58, 365-381.
- Peacocke, Christopher (2004). *The realm of reason*. Oxford University Press.

- Pearce, David (pre-2014). *Social media unsorted postings*. Online manuscript. URL: <https://www.hedweb.com/social-media/pre2014.html>.
- Penrose, Roger (1989). *The emperor's new mind*. Oxford University Press.
- Penrose, Roger (2004). *The road to reality*. Knopf.
- Penrose, Roger (2006). Before the Big Bang: An outrageous new perspective and its implications for particle physics. *Proceedings of EPAC 2006*, 2759-2762.
- Perry, John (1979). The problem of the essential indexical. *Noûs*, 13, 3-21.
- Pettit, Philip (2018). Consciousness incorporated. *Journal of Social Philosophy*, 49, 12-37.
- Petty, Richard E., Russell H. Fazio, and Pablo Briñol, eds. (2009). *Attitudes: Insights from the new implicit measures*. Taylor and Francis.
- Phelan, Mark, Adam Arico, and Shaun Nichols (2013). Thinking things and feeling things: On an alleged discontinuity in folk metaphysics of mind. *Phenomenology & the Cognitive Sciences*, 12, 703-725.
- Phillips Ian (2018). The methodological puzzle of phenomenal consciousness. *Philosophical Transactions of the Royal Society B*, 373, 20170347.
- Phillips, Ian, and Jorge Morales (2020). The fundamental problem with no-cognition paradigms. *Trends in Cognitive Sciences*, 24, 165-167.
- Piccinini, Gualtiero (2009). First-person data, publicity, and self-measurement. *Philosophers' Imprint*, 9, 1-16.
- Piccinini, Gualtiero, and Sonya Bahar (2013). Neural computation and the computational theory of cognition. *Cognitive Science*, 34, 453-488.
- Piccinini, Gualtiero, and Andrea Scarantino (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, 37, 1-38.
- Planck Collaboration (2014). Planck 2013 results: XVI cosmological parameters. *Astronomy & Astrophysics*, 571, A16.
- Plato (4th c. BCE/1992). *Republic*, trans. G. M. A. Grube and C. D. C. Reeve. Hackett.
- Polger, Thomas W. and Lawrence A. Shapiro (2016). *The multiple realization book*. Oxford University Press.
- Popper, Karl R. (1935/1959//2002). *The logic of scientific discovery*. Routledge.
- Popper, Karl R. (1983). *Realism and the aim of science*. Hutchinson.
- Popper, Karl R. (1994). *Knowledge and the body-mind problem*. Routledge.
- Price, H. H. (1938). Our evidence for the existence of other minds. *Philosophy*, 13, 425-456.
- Prinz, Jesse J. (2012). *The conscious brain*. Oxford University Press.
- Pritchard, Duncan (2016). *Epistemic angst*. Princeton University Press.
- Prosser, Simon (2011). Affordances and phenomenal character in spatial perception. *Philosophical Review*, 120, 475-513.
- Putnam, Hilary (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill, eds., *Art, mind, and religion*. University of Pittsburgh Press / C. Tinling.
- Putnam, Hilary (1975). *Mind, language and reality*. Cambridge University Press.
- Putnam, Hilary (1981). *Reason, truth and history*. Cambridge: Cambridge.
- Putnam, Hilary (1988). *Representation and reality*. MIT Press.
- Quine, W.V.O. (1966/1976). *Ways of paradox*. Cambridge, MA: Harvard.
- Radford, Colin (1966). Knowledge – by examples. *Analysis*, 27, 1-11.
- Radin, Dean, Leena Michel, Karla Galdamez, Paul Wendland, Robert Rickenbach, and Arnaud Delorme (2012). Consciousness and the double-slit interference pattern: Six experiments. *Physics Essays*, 25, 157-171.

- Raymont, Paul (1999). The know-how response to Jackson's knowledge argument. *Journal of Philosophical Research*, 24, 113-126.
- Reader, Soran (2010). Agency, patiency, and personhood. In T. O'Connor and C. Sandis, eds., *A companion to the philosophy of action*. Wiley.
- Reck, Erich H. (2005). Frege on numbers: Beyond the Platonist picture. *Harvard Review of Philosophy*, 13 (2), 25-40.
- Reichenbach, Hans (1938/2006) *Experience and prediction*. Notre Dame, IN: University of Notre Dame.
- Reid, Thomas 1774-1778 (1995). Materialism, in *Thomas Reid on the Animate Creation*, ed. P. Wood, University Park, PA: Pennsylvania State University.
- Reid, Thomas 1785 (2002). *Essays on the Intellectual Powers of Man*, ed. D.R. Brookes., University Park, PA: Pennsylvania State University.
- Reid, Thomas 1788 (2010). *Essays on the Active Powers of Man*, ed. K Haakonssen and J.A. Harris, University Park, PA: Pennsylvania State University.
- Revonsuo, Antti (1995). Consciousness, dreams and virtual realities. *Philosophical Psychology*, 8, 35-58.
- Richert, Rebekah A., and Paul L. Harris (2008). Dualism revisited: Body vs. mind vs. soul. *Journal of Cognition and Culture*, 8, 99-115.
- Rinard, Susanna (2019). Reasoning one's way out of skepticism. In K. McCain and T. Poston, eds., *The mystery of skepticism*. Brill.
- Robinson, Howard (2016). *From the knowledge argument to mental substance*. Cambridge University Press.
- Robinson, William S. (2019). *Epiphenomenal mind*. Routledge.
- Rockwell, Teed (2005). *Neither brain nor ghost*. MIT Press.
- Roelofs, Luke (2019). *Combining minds*. Oxford University Press.
- Rosen, Melanie G. (2018). How bizarre? A pluralist approach to dream content. *Consciousness and Cognition*, 62, 148-162.
- Rosen, Melanie G. (2019). Dreaming of a stable world: Vision and action in sleep. *Synthese*. <https://doi.org/10.1007/s11229-019-02149-1>.
- Rosenberg, Eugene (1971). *Cell and molecular biology*. Holt, Rinehart, and Winston.
- Rosenthal, David M. (2005). *Consciousness and mind*. Oxford: Oxford.
- Rupert, Robert D. (2005). Minding one's own cognitive system: When is a group of minds a single cognitive unit? *Episteme*, 1, 177-88.
- Russell, Bertrand (1912). *The problems of philosophy*. Oxford: Oxford.
- Russell, Bertrand (1914). *Our knowledge of the external world*. George Allen & Unwin.
- Russell, Bertrand (1921). *The Analysis of Mind*. George Allen & Unwin.
- Russell, Bertrand (1927). *The Analysis of Matter*. Paul, Trench, and Trubner.
- Sahley, Christie, Alan Gelperin, and Jerry W. Rudy (1981). One-trial associative learning modifies food odor preferences of a terrestrial mollusc. *Proceedings of the National Academy of Sciences*, 78, 640-642.
- Sartre, Jean-Paul (1936/1962). *Imagination*, trans. F. Williams. University of Michigan Press.
- Saul, Jennifer (2013). Implicit bias, stereotype threat and women in philosophy. In *Women in philosophy*, ed. F. Jenkins and K. Hutichson.
- Schaffer, Jonathan (2015). What not to multiply without necessity. *Australasian Journal of Philosophy*, 93, 644-664.
- Schäffle, Albert E. F. (1896). *Bau und Leben des sozialen Körpers*, 2nd ed. Laupp'schen.

- Schechter, Elizabeth (2018). *Self-consciousness and “split” brains*. Oxford University Press.
- Schelling, Friedrich W. J. von (1800/1978). *System of transcendental idealism*, trans. P. Heath. University Press of Virginia.
- Schmid, Hans B. (2014). The feeling of being a group: Corporate emotions and collective consciousness. In C. v Scheve and M. Salmela, eds., *Collective emotions*. Oxford University Press.
- Schneider, Susan (2019). *Artificial you*. Princeton University Press.
- Scholl, Brian (2007). Object persistence in philosophy and psychology. *Mind & Language*, 22, 563-591.
- Schwitzgebel, Eric (1996). Zhuangzi’s attitude toward language and his skepticism. In P. Kjellberg and P.J. Ivanhoe, eds., *Essays on skepticism, relativism, and ethics in the Zhuangzi*. Albany, NY: SUNY.
- Schwitzgebel, Eric (2005). Difference tone training: A demonstration adapted from Titchener’s *Experimental Psychology*. *Psyche*, 11 (6). Available at: <https://faculty.ucr.edu/~eschwitz/SchwitzAbs/DiffTone.htm>.
- Schwitzgebel, Eric (2010). Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, 91, 531-553.
- Schwitzgebel, Eric (2011a). Bostrom’s response to my discussion of the simulation argument. Blog post at *The Splintered Mind* (Sep. 2).
- Schwitzgebel, Eric (2011b). *Perplexities of consciousness*. MIT Press.
- Schwitzgebel, Eric (2012a). Group minds on Ringworld. Blog post at *The Splintered Mind* (Oct. 24.)
- Schwitzgebel, Eric (2012b). Introspection, what?. In D. Smithies and D. Stoljar (eds.), *Introspection and consciousness*. Oxford University Press.
- Schwitzgebel, Eric (2012c). Self-ignorance. In J. Liu and J. Perry, eds., *Consciousness and the self*. Cambridge University Press.
- Schwitzgebel, Eric (2012d). Why Dennett should think that the United States is conscious. Blog post at *The Splintered Mind* (Feb. 9).
- Schwitzgebel, Eric (2012e). Why Dretske should think that the United States is conscious. Blog post at *The Splintered Mind* (Feb. 17).
- Schwitzgebel, Eric (2012f). Why Humphrey should think that the United States is conscious. Blog post at *The Splintered Mind* (Mar. 8).
- Schwitzgebel, Eric (2012g). Why Tononi should allow that conscious entities can have conscious parts. Blog post at *The Splintered Mind* (Jun. 6).
- Schwitzgebel, Eric (2012h). Why Tononi should think that the United States is conscious. Blog post at *The Splintered Mind* (Mar. 23).
- Schwitzgebel, Eric (2013a). Hans Reichenbach’s cubical world and Elliott Sober’s beach. Blog post at *The Splintered Mind* (Mar. 19).
- Schwitzgebel, Eric (2013b). Two types of hallucination. Blog post at *The Splintered Mind* (May 6).
- Schwitzgebel, Eric (2014a). Stanislaw Lem’s proof that the external world exists. Blog post at *The Splintered Mind* (Jan. 21).
- Schwitzgebel, Eric (2014b). The crazyist metaphysics of mind. *Australasian Journal of Philosophy*, 92, 665-682.

- Schwitzgebel, Eric (2014c). Tononi's Exclusion Postulate would make consciousness (nearly) irrelevant. Blog post at *The Splintered Mind* (Jul. 16).
- Schwitzgebel, Eric (2015a). Against the "still here" reply to the Boltzmann Brains problem. Blog post at *The Splintered Mind* (Oct. 1).
- Schwitzgebel, Eric (2015b). Duplicating the universe. Blog post at *The Splintered Mind* (Apr. 29).
- Schwitzgebel, Eric (2015c). If Materialism is true, the United States is probably conscious. *Philosophical Studies*, 172, 1697-1721.
- Schwitzgebel, Eric (2015d). Out of the jar. *The Magazine of Fantasy & Science Fiction*, 128, 118-128.
- Schwitzgebel, Eric (2015e). The tyrant's headache. *Sci Phi Journal*, issue #3, 78-83.
- Schwitzgebel, Eric (2016a). Fish Dance. *Clarksworld*, issue #118
- Schwitzgebel, Eric (2016b). Is the United States phenomenally conscious? Reply to Kammerer. *Philosophia*, 44, 877-883.
- Schwitzgebel, Eric (2016c). Phenomenal consciousness, defined and defended as innocently as I can manage. *Journal of Consciousness Studies*, 23 (11-12), 224-235.
- Schwitzgebel, Eric (2017a). 1% skepticism. *Noûs*, 51, 271-290.
- Schwitzgebel, Eric (2017b). THE TURING MACHINES OF BABEL. *Apex*, # 98. URL: <https://www.apex-magazine.com/the-turing-machines-of-babel>
- Schwitzgebel, Eric (2019a). Kant meets cyberpunk. *Disputatio*, 55, 411-435.
- Schwitzgebel, Eric (2019a). The 295 Most-Cited Contemporary Authors in the Stanford Encyclopedia of Philosophy. Blog post at *The Splintered Mind* (Aug. 20).
- Schwitzgebel, Eric (2019b). *A theory of jerks and other philosophical misadventures*. MIT Press.
- Schwitzgebel, Eric (2020a). *Inflate and explode*. Manuscript at <http://faculty.ucr.edu/~eschwitz/SchwitzAbs/InflateExplode.htm>.
- Schwitzgebel, Eric (2020b). The Copernican principle of consciousness. Blog post at *The Splintered Mind* (Sep. 24).
- Schwitzgebel, Eric (2022a). *Borderline consciousness, when it's neither determinately true nor determinately false that experience is present*. Manuscript. URL: <http://faculty.ucr.edu/~eschwitz/SchwitzAbs/BorderlineConsciousness.htm>.
- Schwitzgebel, Eric (2022b). Our infinite predecessors: Flipping the Doomsday Argument on its head. Blog post at *The Splintered Mind* (May 17).
- Schwitzgebel, Eric (forthcoming). How to Wave at Future Versions of Yourself in an Infinite Cosmos. Blog post at *The Splintered Mind*.
- Schwitzgebel, Eric, and Mara Garza (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39, 98-119.
- Schwitzgebel, Eric, and Mara Garza (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In S. M. Liao, ed., *Ethics of artificial intelligence*. Oxford University Press.
- Scott, James C. (1998). *Seeing like a state*. Yale University Press.
- Seager, William E., ed. (2020). *The Routledge handbook of panpsychism*. Routledge.
- Searle, John (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3, 417-457.
- Searle, John (1984). *Minds, brains, and science*. Cambridge, MA: Harvard.
- Searle, John (1992). *The rediscovery of the mind*. MIT Press.
- Searle, John (2004). *Mind*. Oxford University Press.

- Searle, John (2010). *Making the social world*. Oxford University Press.
- Segal, Aaron (2020). Lost at sea: A new route to metaphysical skepticism. *Pacific Philosophical Quarterly*, 101, 256-275.
- Sextus Empiricus (c. 200 CE/1994). *Outlines of skepticism*, trans. J. Annas and J. Barnes. Cambridge: Cambridge.
- Shapiro, J. A. (2007). Bacteria are small but not stupid: Cognition, natural genetic engineering, and socio-bacteriology. *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 38, 807-819.
- Shepard, Joshua (2018). *Consciousness and moral status*. Routledge.
- Shevlin, Henry (2021a). Non-human consciousness and the specificity problem. *Mind & Language*, 36, 297-314.
- Shevlin, Henry (2021b). *Uncanny believers: Chatbots, beliefs, and folk psychology*. Unpublished manuscript.
- Shimony, Abner. Scientific inference. In R.G. Colodny, ed., *The nature and function of scientific theories*. Pittsburgh: University of Pittsburgh.
- Sider, Theodore, (2003). Maximality and microphysical supervenience. *Philosophy and Phenomenological Research*, 66, 139-149.
- Siderits, Mark (2007). *Buddhism as philosophy*. Ashgate.
- Siegel, Susanna (2014). Affordances and the contents of perception. In B. Brogaard, ed., *Does perception have content?* Oxford University Press.
- Siegel, Susanna, and Alex Byrne (2017). Rich or thin? In B. Nanay, ed., *Current controversies in philosophy of perception*. Routledge.
- Siewert, Charles (1998). *The significance of consciousness*. Princeton, NJ: Princeton.
- Silcox, Mark (2019). *A defense of simulated experience*. Routledge.
- Silverman, Allan (2003/2014). Plato's middle period metaphysics and epistemology. *Stanford Encyclopedia of Philosophy* (Fall 2014 edition).
- Simon, Jonathan A. (2017a). The hard problem of the many. *Philosophical Perspectives*, 31, 449-468.
- Simon, Jonathan A. (2017b). Vagueness and zombies: Why 'phenomenally conscious' has no borderline cases. *Philosophical Studies*, 174, 2105-2123.
- Singer, Peter (1975/2009). *Animal liberation, updated ed.* HarperCollins.
- Singer, Peter (1980/2011). *Practical ethics, 3rd ed.* Cambridge University Press.
- Singer, Peter W. (2009). *Wired for war*. Penguin.
- Sinhababu, Neil (2008). Possible girls. *Pacific Philosophical Quarterly*, 89, 254-260.
- Slingerland, Edward, and Maciej Chudek (2011). The prevalence of mind-body dualism in early China. *Cognitive Science*, 35, 997-1007.
- Sliwa, Paulina, and Sophie Horowitz (2015). Respecting *all* the evidence. *Philosophical Studies*, 172, 2835-2858.
- Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review*, 68, 141-156.
- Smart, J. J. C. (2000/2017). The mind/brain identity theory. *Stanford Encyclopedia of Philosophy* (Spring 2017 edition).
- Smith, David L. (2021). *Making monsters*. Harvard University Press.
- Snodgrass, Melinda M., and Robert Scheerer (1989). The measure of a man. *Star Trek: The Next Generation*, season 2, episode 9.
- Sober, Elliott (1975). *Simplicity*. Oxford University Press.

- Sober, Elliott (2011) 'Reichenbach's cubical universe and the problem of the external world'. *Synthese*, 181, 3-21.
- Sober, Elliott, and David Sloan Wilson (1998). *Unto others*. Harvard University Press.
- Sosa, Ernest (2000) 'Skepticism and contextualism'. *Philosophical Issues*, 10, 1-18.
- Sosa, Ernest (2007). *A virtue epistemology*. Oxford: Oxford.
- Spelke, Elizabeth S., Karen Breinlinger, Janet Macomber, and Kristen Jacobson (1992). Origins of knowledge. *Psychological Review*, 99, 605-632.
- Stalnaker, Aaron (2006). *Overcoming our evil*. Washington, DC: Georgetown University.
- Stanford, P. Kyle (2006). *Exceeding our grasp*. Oxford University Press.
- Stang, Nicholas F. (2016). Kant's transcendental idealism. *Stanford Encyclopedia of Philosophy* (Spring 2016 edition).
- Steinhardt, Paul J., and Neil Turok (2002). A cyclic model of the universe. *Science*, 296 (5572), 1436-1439.
- Steinhart, Eric C. (2014). *Your digital afterlives*. Palgrave.
- Stenger, Victor J. (2011). *The fallacy of fine-tuning*. Prometheus.
- Stephenson, Richard, and Vern Lewis (2011). Behavioural evidence for a sleep-like quiescent state in a pulmonate mollusc, *Lymnaea stagnalis* (Linnaeus). *Journal of Experimental Biology*, 214, 747-756
- Stern, Robert (2012) 'Is Hegel's master-slave dialectic a refutation of solipsism?' *British Journal for the History of Philosophy*, 20, 333-361.
- Stich, Stephen (1983). *From folk psychology to cognitive science*. MIT Press.
- Stich, Stephen (2009). Five answers. In *Mind and consciousness*, ed. S. Grim. Automatic Press.
- Stock, Gregory (1993). *Metaman*. Doubleday Canada.
- Stoljar, Daniel (2010). *Physicalism*. Routledge.
- Stratton, George M. (1896). Some preliminary experiences on vision without inversion of the retinal image. *Psychological Review*, 3, 611-617.
- Stratton, George M. (1897a). Upright vision and the retinal image. *Psychological Review*, 3, 182-187.
- Stratton, George M. (1897b). Vision without inversion of the retinal image. *Psychological Review*, 4, 341-360.
- Stratton, George M. (1897c). Vision without inversion of the retinal image [concluded]. *Psychological Review*, 4, 463-481.
- Strawson, Galen (2006). *Consciousness and its place in nature*. Imprint Academic.
- Strawson, Galen (2012). Real Naturalism, *Proceedings of the American Philosophical Association* 86/2: 125-154.
- Strawson, P. F. (1959). *Individuals*. Methuen.
- Strawson, P. F. (1985). *Skepticism and Naturalism*. Columbia University Press.
- Stringer, Ian Alexander Noel, Glenn Richard Parrish, and Gregory Howard Sherley (2018). Homing, dispersal and mortality after translocation of long-lived land snails *Placostylus ambagiosus* and *P. hongii* (Gastropoda: Bothriembryontidae) in New Zealand. *Molluscan Research*, 38, 56-76,
- Stroud, Barry (1984). *The significance of philosophical scepticism*. Oxford: Oxford.
- Stroud, Barry (1994). Kantian arguments, conceptual capacities and invulnerability. In P. Parrini, ed., *Kant and contemporary epistemology*. Kluwer.
- Sussman, David (2003). The authority of humanity. *Ethics*, 113, 350-366.
- Šuster, Danilo (2016). Dreams in a vat. *European Journal of Analytic Philosophy*, 12, 89-106.

- Sutton, C. S. (2014). The supervenience solution to the too-many-thinkers problem. *Philosophical Quarterly*, 64, 619-639.
- Sytsma, Justin M., and Edouard Machery (2010). Two conceptions of subjective experience. *Philosophical Studies*, 151, 299-327.
- Talbott, William (2001/2016). Bayesian epistemology. *Stanford Encyclopedia of Philosophy* (Winter 2016 edition).
- Tarrow, Sidney G. (1994/2011). *Power in movement*, 3rd ed. Cambridge University Press.
- Taylor, James G. (1962). *The behavioral basis of perception*. New Haven, CT: Yale.
- Tegmark, Max (2009). The multiverse hierarchy. In B. Carr, ed., *Universe or multiverse?* Cambridge University Press.
- Tegmark, Max (2014). *Our mathematical universe*. Random House.
- Teilhard de Chardin, Pierre (1955/1965). *The phenomenon of man*, rev. English ed., trans. B. Wall. Harper & Row.
- Thomas, Nigel J. T. (1999). Are theory of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive Science*, 23, 207-245.
- Thomas, Nigel J. T. (2014). The multidimensional spectrum of imagination: Images, dreams, hallucinations, and active, imaginative perception. *Humanities*, 3 (2), 132-184.
- Thompson, Brad (2010). The spatial content of experience. *Philosophy and Phenomenological Research*, 81, 146-184.
- Thompson, Evan (2015). *Dreaming, waking, being*. Columbia University Press.
- Thornton, Stephen P. (2004). Solipsism and the problem of other minds. *Internet Encyclopedia of Philosophy*: www.iep.utm.edu/solipsis.
- Titchener, Edward B. (1915). *A text-book of psychology*. Macmillan.
- Tomiyama, Kiyonori (1992). Homing behaviour of the giant African snail, *Achatina fulica* (Ferussac) (Gastropoda; Pulmonata). *Journal of Ethology*, 10, 139-146.
- Tononi, Giulio (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5: 42.
- Tononi, Giulio (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215, 216-242.
- Tononi, Giulio (2010). Information integration: Its relevance to brain function and consciousness. *Archives Italiennes de Biologie*, 148, 299-322.
- Tononi, Giulio (2012). The integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie*, 150, 290-326.
- Tononi, Giulio, and Christof Koch (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B* 370: 20140167.
- Townsend, James T., and Michael J. Wenger (2004). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review*, 11, 391-418.
- Trewavas, Anthony (2014). *Plant behavior and intelligence*. Oxford University Press.
- Tsuchiya, Naotsugu, Stefan Frässle, Melanie Wilke, and Victor Lamme. (2016). No-report and report-based paradigms jointly unravel the NCC: response to Overgaard and Fazekas. *Trends in Cognitive Sciences*, 20, 242-243.
- Tuomela, Raimo (2007). *The philosophy of sociality*. Oxford University Press.
- Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2*, 42, 230-265.
- Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.

- Tye, Michael (1995). *Ten Problems of Consciousness*. MIT Press.
- Tye, Michael (2000). *Consciousness, color, and content*. MIT Press.
- Tye, Michael (2009). *Consciousness revisited*. MIT Press.
- Tye, Michael (2017). *Tense bees and shell-shocked crabs*. Oxford University Press.
- Tye, Michael (2021). *Vagueness and the evolution of consciousness*. Oxford University Press.
- Udell, David B., and Eric Schwitzgebel (forthcoming). Susan Schneider's proposed tests for AI consciousness: Promising but flawed. *Journal of Consciousness Studies*.
- Unger, Peter (1999). The mental problems of the many. In D. Zimmerman, ed., *Oxford Studies in Metaphysics, Vol. 1*.
- United Nations (1948). *Universal declaration of human rights*. Available at <https://www.un.org/sites/un2.un.org/files/udhr.pdf>.
- Valberg, J. J. (2007). *Dream, death, and self*. Princeton: Princeton.
- Van Cleve, James (1995). Putnam, Kant, and secondary qualities. *Philosophical Papers*, 24, 83-109.
- Van der Deijl, Willem (2021). The sentience argument for experientialism about welfare. *Philosophical Studies*, 178, 187-208.
- Van Inwagen, Peter (1996). "It is wrong, everywhere, always, and for anyone, to believe anything upon insufficient evidence". In J. Jordan and D. Howard-Snyder, eds., *Faith, freedom, and rationality*. Rowman & Littlefield.
- Vardanyan, M., R. Trotta, and J. Silk (2011). Applications of Bayesian model averaging to the curvature and size of the Universe. *Monthly Notices of the Royal Astronomical Society: Letters*, 413, L91–L95.
- Vilenkin, Alex (2006). *Many worlds in one*. Hill and Wang.
- Vinge, Vernor (1992). *A fire upon the deep*. Tor.
- Vinge, Vernor (2011). *Children of the sky*. Tor.
- Vogel, Jonathan (1990). Cartesian skepticism and inference to the best explanation. *Journal of Philosophy*, 87, 658-666.
- Vogel, Jonathan (1999). The new Relevant Alternatives Theory. *Philosophical Perspectives*, 13, 155-180.
- Vogel, Jonathan (2005). The refutation of skepticism. In M. Steup and E. Sosa, eds., *Contemporary debates in epistemology*. Blackwell.
- Vogel, Jonathan (2008). Internalist responses to skepticism. In J. Greco, ed., *The Oxford handbook of skepticism*. Oxford: Oxford.
- Vold, Karina (2015). The parity argument for extended consciousness. *Journal of Consciousness Studies*, 22 (3-4), 16-33.
- Wagner, Stephen I. (2014). *Squaring the circle in Descartes' Meditations*. Cambridge University Press.
- Wallace, David (2008). Philosophy of quantum mechanics. In D. Rickles, ed., *The Ashgate companion to contemporary philosophy of physics*. Ashgate.
- Wallace, David (2012). *The emergent multiverse*. Oxford University Press.
- Walleczek, Jan, and Nikolaus von Stillfried (2019). False-positive effect in the Radin double-slit experiment on observer consciousness as determined with the advanced meta-experimental protocol. *Frontiers in Psychology*, 10, #1891.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273-281.

- Wasserman, David, Adrienne Asch, Jeffrey Blustein, and Daniel Putnam (2012/2017). Cognitive disability and moral status. *Stanford Encyclopedia of Philosophy* (Fall 2017 edition).
- Weatherson, Brian (2003). Are you a sim? *Philosophical Quarterly*, 53, 425-431.
- Weinberg, Jonathan M., Chad Gonnerman, Cameron Buckner, and Joshua Alexander (2010). Are philosophers expert intuiters? *Philosophical Psychology*, 23, 331-355.
- Weinberg, Justin, ed., (2020). Philosophers on GPT-3 (updated with replies by GPT-3). Blog post at *Daily Nous* (Jul. 30).
- Wenmackers, Sylvia, and Jan-Willem Romeijn (2016). New theory about old evidence: Humbly catching them all. *Synthese*, 193, 1225–1250
- Whiteley, Cecily M. K. (2021). Aphantasia, imagination and dreaming. *Philosophical Studies*, 178, 2111–2132.
- Wierwille, Walter W., W.A. Schaudt, S. Gupta, J.M. Spaulding, D.S. Bowman, G.M. Fitch, D M. Wiegand, and R.J. Hanowski (2008). *Study of driver performance/acceptance using aspheric mirrors in light vehicle applications*. United States Department of Transportation report DOT HS 810 959.
- Wigner, Eugene P. (1961). Remarks on the mind-body question. In I.J. Good, A.J. Mayne, and J.M. Smith (eds.), *The Scientist Speculates*. Heinemann.
- Williams, Bernard (1973). *Problems of the self*. Cambridge: Cambridge.
- Williams, Michael (1991) *Unnatural doubts*. Cambridge, MA: Blackwell.
- Wilson, Robert A. (2004). *Boundaries of the mind*. Cambridge University Press.
- Wilson, Robert A. (2005). *Genes and the agents of life*. Cambridge University Press.
- Windt, Jennifer M. (2015). *Dreaming*. MIT Press.
- Windt, Jennifer M. (2017). Predictive brains, dreaming selves, sleeping bodies: How the analysis of dream movement can inform a theory of self- and world-simulation in dreams. *Synthese* 195, 2577–2625.
- Windt, Jennifer M., Tore Nielsen, and Evan Thompson (2016). Does consciousness disappear in dreamless sleep? *Trends in Cognitive Sciences*, 20, 871-882.
- Wittenbrink, Bernd, and Norbert Schwarz, eds. (2007). *Implicit measures of attitudes*. Guilford.
- Wittgenstein, Ludwig (1945-1949/1958). *Philosophical Investigations*, trans. G. E. M. Anscombe. Macmillan.
- Wittgenstein, Ludwig (1947/1980). *Remarks on the Philosophy of Psychology, vol. 1*, trans. G. E. M. Anscombe. University of Chicago Press.
- Wittgenstein, Ludwig (1951/1969). *On certainty*, ed. G. E. M. Anscombe and G. H. von Wright. Harper.
- Wundt, Wilhelm (1897/1897). *Outlines of psychology*, trans. C. H. Judd. Wilhelm Engelmann.
- Yablo, Stephen (1987). Identity, essence, and indiscernibility. *Journal of Philosophy*, 84, 293-314.
- Yudkowsky, Eliezer S. (2002). *The AI-box experiment*. Online manuscript. URL: <https://www.yudkowsky.net/singularity/aibox>.
- Zellner, Arnold, Hugo A. Zeuzenkamp, and Michael McAleer, eds. (2001). *Simplicity, inference, and modeling*. Cambridge: Cambridge.
- Zieger, Marina V. and Victor Benno Meyer-Rochow (2008). Understanding the cephalic eyes of pulmonate gastropods: A review. *American Malacological Bulletin*, 26 (1-2), 47-6.
- Zuckerman, Phil (2007). Atheism: Contemporary numbers and patterns. In M. Martin, ed., *The Cambridge companion to atheism*. Cambridge University Press.