



# REALITY +

virtual worlds  
and the  
problems of philosophy

DAVID J.  
CHALMERS

# REALITY+

---

VIRTUAL WORLDS  
AND THE PROBLEMS  
OF PHILOSOPHY

---

**David J. Chalmers**

Illustrations by Tim Peacock



**W. W. NORTON & COMPANY**  
*Independent Publishers Since 1923*

For Claudia

# Contents

---

*Introduction* Adventures in technophilosophy

## **Part 1 VIRTUAL WORLDS**

*Chapter 1* Is this the real life?

*Chapter 2* What is the simulation hypothesis?

## **Part 2 KNOWLEDGE**

*Chapter 3* Do we know things?

*Chapter 4* Can we prove there is an external world?

*Chapter 5* Is it likely that we're in a simulation?

## **Part 3 REALITY**

*Chapter 6* What is reality?

*Chapter 7* Is God a hacker in the next universe up?

*Chapter 8* Is the universe made of information?

*Chapter 9* Did simulation create its from bits?

## **Part 4 REAL VIRTUAL REALITY**

*Chapter 10* Do virtual reality headsets create reality?

*Chapter 11* Are virtual reality devices illusion machines?

*Chapter 12* Does augmented reality lead to alternative facts?

*Chapter 13* Can we avoid being deceived by deepfakes?

## **Part 5 MIND**

*Chapter 14* How do mind and body interact in a virtual world?

*Chapter 15* Can there be consciousness in a digital world?

*Chapter 16* Does augmented reality extend the mind?

## **Part 6 VALUE**

*Chapter 17* Can you lead a good life in a virtual world?

*Chapter 18* Do simulated lives matter?

*Chapter 19* How should we build a virtual society?

## **Part 7 FOUNDATIONS**

*Chapter 20* What do our words mean in virtual worlds?

*Chapter 21* Do dust clouds run computer programs?

*Chapter 22* Is reality a mathematical structure?

*Chapter 23* Have we fallen from the Garden of Eden?

*Chapter 24* Are we Boltzmann brains in a dream world?

*Acknowledgments*

*Glossary*

*Notes*

*Index*

## Introduction

# Adventures in technophilosophy

**W**HEN I WAS TEN YEARS OLD, I DISCOVERED COMPUTERS. MY first machine was a PDP-10 mainframe system at the medical center where my father worked. I taught myself to write simple programs in the BASIC computer language. Like any ten-year-old, I was especially pleased to discover games on the computer. One game was simply labeled “ADVENT.” I opened it and saw:

*You are standing at the end of a road before a small brick building.*

*Around you is a forest.*

*A small stream flows out of the building and down a gully.*

I figured out that I could move around with commands like “go north” and “go south.” I entered the building and got food, water, keys, a lamp. I wandered outside and descended through a grate into a system of underground caves. Soon I was battling snakes, gathering treasures, and throwing axes at pesky attackers. The game used text only, no graphics, but it was easy to imagine the cave system stretching out below ground. I played for months, roaming farther and deeper, gradually mapping out the world.

It was 1976. The game was *Colossal Cave Adventure*. It was my first virtual world.

In the years that followed, I discovered video games. I started with *Pong* and *Breakout*. When *Space Invaders* came to our local shopping mall,

it became an obsession for my brothers and me. Eventually I got an Apple II computer, and we could play *Asteroids* and *Pac-Man* endlessly at home.

Over the years, virtual worlds have become richer. In the 1990s, games such as *Doom* and *Quake* pioneered the use of a first-person perspective. In the 2000s, people began spending vast amounts of time in multiplayer virtual worlds like *Second Life* and *World of Warcraft*. In the 2010s, there arrived the first rumblings of consumer-level virtual reality headsets, like the Oculus Rift. That decade also saw the first widespread use of augmented reality environments, which populate the physical world with virtual objects in games like *Pokémon Go*.

These days, I have numerous virtual reality systems in my study, including an Oculus Quest 2 and an HTC Vive. I put on a headset, open an application, and suddenly I'm in a virtual world. The physical world has disappeared entirely, replaced by a computer-generated environment. Virtual objects surround me, and I can move among them and manipulate them.

Like ordinary video games from *Pong* to *Fortnite*, virtual reality (or VR) involves a virtual world: an interactive, computer-generated space. What's distinctive about VR is that its virtual worlds are *immersive*. Instead of showing you a two-dimensional screen, VR immerses you in a three-dimensional world you can see and hear as if you existed within it. Virtual reality involves an immersive, interactive, computer-generated space.

I've had all sorts of interesting experiences in VR. I've assumed a female body. I've fought off assassins. I've flown like a bird. I've traveled to Mars. I've looked at a human brain from the inside, with neurons all around me. I've stood on a plank stretched over a canyon—terrified, though I knew perfectly well that if I were to step off, I'd step onto a nonvirtual floor just below the plank.

Like many other people, during the recent pandemic I've spent a great deal of time talking to friends, family, and colleagues using Zoom and other videoconferencing software. Zoom is convenient, but it has many limitations. Eye contact is difficult. Group interactions are choppy rather than cohesive. There is no sense that we are inhabiting a common space. One underlying issue is that videoconferencing is not virtual reality. It is interactive but not immersive, and there is no common virtual world.

During the pandemic, I've also met up once a week with a merry band of fellow philosophers in VR. We've tried many different platforms and

activities—flying with angel wings in *Altspace*, slicing cubes to a rhythm in *Beat Saber*, talking philosophy on the balcony in *Bigscreen*, playing paintball in *Rec Room*, giving lectures in *Spatial*, trying out colorful avatars in *VRChat*. VR technology is still far from perfect, but we’ve had the sense of inhabiting a common world. When five of us were standing around after a short presentation, someone said, “This is just like coffee break at a philosophy conference.” When the next pandemic arrives in a decade or two, it’s likely that many people will hang out in immersive virtual worlds designed for social interaction.

Augmented reality (or AR) systems are also progressing fast. These systems offer a world that is partly virtual and partly physical. The ordinary physical world is augmented by virtual objects. I don’t yet have my own augmented reality glasses, but companies like Apple, Facebook, and Google are said to be working on them. Augmented reality systems have the potential to replace screen-based computing entirely, or at least replace physical screens with virtual screens. Interacting with virtual objects may become part of everyday life.

Today’s VR and AR systems are primitive. The headsets and glasses are bulky. The visual resolution for virtual objects is grainy. Virtual environments offer immersive vision and sound, but you can’t touch a virtual surface, smell a virtual flower, or taste a virtual glass of wine when you drink it.

These temporary limitations will pass. The physics engines that underpin VR are improving. In years to come, the headsets will get smaller, and we will transition to glasses, contact lenses, and eventually retinal or brain implants. The resolution will get better, until a virtual world looks exactly like a nonvirtual world. We will figure out how to handle touch, smell, and taste. We may spend much of our lives in these environments, whether for work, socializing, or entertainment.

My guess is that within a century we will have virtual realities that are indistinguishable from the nonvirtual world. Perhaps we’ll plug into machines through a brain-computer interface, bypassing our eyes and ears and other sense organs. The machines will contain an extremely detailed simulation of a physical reality, simulating laws of physics to track how every object within that reality behaves.

Sometimes VR will place us in other versions of ordinary physical reality. Sometimes it will immerse us in worlds entirely new. People will



enter some worlds temporarily for work or for pleasure. Perhaps Apple will have its own workplace world, with special protections so that no one can leak its latest Reality system under development. NASA will set up a world with spaceships in which people can explore the galaxy at faster-than-light speed. Other worlds will be worlds in which people can live indefinitely. Virtual real estate developers will compete to offer worlds with perfect weather near the beach, or with glorious apartments in a vibrant city, depending on what customers want.

Perhaps, as in the novel and movie *Ready Player One*, our planet will be crowded and degraded, and virtual worlds will provide us with new landscapes and new possibilities. In centuries past, families often faced a decision: “Should we emigrate to a new country to start a new life?” In centuries to come, we may face an equivalent decision: “Should we move our lives to a virtual world?” As with emigration, the reasonable answer may often be yes.

Once simulation technology is good enough, these simulated environments may even be occupied by simulated people, with simulated brains and bodies, who will undergo the whole process of birth, development, aging, and death. Like the nonplayer characters that one encounters in many video games, simulated people will be creatures of the simulation. Some worlds will be simulations set up for research or to make predictions about the future. For instance, a dating app (as seen on the TV series *Black Mirror*) could simulate many futures for a couple in order to see whether they are compatible. A historian might study what would have happened if Hitler had chosen not to start a war with the Soviet Union. Scientists might simulate whole universes from the Big Bang onward, with small variations to study the range of outcomes: How often does life develop? How often is there intelligence? How often is there a galactic civilization?

One can imagine that a few curious 23rd-century simulators might focus on the early 21st century. Let’s suppose the simulators live in a world in which Hillary Clinton defeated Jeb Bush in the US presidential election of 2016. They might ask: How would history have been different if Clinton had lost? Varying a few parameters, the simulators might go so far as to simulate a world where the 2016 victor was Donald Trump. They might even simulate Brexit and a pandemic.

Simulators interested in the history of simulation might also be interested in the 21st century as a period when simulation technology was coming into its own. Perhaps they might occasionally simulate people who are writing books about possible future simulations, or people who are reading them! Narcissistic simulators might nudge the parameters so that some simulated 21st-century philosophers speculate wildly about simulations built in the 23rd century. They might be especially interested in simulating the reactions of 21st-century readers reading thoughts about 23rd-century simulators, as you are right now.

Someone in such a virtual world would believe themselves to be living in an ordinary world in the early 21st century—a world in which Trump was elected president, the UK left the European Union, and there was a pandemic. Those events may have been surprising at the time, but humans have a remarkable capacity to adjust, and after a while these things become normal. Although simulators may have nudged them into reading a book on virtual worlds, it will seem to them as if they are reading the book out of their own free choice. The book they're reading now is perhaps a little unsubtle in trying to convey the message that they may be in a virtual world, but they will take this in stride and start thinking about the idea all the same.

At this point, we can ask, “How do you know you're not in a computer simulation right now?”



This idea is often known as the *simulation hypothesis*. It is famously depicted in the *Matrix* movies, in which what seems an ordinary physical world turns out to be the result of connecting human brains to a giant bank of computers. Inhabitants of the Matrix experience their world very much as we do, but the Matrix is a virtual world.

Could you be in a virtual world right now? Stop and think about this question for a moment. When you do, you're doing philosophy.

*Philosophy* translates as *love of wisdom*, but I like to think of it as *the foundations of everything*. Philosophers are like the little kid who keeps asking, *Why?* or *What is that?* or *How do you know?* or *What does that mean?* or *Why should I do that?* Ask those questions a few times in a row

and you rapidly reach the foundations. You're examining the assumptions that underlie things we take for granted.

I was that kid. It took me a while to realize that what I was interested in was philosophy. I started off studying mathematics, physics, and computer science. These take you a fair distance into the foundations of everything, but I wanted to go deeper. I turned to studying philosophy, along with cognitive science to keep an anchor in the solid ground of science while I explored the foundations underneath.

I was first drawn to address questions about the mind, like *What is consciousness?* I've spent much of my career focusing on those questions. But questions about the world, like *What is reality?*, are just as central to philosophy. Perhaps most central of all are questions about the relation between mind and world, such as *How can we know about reality?*

This last question was at the heart of the challenge posed by René Descartes in his *Meditations on First Philosophy* (1641), which set the agenda for centuries of Western philosophy to come. Descartes posed what I'll call the problem of the external world: How do you know anything at all about the reality outside you?

Descartes approached the problem by asking: How do you know that your perception of the world is not an illusion? How do you know that you are not dreaming right now? How do you know you're not being deceived by an evil demon into thinking all this is real, when it's not? These days, he might approach the problem by asking the question I just asked you: How do you know you're not in a virtual world?

For a long time I thought I didn't have much to say about Descartes's problem of the external world. Thinking about virtual reality gave me a new perspective. It was reflecting on the simulation hypothesis that led me to realize that I had underestimated virtual worlds. In their own way, so had Descartes and many others. I concluded that if we think more clearly about virtual worlds, this might lead us to the beginnings of a solution to Descartes's problem.



The central thesis of this book is: *Virtual reality is genuine reality*. Or at least, *virtual realities are genuine realities*. Virtual worlds need not be second-class realities. They can be first-class realities.

We can break down this thesis into three parts:

- Virtual worlds are not illusions or fictions, or at least they need not be. What happens in VR really happens. The objects we interact with in VR are real.
- Life in virtual worlds can be as good, in principle, as life outside virtual worlds. You can lead a fully meaningful life in a virtual world.
- The world we're living in could be a virtual world. I'm not saying it is. But it's a possibility we can't rule out.

The thesis—especially the first two parts—has practical consequences for the role of VR technology in our lives. In principle, VR can be much more than escapism. It can be a full-blooded environment for living a genuine life.

I'm not saying that virtual worlds will be some sort of utopia. Like the internet, VR technology will almost certainly lead to awful things as well as wonderful things. It's certain to be abused. Physical reality is abused, too. Like physical reality, virtual reality has room for the full range of the human condition—the good, the bad, and the ugly.

I'll focus more on VR in principle than VR in practice. In practice, the road to full-scale virtual reality is sure to be bumpy. It won't surprise me if widespread adoption of VR is limited for a decade or two, while the technology matures. No doubt it will move in all sorts of directions I haven't anticipated. But once a mature VR technology is developed, it should be able to support lives that are on a par with or even surpass life in physical reality.



The title of this book captures my main claims. You can understand it in a number of ways. Each virtual world is a new reality: Reality+. Augmented reality involves additions to reality: Reality+. Some virtual worlds are as good as or better than ordinary reality: Reality+. If we're in a simulation, there is more to reality than we thought: Reality+. There will be a smorgasbord of multiple realities: Reality+.

I know that what I'm saying is counterintuitive to many people. Perhaps you think that VR is Reality−, or Reality Minus. Virtual worlds are fake realities, not genuine realities. No virtual world is as good as ordinary reality. Over the course of this book, I'll try to convince you that Reality+ is closer to the truth.



This book is a project in what I call *technophilosophy*. Technophilosophy is a combination of (1) asking philosophical questions about technology and (2) using technology to help answer traditional philosophical questions.

The name is inspired by what the Canadian-American philosopher Patricia Churchland called *neurophilosophy* in her landmark 1987 book of the same title. Neurophilosophy combines asking philosophical questions about neuroscience with using neuroscience to help answer traditional questions in philosophy. Technophilosophy does the same with technology.

There's a thriving area, often called the philosophy of technology, that carries out the first project—asking philosophical questions about technology. What's especially distinctive about technophilosophy is the second project—using technology to answer traditional philosophical questions. The key to technophilosophy is a two-way interaction between philosophy and technology. Philosophy helps to shed light on (mostly new) questions about technology. Technology helps to shed light on (mostly old) questions about philosophy. I wrote this book in order to shed light on both sorts of question at once.



First, I want to use technology to address some of the oldest questions in philosophy, especially the problem of the external world. At a minimum, virtual reality technology helps *illustrate* Descartes's problem—that is, how can we know anything about the reality around us? How do we know that reality is not an illusion? In [chapters 2 and 3](#), I lay out these problems by introducing the simulation hypothesis and asking, “How do we know we're not in a simulation right now?”

The simulation idea does more than illustrate the problem, however. It also *sharpens* the problem by turning Descartes's far-fetched scenarios

involving evil demons into more realistic scenarios involving computers—scenarios we have to take seriously. In [chapter 4](#), I make the case that the simulation idea undercuts many common responses to Descartes. In [chapter 5](#), I use statistical reasoning about simulations to argue that we cannot know we're not in a simulation. All this makes Descartes's problem even harder.

Most importantly, reflection on virtual reality technology can help us *respond* to the problem of the external world. In [chapters 6](#) through [9](#), I argue that if indeed we're in a simulation, tables and chairs are not illusions but perfectly real objects: they are digital objects that are made of bits. This leads us to what is sometimes called, in modern physics, the *it-from-bit hypothesis*: Physical objects are real and they are digital. Thinking about the simulation hypothesis and the it-from-bit hypothesis—two ideas inspired by modern computers—yields the beginnings of a response to Descartes's classic problem.

We can put Descartes's argument as follows: We don't know that we're not in a virtual world, and in a virtual world nothing is real, so we don't know that anything is real. This argument turns on the assumption that virtual worlds are not genuine realities. Once we make the case that virtual worlds are indeed genuine realities—and especially that objects in a virtual world are real—we can respond to Descartes's argument.

I shouldn't overstate the case. My analysis doesn't address everything Descartes says, and it doesn't prove that we know a great deal about the external world. Still, if the analysis works, it dissolves what is perhaps the Western tradition's prime reason for doubting that we can know anything about the external world. So it gives us at least a foothold in establishing that we have knowledge of the reality around us.

We'll also use technology to illuminate traditional questions about the mind: How do mind and body interact? (See [chapter 14](#).) What is consciousness? (See [chapter 15](#).) Does the mind extend beyond the body? (See [chapter 16](#).) In each case, thinking about a technology—VR, artificial intelligence (AI), and augmented reality (AR), respectively—can illuminate those questions. And conversely, thinking about the questions can illuminate these technologies.

It's worth saying that my views about consciousness and the mind are not the main focus of this book. I've explored those issues in other work, and this book is independent of them to a large degree. I hope that even people who disagree with me about consciousness may find my picture of

reality appealing. That said, there are many connections between the two domains. You can think of [chapters 15 and 16](#), in particular, as adding a fourth plank to the thesis that virtual reality is genuine reality: namely, *virtual and augmented minds are genuine minds*.

Technology can also illuminate traditional questions about value and ethics. Value is the domain of good and bad, better and worse. Ethics is the domain of right and wrong. What makes for a good life? (See [chapter 17](#).) What is the difference between right and wrong? (See [chapter 18](#).) How should society be organized? (See [chapter 19](#).) I'm by no means an expert on these issues, but technology provides at least an interesting angle on them.

Other time-hallowed philosophical questions will come up along the way. Is there a God? (See [chapter 7](#).) What is the universe made of? (See [chapter 8](#).) How does language describe reality? (See [chapter 20](#).) What does science tell us about reality? (See [chapters 22 and 23](#).) It turns out that to make our case that virtual reality is genuine reality, we have to think hard about those old questions. As always, the illumination flows both ways; thinking about technology throws light on the old questions in turn.



I also want to use philosophy to address new questions about technology, especially the technology of virtual worlds. These include questions about everything from video games through augmented reality glasses and virtual reality headsets to simulations of entire universes.

I've already outlined my central thesis that virtual reality is genuine reality. Where VR is concerned, I'll ask questions like: Is virtual reality an illusion? (See [chapters 6, 10, and 11](#).) What are virtual objects? (See [chapter 10](#).) Does augmented reality genuinely augment reality? (See [chapter 12](#).) Can you live a good life in VR? (See [chapter 17](#).) How should you behave in a virtual world? (See [chapter 19](#).)

I'll also discuss other technologies: artificial intelligence, smartphones, the internet, deepfakes, and computers in general. How can we know we're not being deceived by deepfakes? (See [chapter 13](#).) Can AI systems be conscious? (See [chapter 15](#).) Do smartphones extend our minds, and is the internet making us smart or stupid? (See [chapter 16](#).) And what is a computer, anyway? (See [chapter 21](#).)

These questions are all philosophical questions. Many of them are also intensely practical questions. We need to make decisions right now about how we use video games, smartphones, and the internet. An increasing number of such practical questions will confront us in decades to come. As we spend more and more time in virtual worlds, we'll have to grapple with the issue of whether life there is fully meaningful. Eventually, we may have to decide whether or not to upload ourselves to the cloud entirely. Thinking philosophically can help us get clear on these decisions about how to live our lives.



By the end of this book, you'll have been introduced to many of the central questions in philosophy. We'll encounter both historical greats from centuries and millennia past and contemporary figures and arguments from recent decades. We'll cover many of the central topics in philosophy: knowledge, reality, mind, language, value, ethics, science, religion, and more. I'll introduce some of the powerful tools that philosophers have developed over the centuries for thinking about these issues. This is only one perspective, and a great deal of important philosophy is left out. But by the end, you'll have a sense of some of the historical and contemporary landscape of philosophy.

To help readers think through these ideas, I've made connections to science fiction and other corners of popular culture whenever I can. Many authors of science fiction have delved into these issues just as deeply as philosophers have. I've often had new philosophical ideas by thinking about science fiction. Sometimes I think science fiction gets these issues right, and sometimes it gets them wrong. Either way, science-fiction scenarios can prompt a lot of fruitful philosophical analysis.

The best way I know to introduce philosophy is to *do* philosophy. So while I'll start many chapters by posing a philosophical question connected to virtual worlds and introducing some philosophical background, I'll usually get down quickly to thinking hard about the issues. I'll analyze the issues both inside and outside virtual worlds, with an eye on building my argument for the Reality+ point of view.

As a result, this book is as full of my own philosophical theses and arguments as anything I've ever written. While some chapters of the book



go over ground I've discussed in academic articles, well over half of it is entirely new. So even if you're an old hand at philosophy, I hope that you'll find rewards here. In an online supplement ([consc.net/reality](http://consc.net/reality)), I've included extensive notes and appendices pursuing the issues in more depth, often including connections to the academic literature.



The book has seven parts. [Part 1 \(chapters 1 and 2\)](#) introduces the central problems of the book and the simulation hypothesis that plays a central role. [Part 2 \(chapters 3–5\)](#) focuses on questions about knowledge, and especially Descartes's arguments for skepticism about the external world. [Part 3 \(chapters 6–9\)](#) focuses on questions about reality, and makes an initial case for my thesis that virtual reality is genuine reality.

The next three parts of the book develop many different aspects of the thesis. [Part 4 \(chapters 10–13\)](#) brings things down to earth to focus on questions about real virtual reality technology: virtual reality headsets, augmented reality glasses, and deepfakes. [Part 5 \(chapters 14–16\)](#) focuses on questions about the mind. [Part 6 \(chapters 17–19\)](#) focuses on questions about value and ethics. Finally, [part 7 \(chapters 20–24\)](#) focuses on foundational issues about language, computers, and science that are required to fully develop the Reality+ vision. The last chapter pulls the pieces together to see where things stand with Descartes's problem of the external world.

Different readers may want to read the book in different ways. Everyone should read [chapter 1](#), but after that you can strike out in many different directions. In the endnotes, I give some possible paths, depending on your interests. Many chapters stand relatively independently. [Chapters 2, 3, 6, and 10](#) may be especially helpful in providing background for the chapters that follow, but they aren't absolutely essential.

Most of the chapters are frontloaded with introductory material toward the start. The discussion sometimes gets denser toward the end of each chapter, and toward the end of the book. If you're after a shorter book and a lighter reading experience, you might try reading the first two or three sections of every chapter, and then skipping to the next chapter whenever you like.



We live in an age in which truth and reality have been under attack. We're sometimes said to be in an era of post-truth politics in which truth is irrelevant. It's common to hear that there's no absolute truth and no objective reality. Some people think that reality is all in the mind, so that what's real is entirely up to us. The multiple realities of this book may initially suggest a view like that on which truth and reality are cheap. This is not my view.

Here's my view of these things. Our minds are part of reality, but there's a great deal of reality outside our minds. Reality contains our world and it may contain many others. We can build new worlds and new parts of reality. We know a little about reality, and we can try to know more. There may be parts of it that we can never know.

Most importantly: Reality exists, independently of us. The truth matters. There are truths about reality, and we can try to find them. Even in an age of multiple realities, I still believe in objective reality.

Part 1

---

# **VIRTUAL WORLDS**

## Chapter 1

# Is this the real life?

IN THE OPENING LINES OF THE 1975 HIT “BOHEMIAN RHAPSODY” by the British rock group Queen, lead singer Freddie Mercury sings in five-part harmony:

*Is this the real life?*

*Is this just fantasy?*

These questions have a history. Three of the great ancient traditions of philosophy—those of China, Greece, and India—all ask versions of Mercury’s questions.

Their questions involve alternative versions of reality. Is this real life, or is it just a dream? Is this real life, or is it just an illusion? Is this real life, or is it just a shadow of reality?

Today we might ask: Is this real life, or is it virtual reality? We can think of dreams, illusions, and shadows as ancient counterparts of virtual worlds—minus the computer, which would not be invented for two millennia.

With or without the computer, these scenarios raise some of the deepest questions in philosophy. We can use them to introduce these questions and to guide our thinking about virtual worlds.

### **Zhuangzi’s butterfly dream**

The ancient Chinese philosopher Zhuangzi (also known as Zhuang Zhou or Chuang Tzu) lived around 300 BCE and was a central figure in the Daoist

tradition. He recounts this famous parable: “Zhuangzi Dreams of Being a Butterfly.”

Once Zhuangzi dreamt he was a butterfly, a butterfly flitting and fluttering around, happy with himself and doing as he pleased. He didn't know he was Zhuangzi. Suddenly he woke up and there he was, solid and unmistakably Zhuangzi. But he didn't know if he was Zhuangzi who had dreamt he was a butterfly, or a butterfly dreaming he was Zhuangzi.



*Figure 1* Zhuangzi's butterfly dream. Was he Zhuangzi who dreamt he was a butterfly, or a butterfly dreaming he was Zhuangzi?

Zhuangzi can't be sure that the life he's experiencing as Zhuangzi is real. Maybe the butterfly was real, and Zhuangzi is a dream.

A dream world is a sort of virtual world without a computer. So Zhuangzi's hypothesis that he is in a dream world right now is a computer-free version of the hypothesis that he's in a virtual world right now.

The plot of the Wachowski sisters' 1999 movie *The Matrix* provides a nice parallel. The main character, Neo, lives an ordinary life until he takes a

red pill and wakes up in another world, where he's told that the world he knew was a simulation. If Neo had thought as deeply as Zhuangzi, he might have wondered, "Maybe my old life was the reality, and my new life is the simulation"—a perfectly reasonable thought. While his old world was a world of drudgery, his new world is a world of battles and adventure, where he's treated as a savior. Maybe the red pill knocked him out just long enough for him to be hooked up to this exciting simulation.

On one interpretation, Zhuangzi's butterfly dream raises a question about knowledge: How do any of us know we aren't dreaming right now? This is a cousin of the question raised in the introduction: How do any of us know we aren't in a virtual world right now? These questions lead to a more basic question: How do we know anything we experience is real?

## **Narada's transformation**

Ancient Indian philosophers in the Hindu tradition were gripped by issues of illusion and reality. A central motif appears in the folk tale of the sage Narada's transformation. In one version of the story, Narada says to the god Vishnu, "I have conquered illusion." Vishnu promises to show Narada the true power of illusion (or *Maya*). Narada wakes up as a woman, Sushila, with no memory of what came before. Sushila marries a king, becomes pregnant, and eventually has eight sons and many grandsons. One day, an enemy attacks, and all her sons and grandsons are killed. As the queen grieves, Vishnu appears and says, "Why are you so sad? This is just an illusion." Narada finds himself back in his original body only a moment after the original conversation. He concludes that his whole life is an illusion, just like his life as Sushila.



Figure 2 Vishnu oversees Narada's transformation into Sushila, in the style of *Rick and Morty*.

Narada's life as Sushila is akin to life in a virtual world—a simulation with Vishnu acting as the simulator. As a simulator, Vishnu is in effect suggesting that Narada's ordinary world is a virtual world too.

Narada's transformation is echoed in an episode of the animated TV series *Rick and Morty*, which chronicles the interdimensional adventures of a powerful scientist, Rick, and his grandson Morty. Morty puts on a virtual reality helmet to play a video game titled *Roy: A Life Well Lived*. (It would be even better if Morty had played *Sue: A Life Well Lived*, but you can't have everything.) Morty lives Roy's entire fifty-five-year life: childhood, football star, carpet salesman, cancer patient, death. When he emerges from the game a moment later as Morty, his grandfather berates him for having made the wrong life decisions in the simulation. This is a recurring theme in the series. Its characters are in apparently normal situations that turn out to be simulations and are often led to ask whether their current reality might be a simulation, too.

Narada's transformation raises deep questions about reality. Is Narada's life as Sushila real, or is it an illusion? Vishnu says it is an illusion, but this

is far from obvious. We can raise an analogous question about virtual worlds, including the world of *Roy: A Life Well Lived*. Are these worlds real or illusory? An even more pressing question looms. Vishnu says that our ordinary lives are as illusory as Narada's transformed life. Is our own world real or an illusion?

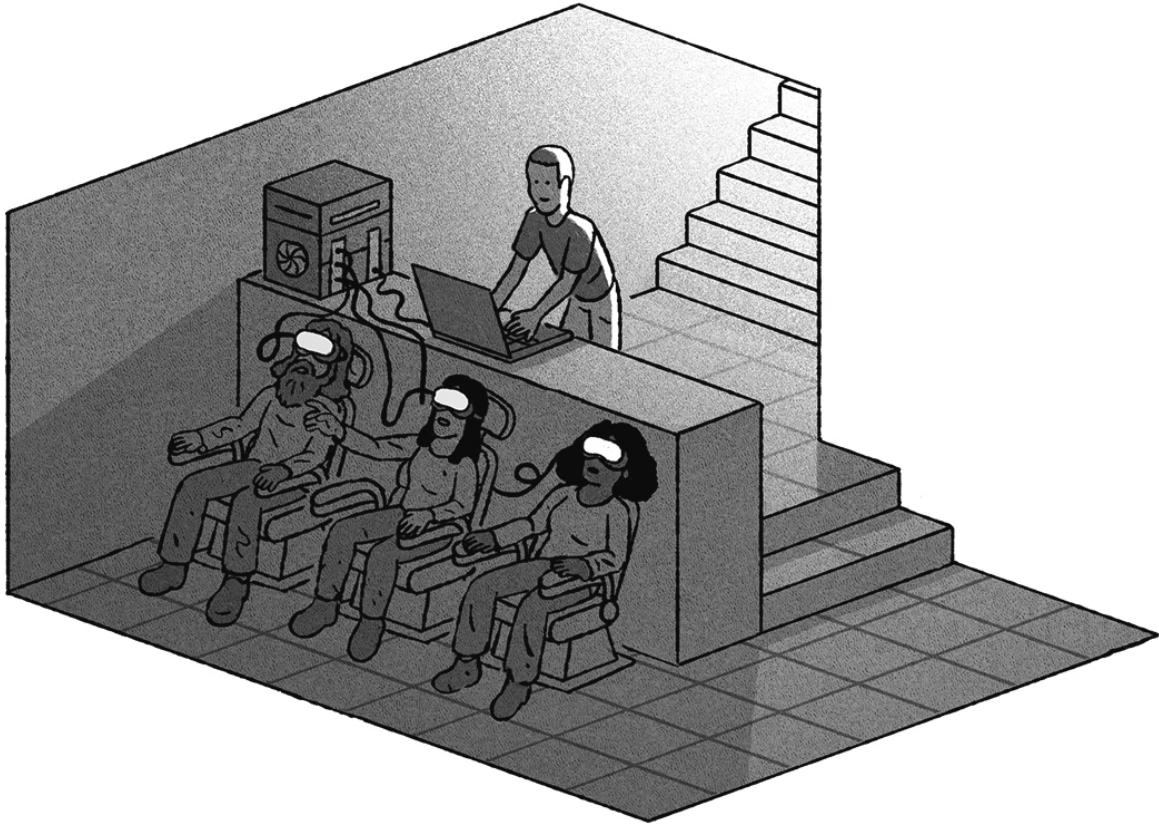
## **Plato's cave**

Around the same time as Zhuangzi, the ancient Greek philosopher Plato put forward his allegory of the cave. In his extended dialogue, the *Republic*, he tells the story of humans who are chained up in a cave, seeing only shadows cast on a wall by puppets that imitate things in the world of sunlight outside. The shadows are all the cave people know, so they take them to be reality. One day, one of them escapes and discovers the glories of the real world outside the cave. Eventually he reenters the cave and tells stories of that world, but no one believes him.

Plato's prisoners watching shadows call to mind viewers in a movie theater. It's as if the prisoners had never watched anything but movies—or, to update the technology, had watched only movies on a virtual reality headset. A 2016 mobile technology conference produced a famous photograph of Facebook chief executive Mark Zuckerberg walking down the aisle past the conference audience. The members of the audience are all wearing virtual reality headsets in the darkened hall, apparently unaware of Zuckerberg as he strides by. It's a contemporary illustration of Plato's cave.

Plato uses his allegory for many purposes. He's suggesting that our own imperfect reality is something like the cave. He's also using it to help us think about what sort of lives we want to live. In a key passage, Plato's spokesman, Socrates, raises the question of whether we should prefer life inside or outside the cave.





*Figure 3 Plato's cave in the 21st century.*

SOCRATES: Do you think the one who had gotten out of the cave would still envy those within the cave and would want to compete with them who are esteemed and who have power? Or would not he much rather wish for the condition that Homer speaks of, namely “to live on the land [above ground] as the paid menial of another destitute peasant”? Wouldn't he prefer to put up with absolutely anything else rather than associate with those opinions that hold in the cave and be that kind of human being?

GLAUCON: I think that he would prefer to endure everything rather than be that kind of human being.

The allegory of the cave raises deep questions about value: that is, about good and bad, or at least about better and worse. Which is better, life inside the cave or life outside the cave? Plato's answer is clear: Life outside the cave, even life as a menial laborer, is vastly better than life inside it. We can ask the same question about virtual worlds. Which is better, life in a virtual

world or life outside it? This leads us to a more fundamental question: What does it mean to live a good life?

### **Three questions**

In one traditional picture, philosophy is the study of *knowledge* (How do we know about the world?), *reality* (What is the nature of the world?), and *value* (What is the difference between good and bad?).

Our three stories raise questions in each of these domains. Knowledge: *How can Zhuangzi know whether or not he's dreaming?* Reality: *Is Narada's transformation real or illusory?* Value: *Can one lead a good life in Plato's cave?*

When we transpose our three stories from their original realms of dreams, transformations, and shadows into the realm of virtuality, they raise three key questions about virtual worlds.

The first question, raised by Zhuangzi's butterfly dream, concerns knowledge. I'll call it the Knowledge Question. *Can we know whether or not we're in a virtual world?*

The second question, raised by Narada's transformation, concerns reality. I'll call it the Reality Question. *Are virtual worlds real or illusory?*

The third question, raised by Plato's cave, concerns value. I'll call it the Value Question. *Can you lead a good life in a virtual world?*

These three questions in turn lead us to three more general questions that are at the heart of philosophy: *Can we know anything about the world around us? Is our world real or illusory? What is it to lead a good life?*

Over the course of this book, these questions about knowledge, reality, and value will be at the heart of our exploration of virtual worlds and at the heart of our exploration of philosophy.

### **The Knowledge Question: Can we know whether or not we're in a virtual world?**

In the 1990 movie *Total Recall* (remade with a few changes in 2012), the viewer is never quite sure which parts of the movie take place in a virtual world and which take place in the ordinary world. The main character, a construction worker named Douglas Quaid (played by Arnold

Schwarzenegger) experiences many outlandish adventures on Earth and on Mars. At the movie's end, Quaid looks out over the surface of Mars and begins to wonder (and so do we) whether his adventures took place in the ordinary world or in virtual reality. The movie hints that Quaid may indeed be in a virtual world. Virtual reality technology that implants memories of adventures plays a key role in the plot. Since heroic adventures on Mars are presumably more likely to take place in virtual worlds than in ordinary life, Quaid, if he is reflective, will conclude that he's probably in virtual reality.

What about you? Can you know whether you're in a virtual or a nonvirtual world? Your life may not be as exciting as Quaid's. But the fact that you're reading a book about virtual worlds should give you pause. (The fact that I'm writing one should give me even more pause.) Why? I suspect that as simulation technology develops, simulators may be drawn to simulate people thinking about simulations, perhaps to see how close they come to realizing the truth about their lives. Even if we seem to be leading perfectly ordinary lives, is there any way we could know whether these lives are virtual?

To put my cards on the table: I don't know whether we're in a virtual world or not. I don't think you know, either. In fact, I don't think we can ever know whether or not we're in a virtual world. In principle, we could confirm that we *are* in a virtual world—for example, the simulators could choose to reveal themselves to us and show us how the simulation works. But if we're *not* in a virtual world, we'll never know that for sure.

I'll discuss the reasons for this uncertainty over the next few chapters. The basic reason is spelled out in [chapter 2](#): We can never prove we're not in a computer simulation because any evidence of ordinary reality—whether the grandeur of nature, the antics of your cat, or the behavior of other people—could presumably be simulated.

Over the centuries, many philosophers have offered strategies that could be used to show that we're not in a virtual world. I'll discuss these strategies in [chapter 4](#) and argue that they don't work. Going beyond this, we should take seriously the possibility that we *are* in a virtual world. The Swedish-born philosopher Nick Bostrom has argued on statistical grounds that under certain assumptions, there will be many more simulated people in the universe than nonsimulated people. If that's right, perhaps we should consider it likely that we're in a simulation. I'll argue in [chapter 5](#) for a

somewhat weaker conclusion: All these considerations mean that we can't know we're *not* in a simulation.

This verdict has major consequences for Descartes's problem: How do we know anything about the external world? If we don't know whether or not we're in a virtual world, and if nothing in a virtual world is real, then it looks like we cannot know if anything in the external world is real. And then it looks like we can't know anything at all about the external world.

That's a shocking consequence. We can't know whether Paris is in France? I can't know that I was born in Australia? I can't know that there's a desk in front of me?

Many philosophers try to avoid this shocking consequence by arguing for a positive answer to the Knowledge Question: we *can* know that we're not in a simulation. If we can know that, then we can know something about the external world after all. If I'm right, though, we can't fall back on this comforting position. We can't know that we're not in a simulation. That makes the problem of knowledge of the external world that much harder.

## **The Reality Question: Are virtual worlds real or illusory?**

Whenever virtual reality is discussed, one hears the same refrain. *Simulations are illusions. Virtual worlds aren't real. Virtual objects don't really exist. Virtual reality isn't genuine reality.*

You can find this idea in *The Matrix*. In a waiting room inside the simulation, Neo sees a child apparently bending a spoon with the power of his mind. They engage in conversation:

CHILD: Do not try and bend the spoon. That's impossible.

Instead . . . only try to realize the truth.

NEO: What truth?

CHILD: There is no spoon.

This is presented as a deep truth. *There is no spoon*. The spoon inside the Matrix is not real but a mere illusion. The implication is that everything one experiences in the Matrix is an illusion.

In a commentary on *The Matrix*, the American philosopher Cornel West, who himself played Councillor West of Zion in *The Matrix Reloaded* and *The Matrix Revolutions*, takes this line of thinking a step further. Speaking of awakening from the Matrix, he says “What you think you’re awakening to may in fact be another species of illusion. It’s illusions all the way down.” Here there is an echo of Vishnu: Simulations are illusions, and ordinary reality may be an illusion, too.

The same line of thinking recurs in the TV series *Atlanta*. Three characters are sitting around a pool late at night discussing the simulation hypothesis. Nadine becomes convinced: “We’re all nothing. It’s a simulation, Van. We’re all fake.” She takes for granted that if we’re living in a simulation, we’re not real.

I think these claims are wrong. Here’s what I think: *Simulations are not illusions. Virtual worlds are real. Virtual objects really exist.* In my view, the Matrix child should have said, “Try to realize the truth. There is a spoon—a digital spoon.” Neo’s world is perfectly real. So is Nadine’s world, even if she is in a simulation.

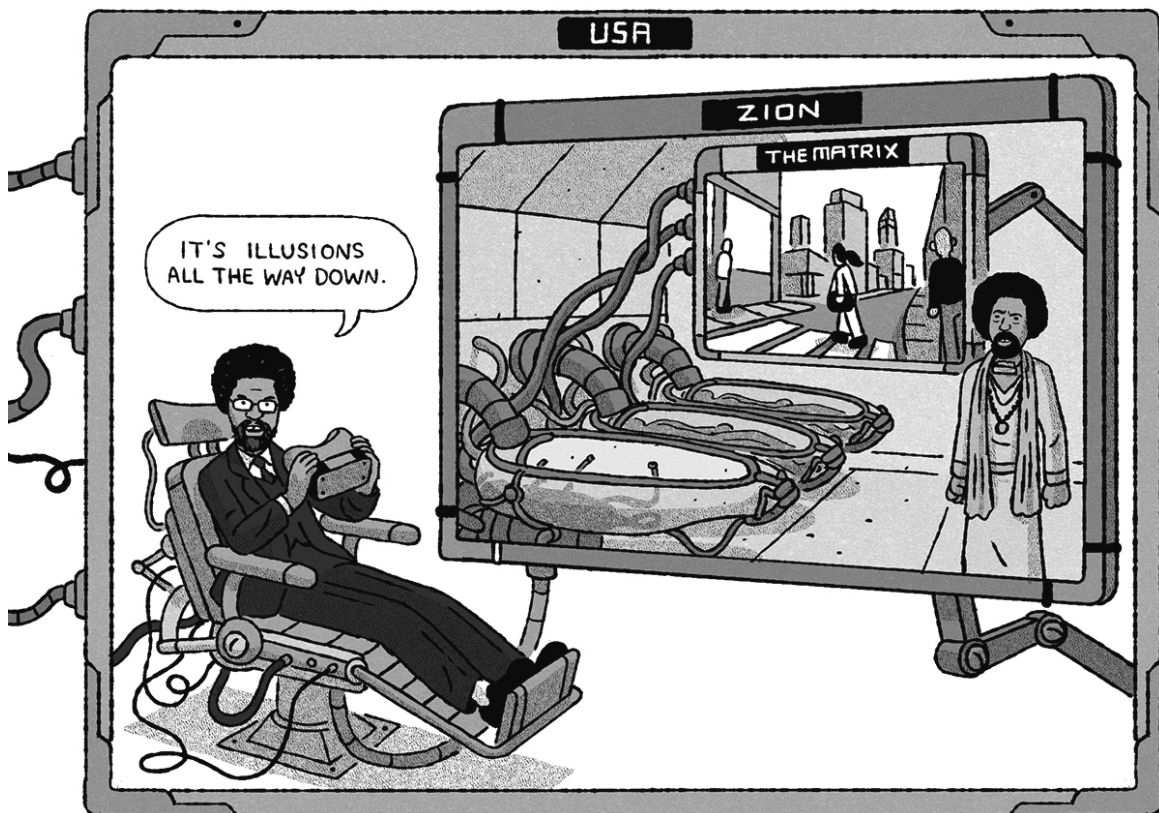


Figure 4 Cornel West, awakening from life as Councillor West of Zion, on illusion and reality.

The same goes for our world. Even if we're in a simulation, our world is real. There are still tables and chairs and people here. There are cities, there are mountains, there are oceans. Of course there may be many illusions in our world. We can be deceived by our senses and by other people. But the ordinary objects around us are real.

What do I mean by "real"? That's complicated—the word "real" doesn't have a single, fixed meaning. In [chapter 6](#), I'll discuss five different criteria for being "real." I'll argue that even if we're in a simulation, the things we perceive meet all these criteria for reality.

What about ordinary virtual reality, experienced through a headset? This can sometimes involve illusion. If you don't know you're in VR and you take the virtual objects to be normal physical objects, you'd be wrong. But I'll argue in [chapter 11](#) that for experienced users of VR, who know they're using VR, there need be no illusion. They're experiencing real virtual objects in virtual reality.

Virtual realities are different from nonvirtual realities. Virtual furniture isn't the same as nonvirtual furniture. Virtual entities are made one way, and nonvirtual entities are made another. Virtual entities are *digital* entities, made of computational and informational processes. More succinctly, they're made of bits. They're perfectly real objects that are grounded in a pattern of bits in a computer. When you interact with a virtual sofa, you're interacting with a pattern of bits. The pattern of bits is entirely real, and so is the virtual sofa.

"Virtual reality" is sometimes taken to mean "fake reality." If I'm right, that's the wrong way to define it. Instead it means something closer to "digital reality." A virtual chair or table is made of digital processes, just as a physical chair or table is made of atoms and quarks and ultimately of quantum processes. The virtual object is different from the nonvirtual one, but both are equally real.

If I'm right, then Narada's life as a woman is not entirely an illusion. Nor is Morty's life as a football star and carpet salesman. The long lives that they experience really happen. Narada really lives a life as Sushila. Morty really lives a life as Roy, albeit in a virtual world.

This view has major consequences for the problem of the external world. If I'm right, then even if I don't know whether or not we're in a simulation, it won't follow that I don't know whether or not the objects around us are real. If we're in a simulation, tables are real (they're patterns

of bits), and if we're not in a simulation, tables are real (they're something else). So either way, tables are real. This offers a new approach to the problem of the external world, one that I will spell out over the course of this book.

## **The Value Question: Can you live a good life in a virtual world?**

In James Gunn's 1954 science-fiction story "The Unhappy Man," a company known as Hedonics, Inc., uses the new "science of happiness" to improve people's lives. People sign a contract to move their life into "sensies," a sort of virtual world where everything is perfect:

We take care of everything; we arrange your life so you never have to worry again. In this age of anxiety, you never have to be anxious. In this age of fear, you never need be afraid. You will always be fed, clothed, housed, and happy. You will love and be loved. Life, for you, will be an unmixed joy.

Gunn's protagonist rejects the offer to hand over his life to Hedonics, Inc.

In his 1974 book *Anarchy, State, and Utopia*, the American philosopher Robert Nozick offers the reader a similar choice:

Suppose there was an experience machine that would give you any experience you desired. Super-duper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life experiences?

Gunn's sensies and Nozick's experience machine are virtual reality devices of a kind. They are asking, "Given the choice, would you spend your life in this kind of engineered reality?"

Like Gunn's protagonist, Nozick says no, and he expects his readers to do the same. His view seems to be that the experience machine is a second-

class reality. Inside the machine, one does not actually do the things one seems to be doing. One is not a genuine autonomous person. For Nozick, life in the experience machine does not have much meaning or value.

Many people would agree with Nozick. In a 2020 survey of professional philosophers, 13 percent of respondents said they would enter the experience machine, and 77 percent said they would not. In broader surveys, most people decline the opportunity, too—although as virtual worlds have become more and more a part of our lives, the number who say they would plug in is increasing.

We can ask the same question of VR more generally. Given the chance to spend your life in VR, would you do it? Could this ever be a reasonable choice? Or we can ask the Value Question directly: Can you lead a valuable and meaningful life in VR?

Ordinary VR differs in some ways from Nozick's experience machine. You know when you're in VR, and many people can enter the same VR environment at once. In addition, ordinary VR is not entirely preprogrammed. In interactive virtual worlds, you make real choices rather than simply living out a script.

Still, in a 2000 article in *Forbes* magazine, Nozick extends his negative assessment of the experience machine to ordinary VR. He says: "even if everybody were plugged into the same virtual reality, that wouldn't be enough to make its contents truly real." He also says of VR: "The pleasures of this may be so great that many people will choose to spend most of their days and nights that way. Meanwhile, the rest of us are likely to find that choice deeply disturbing."

Where VR is concerned, I'll argue (in [chapter 17](#)) that Nozick's answer is the wrong answer. In full-scale VR, users will build their own lives as they choose, genuinely interacting with others around them and leading a meaningful and valuable life. Virtual reality need not be a second-class reality.

Even existing virtual worlds—such as *Second Life*, which has been perhaps the leading virtual world for building a day-to-day life since it was founded in 2003—can be highly valuable. Many people have meaningful relationships and activities in today's virtual worlds, although much that matters is missing: proper bodies, touch, eating and drinking, birth and death, and more. But many of these limitations will be overcome by the



fully immersive VR of the future. In principle, life in VR can be as good or as bad as life in a corresponding nonvirtual reality.

Many of us already spend a great deal of time in virtual worlds. In the future, we may well face the option of spending more time there, or even of spending most of our lives there. If I'm right, this will be a reasonable choice.

Many would see this as a dystopia. I do not. Certainly virtual worlds can be dystopian, just as the physical world can be, but they won't be dystopian merely because they're virtual. As with most technologies, whether VR is good or bad depends entirely on how it's used.

## Central philosophical questions

To recap, our three main questions about virtual worlds are the following. The Reality Question: *Are virtual worlds real?* (My answer: yes.) The Knowledge Question: *Can we know whether or not we're in a virtual world?* (My answer: no.) The Value Question: *Can you lead a good life in a virtual world?* (My answer: yes.)

The Reality Question, the Knowledge Question, and the Value Question match up with three of the central divisions of philosophy.

- (1) *Metaphysics*, the study of reality. Metaphysics asks questions like "What is the nature of reality?"
- (2) *Epistemology*, the study of knowledge. Epistemology asks questions like "How can we know about the world?"
- (3) *Value theory*, the study of values. Value theory asks questions like "What is the difference between good and bad?"

Or, to simplify: *What is this?* That's metaphysics. *How do you know?* That's epistemology. *Is it any good?* That's value theory.

When we ask the Reality Question, the Knowledge Question, and the Value Question, we're doing the metaphysics, epistemology, and value theory of virtual worlds.

Other philosophical questions we'll ask about virtual worlds include:

The Mind Question: *What is the place of minds in virtual worlds?*

The God Question: *If we're in a simulation, is there a god?*

The Ethics Question: *How should we act in a virtual world?*

The Politics Question: *How should we build a virtual society?*

The Science Question: *Is the simulation hypothesis a scientific hypothesis?*

The Language Question: *What is the meaning of language in a virtual world?*

Like our three main questions, these six further questions each correspond to an area of philosophy: the philosophy of mind, the philosophy of religion, ethics, political philosophy, the philosophy of science, and the philosophy of language.

The traditional questions in each of these areas are more general: What is the place of minds in reality? Is there a God? How should we treat other people? How should society be organized? What does science tell us about reality? What is the meaning of language?

In addressing the questions about virtual worlds, I'll do my best to connect them to these bigger questions, too. That way, our answers will not just help us come to grips with the role of virtual worlds in our lives. They'll also help us to get clearer on reality itself.

## **Answering philosophical questions**

Philosophers are good at asking questions. We're less good at answering them. In 2020, my colleague David Bourget and I conducted a survey of around two thousand professional philosophers on one hundred central philosophical questions. To no one's surprise, we found large disagreement on the answers to almost all of them.

Every now and then a philosopher answers a question. Isaac Newton considered himself a philosopher. He worked on philosophical questions about space and time. He figured out how to answer some of them. As a result the new science of physics emerged. Something similar happened later with economics, sociology, psychology, modern logic, formal semantics, and more. All were founded or cofounded by philosophers who got clear enough on some central questions to help spin off a new discipline.

In effect, philosophy is an incubator for other disciplines. When philosophers figure out a method for rigorously addressing a philosophical question, we spin that method off and call it a new field. Because philosophy has been so successful at this over the centuries, what's now left in philosophy is a basket of hard questions that people are still figuring out. That's why philosophers disagree as much as they do.

Still, we can at least pose the questions and try our best to answer them. Occasionally a question is ready to be answered, and we'll get lucky. If we don't answer it, there's often value in the attempt. At the least, posing a question and exploring potential answers can lead us to understand the subject matter better. Others can build on that understanding, and eventually the question might be answered properly.

In this book, I'll try to answer some of the questions I've posed. I can't expect you to agree with all of my answers. Still, I hope you might find understanding in the attempt. With luck, there will be something here that someone can build on. One way or another, we can hope that some of these questions about virtual worlds will eventually migrate from philosophy to a new discipline of their own.

## Chapter 2

# What is the simulation hypothesis?

**T**HE ANTIKYTHERA MECHANISM WAS FOUND IN A SHIPWRECK off the coast of the Greek island of Antikythera in 1901. It dates from two thousand years earlier. The mechanism is a bronze device that was originally mounted in a wooden box about 13 inches across. Superficially, it resembles a clock, with a complex system of 30 or more gears that once drove pointers and dials on the front and the back. Through painstaking analysis over the last century, researchers have discovered that the pointers simulate the day-by-day positions of the Sun and Moon in the zodiac according to the theories of the astronomer Hipparchus of Rhodes. Recently, mathematical analysis of surviving text and gear fragments has provided strong evidence that the system simulated the five known planets as well. It appears that the Antikythera mechanism is an attempt to simulate the solar system. It is the first known cosmic simulation.

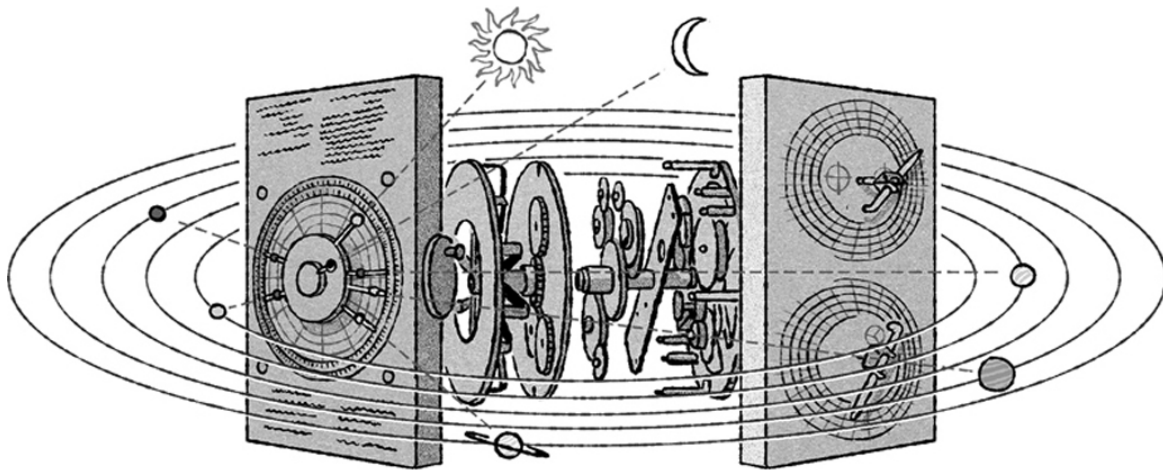


Figure 5 A reconstruction of the Antikythera mechanism, which simulated the position of the Sun and the Moon and probably the five known planets.

The Antikythera mechanism is a *mechanical simulation*. In a mechanical simulation, the positions of components reflect the positions of the entities they're simulating. In the Antikythera, the motion of the gears is intended to reflect the motion of the Sun and Moon against the stars. One could use it to predict a solar eclipse years in the future.

Mechanical simulations are still used from time to time. One prominent example is a mechanical simulation of the San Francisco Bay and its environs, erected in a giant warehouse taking up more than an acre just outside San Francisco. It's a scale model, with enormous amounts of water moved by hydraulic mechanisms to simulate tides, currents, and other forces. It was built to test whether a plan for building dams on the bay would work. The mechanical simulation showed that it wouldn't, and the dams were never built.

Mechanical simulations of highly complex systems are difficult to build, and the art and science of simulation didn't flourish until the start of the computer age in the mid-20th century. In the celebrated code-breaking unit in Bletchley Park (depicted in the film *The Imitation Game*), the British mathematician Alan Turing and other researchers built some of the first computers in order to simulate and analyze German code systems. After the war, the mathematical physicists Stanislaw Ulam and John von Neumann used the ENIAC computer to simulate the behavior of neutrons in a nuclear explosion.

These models were among the first computer simulations. Whereas a mechanical simulation is driven by physical mechanisms, a computer

simulation is driven by algorithms. Instead of using pointers and gears to reflect the positions of the planets, a modern computer simulation uses patterns of bits. An algorithmic simulation of the observed laws of planetary motion makes sure that the bits evolve in a way that reflects the positions of the planets. Using this method, we now have accurate simulations of the solar system allowing us to predict the position of Mars with uncanny precision.

Computer simulations are ubiquitous in science and engineering. In physics and chemistry, we have simulations of atoms and molecules. In biology, we have simulations of cells and organisms. In neuroscience, we have simulations of neural networks. In engineering, we have simulations of cars, planes, bridges, and buildings. In planetary science, we have simulations of Earth's climate over many decades. In cosmology, we have simulations of the known universe as a whole.

In the social sphere, there are many computer simulations of human behavior. As early as 1955, Daniel Gerlough completed a PhD thesis on computer simulation of freeway traffic. In 1959, the Simulmatics Corporation was founded to simulate and predict how a political campaign's messaging would affect various groups of voters. It was said that this effort had a significant effect on the 1960 US presidential election. The claim may have been overblown, but since then, social and political simulations have become mainstream. Advertising companies, political consultants, social media companies, and social scientists build models and run simulations of human populations as a matter of course.

Simulation technology is improving fast, but it's far from perfect. A simulation usually concentrates on a certain level. A population-level simulation approximates human behavior with simple psychological models, but it doesn't usually try to simulate the neural networks that underlie the psychology. A hot topic in the science of simulation involves multiscale simulations, which are increasingly able to simulate systems at a number of levels simultaneously, but there are limits. There are no useful simulations of human behavior that also simulate the atoms within the human brain. Most simulations give at best a rough approximation of the behavior of the systems they simulate.

The same goes for simulations of the whole universe. To date, most cosmic simulations focus on the development of galaxies, typically laying a mesh over an area of the cosmos that divides it into huge units (or cells).

The simulation indicates how these cells evolve and interact over time. In some systems, the size of the mesh is flexible, so that cells can become smaller in certain areas for a more fine-grained analysis. But it is rare for a cosmic simulation to descend to the level of simulating individual stars, let alone planets or organisms on those planets.

Within the next century, however, we may construct reasonably accurate simulations of human brains and behavior. Sometime after that, we might have plausible simulations of a whole human society. Eventually we might simulate a solar system or even a universe, from the level of atoms to the level of the cosmos. In such a system, there will be bits corresponding to every entity in the universe being simulated.

Once we have fine-grained simulations of all the activity in a human brain, we'll have to take seriously the idea that the simulated brains are themselves conscious and intelligent. After all, a perfect simulation of my brain and body will behave exactly like me. Perhaps it might have its own subjective point of view. Perhaps it will experience an environment exactly like the one I experience. At this point, we're just a step away from entertaining the hypothesis that we're living in a simulation ourselves.

## **Possible worlds and thought experiments**

Some simulations are based on reality, while others are not. In his 1981 book *Simulacra and Simulation*, the French philosopher Jean Baudrillard distinguished four phases of simulation according to how closely they mirror reality. The first phase is *representation*, which is the "reflection of a profound reality." The last phase is a *simulacrum*, which "has no relation to any reality whatsoever." Baudrillard is talking about cultural symbols and not computer simulations, but a distant cousin of his distinction can be used to classify four sorts of computer simulation as well.

Some simulations (akin to Baudrillard's representations) aim to simulate a particular aspect of reality as closely as possible, the way a map represents a territory as closely as possible. A historical simulation of the Big Bang or the Second World War aims to replicate those past events closely. A scientific simulation of water boiling aims to simulate what happens when water really boils.

Some simulations aim to simulate something that *could* happen in reality. A flight simulator usually doesn't aim to simulate a flight that has already happened, but to simulate one that could happen. A military simulation may try to simulate what could happen to the United States if there were a nuclear war.

Some simulations aim to simulate something that *could have* happened but didn't. An evolutionary simulation might simulate what would have happened if a massive asteroid impact hadn't led to the extinction of the dinosaurs. A sporting simulation might simulate what would have happened if the United States hadn't boycotted the 1980 Moscow Olympic Games.

Finally, some simulations (akin to Baudrillard's simulacra) aim to simulate worlds that bear no resemblance to reality. A scientific simulation might simulate a world without gravity. We might try to simulate a universe with seven dimensions of space and time.

As a result, simulations are not just a guide to our actual universe. They are also a guide to the vast cosmos of possible universes. Philosophers call these *possible worlds*.

In the world (that is, the universe) we live in, I became a professional philosopher. There are nearby possible worlds in which I became a professional mathematician. There are much more distant possible worlds in which I became a professional athlete. In the actual world, Hitler became leader of Germany and there was a Second World War. There are possible worlds where Hitler never took over and the Second World War never happened. In the actual world, life developed on Earth. There are possible worlds where the solar system never formed. There are even possible worlds where there was no Big Bang.

Computer simulations can help us to explore all of these possible worlds. A cosmic simulation can simulate a universe in which our own galaxy never formed. An evolutionary simulation can simulate a version of Earth in which humans never evolved. A military simulation can simulate a world in which Hitler never invaded the Soviet Union. Eventually, a personal simulation might simulate what would have happened if I had stayed in mathematics and never moved into philosophy.

Another device for exploring possible worlds is the *thought experiment*, an experiment you carry out simply by thinking. You describe a possible world (or at least part of one) and see what follows. Plato's cave is a thought experiment. He imagines a world where prisoners can see only



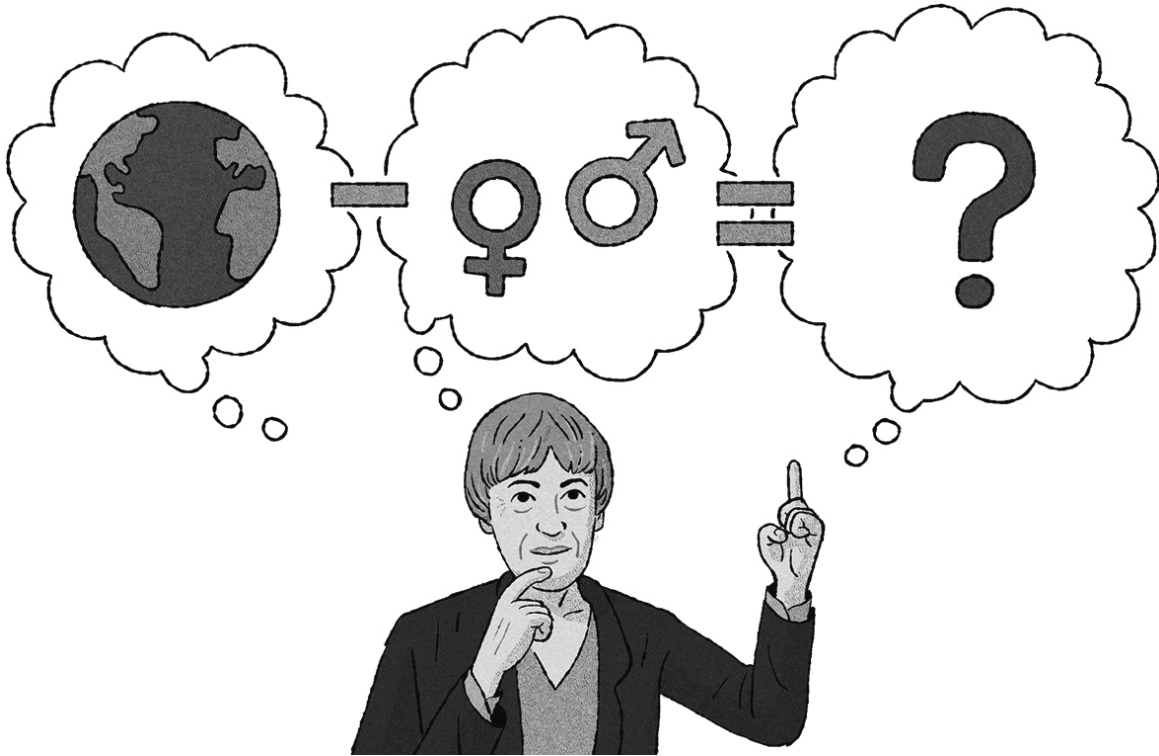
shadows cast on a cave wall, and asks how their lives compare to the lives of people outside the cave. Zhuangzi's butterfly is a thought experiment. Zhuangzi describes a world in which he remembers dreaming about being a butterfly, and asks how he can know he is not a butterfly that dreams he is Zhuangzi.

Thought experiments fuel science fiction. Like philosophy, science fiction explores the world as it could be. Any given science-fiction story is a thought experiment; the author conjures up a scenario and watches what follows. H. G. Wells' *The Time Machine* conjures up a world containing a time machine and then lays out the consequences. Isaac Asimov's stories in *I, Robot* conjure up a world containing intelligent robots, and Asimov then reasons about how we should interact with them.

Ursula Le Guin's classic 1969 novel *The Left Hand of Darkness* describes a possible world where humans on the planet Gethen have no fixed gender. As Le Guin puts it in her 1976 article "Is Gender Necessary?": "I eliminated gender to find out what would be left." In an introduction to the novel, she writes:

If you like you can read [this book], and a lot of other science fiction, as a thought-experiment. Let's say (says Mary Shelley) that a young doctor creates a human being in his laboratory; let's say (says Philip K. Dick) that the Allies lost the Second World War; let's say this or that is such and so, and see what happens. . . . In a story so conceived, the moral complexity proper to the modern novel need not be sacrificed, nor is there any built-in dead end; thought and intuition can move freely within bounds set only by the terms of the experiment, which may be very large indeed.

Thought experiments yield many insights. Le Guin's thought experiment gives us insight into a possibility: It tells us something about gender as it could be. Robert Nozick's thought experiment about the experience machine gives us insight into value: It helps clarify what is valuable to us and what isn't. Zhuangzi's butterfly dream gives us insight into knowledge: What can we know, and what can't we know?



*Figure 6* Ursula Le Guin's thought experiment: "I eliminated gender to find out what would be left."

Thought experiments can stretch the boundaries of some concepts (time and intelligence) and help delimit the boundaries of others (knowledge and value). By exploring these boundaries, they teach us something about the very nature of time, or about what it is to know something.

Thought experiments can be far-fetched, but they often teach us something about reality. Le Guin says that in writing about gender she is "describing certain aspects of psychological reality in the novelist's way, which is by inventing elaborately circumstantial lies." Le Guin's Gethenians may not exist, but aspects of their nature may resonate with the lived experience of many people, including some nonbinary people. Asimov's exploration of artificial intelligence in robots can advise us about how to interact with real AI systems once they're developed. Plato's cave helps us to analyze the complex relation between appearance and reality. This is part of why thought experiments are so central in philosophy, in science, and in literature.

## **Simulations in science fiction**

One especially powerful thought experiment in both science fiction and philosophy is the idea of a simulated universe. What if our universe is a simulation? What follows?

James Gunn's 1955 story "The Naked Sky" was a sequel to the story about Hedonics, Inc. described in [chapter 1](#). Both were later included in his 1961 novel *The Joy Makers*. After apparently destroying the Hedonic Council's dream machines ("In great gobs of blue, the sky began to melt"), the characters wonder whether they're still in a machine or in reality.

How could they be sure that this was reality, not another wish-fulfillment dream from the Council-mech? How could they be sure that they had really conquered it and were not just living an illusion in a watery cell? The answer was: they could never be sure.

Gunn's passage is a contender for the first explicit statement of the simulation hypothesis: the hypothesis that we're living in a computer simulation. Admittedly, computers were new at the time, and Gunn's machines are not explicitly described as computer simulations. His "sensies," in the first story, are akin to highly immersive movies, which in later stories become perfectly convincing "realies." Computer simulations play a small role in Arthur C. Clarke's 1956 novel *The City and the Stars*, but the simulation hypothesis is not entertained there.

The two ideas—computer simulation and the simulation hypothesis—may have come together for the first time in David Duncan's obscure but sophisticated 1960 short story "The Immortals." Roger Staghorn devises a computer-simulation system, Humanac, to predict the future consequences of hypothetical events. He and a colleague, Dr. Peccary, enter the simulation and interact with people predicted to live one hundred years in the future. They have adventures and escape by the skin of their teeth. Back in the ordinary world, they turn off the simulation. The story ends:

"I can't help wondering," mused Staghorn, "of whose computer we're a part right now—slight factors in the chain of causation that started God knows when and will end . . ."

"When someone pulls the switch," said Dr. Peccary.

The deepest development of the computer simulation idea in these early years is the novel *Simulacron-3* (also known as *Counterfeit World*),

published in 1964 by Daniel F. Galouye. This complex work of simulated worlds within simulated worlds was adapted by the great German director Rainer Werner Fassbinder into the German TV production *Welt am Draht* in 1973, later released with English subtitles as the film *World on a Wire*. It appears to be the debut of the simulation hypothesis in film or TV. Fassbinder's film was later remade into the 1999 Hollywood film *The Thirteenth Floor* and is widely credited with inspiring many other films in the simulation genre.

Premiering the same year, *The Matrix*, written and directed by Lana and Lilly Wachowski, remains the best-known depiction of the simulation idea on film. The main character, Neo (in a memorable performance by Keanu Reeves) experiences an ordinary world. He goes to work, he reads books, he hangs out at parties, more or less as we do. He has a few clues that something is strange; his world has a faint green tinge, and he has a perpetual feeling of unease. Tellingly, he has been reading Baudrillard's book *Simulacra and Simulation*. Eventually he takes the red pill and learns that he's been living in a computer simulation all along.

*The Matrix* was partly responsible for my own entry into the simulation arena. The directors and producers of the movie had a significant interest in philosophy, and a number of philosophers were invited to write about philosophical ideas for its official website. I accepted the invitation and in 2003 published an article there called "The Matrix as Metaphysics," all about how the Matrix is not really an illusion. It was an early version of some of the ideas in [part 3](#) of this book.

In "The Matrix as Metaphysics," I introduced my own name for the simulation hypothesis. I called it the "Matrix Hypothesis" and defined it as the hypothesis that I am and always have been in a matrix. I defined a matrix as an artificially designed computer simulation of a world.

In the same year, Nick Bostrom published his important article "Are You Living in a Computer Simulation?," which gave a statistical argument for why we should take the simulation idea seriously. (I'll discuss the argument in [chapter 5](#).) In another 2003 article, Bostrom introduced the name "simulation hypothesis" for the idea. This proved to be a better name than mine; the simulation idea is universal, whereas a movie is ephemeral. In this book I'm following now-standard practice in talking of the simulation hypothesis.

## The simulation hypothesis

What exactly is the simulation hypothesis? Bostrom's version says simply, "We are living in a computer simulation." Mine says, "We are and always have been in an artificially designed computer simulation of a world." I think the two are consistent. My version just makes explicit a couple of things that Bostrom's does not. First, the simulation needs to be lifelong, or at least for as long as we can remember. Being in a simulation since yesterday doesn't count. Second, the simulation needs to have been designed by a simulator. A computer program that popped up randomly without a simulator wouldn't count. Both of these factors are part of the simulation hypothesis as people ordinarily think of it.

What is it to be in a simulation? As I understand this notion, it's all about interacting with the simulation. When you're in a simulation, your sensory inputs come from the simulation, and your motor outputs affect the simulation. You're fully immersed in the simulation through these interactions.

At the start of *The Matrix*, Neo's biological body and brain are in a pod in a nonsimulated world, connected to a simulation somewhere else. In the ordinary spatial sense of "in," Neo's brain is not "in" the simulation. However, all his sensory inputs are coming from the simulation, and his outputs are going there, so he's in a simulation in the sense that matters. After he takes the red pill, his senses respond to the nonsimulated world, so he is no longer in a simulation.

I will use the word *sim* for someone who is in a simulation. There are at least two sorts of sims. First, there are *biosims*: biological beings outside the simulation (in the spatial sense) and connected to it. Neo is a biosim. So is a brain in a vat, connected to a computer. A simulation that includes biosims is an *impure simulation*, since it includes elements (the biosims) that aren't simulated.

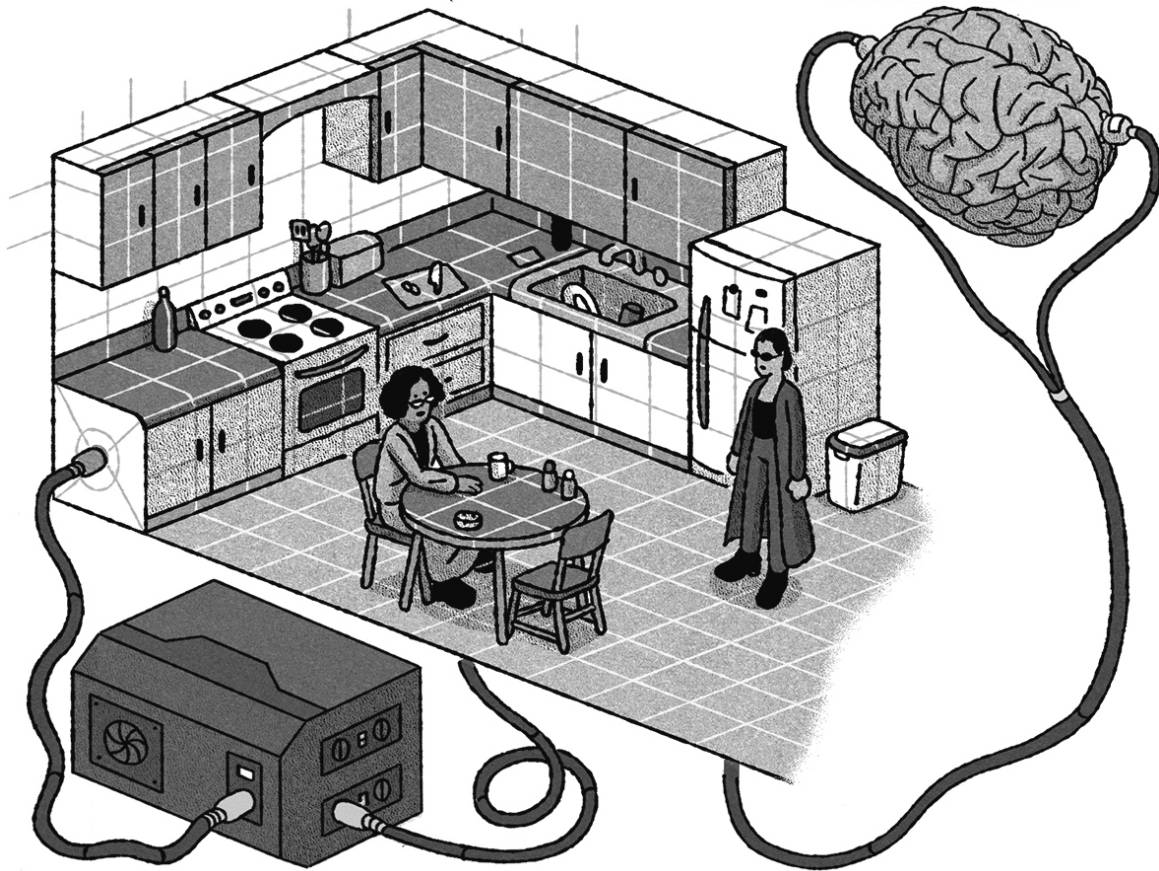


Figure 7 A simulated world with a biosim (controlled by a brain) and a pure sim (controlled by a computer), inspired by Trinity and the Oracle in *The Matrix*.

Second, there are *pure sims*. These are simulated beings who are wholly *inside* the simulation. Most of the people in Galouye’s novel *Simulacron-3* are pure sims. They receive direct sensory inputs from the simulation because they’re part of the simulation. Importantly, their brains are simulated, too. Simulations containing only pure sims may be *pure simulations*—simulations in which everything that happens is simulated.

There can also be *mixed simulations*, which contain both biosims and pure sims. Inside the Matrix, the leading characters Neo and Trinity are biosims, whereas the “machine” characters Agent Smith and the Oracle are pure sims. In the 2021 movie *Free Guy*, the main character, Guy (played by Ryan Reynolds), is a fully digital nonplayer character in a video game, while his in-game partner, Molotov Girl (played by Jodie Comer), is a video game player and designer with an ordinary life outside the game. So Guy is a pure sim, while Molotov Girl is a biosim.

The simulation hypothesis applies equally to pure, impure, and mixed simulations. Occurrences of the simulation idea in science fiction and philosophy are split fairly evenly among them. In the short term, impure simulations will be more common than pure simulations, since we know how to connect people to simulations but we don't yet know how to simulate people. In the long term, pure simulations may well be more common. The supply of brains for impure simulations is not endless, and in any case hooking them up may be tricky. By contrast, pure simulations are easy in the long term. We need only set up the right simulation program and watch it go.

Here's another distinction. The *global*-simulation hypothesis says that the simulation simulates a whole universe in detail. For example, a global simulation of our universe will simulate me, you, everyone on Earth, planet Earth itself, the whole solar system, the galaxy, and everything beyond. The *local* simulation hypothesis says that the simulation simulates only a part of the universe in detail. It might simulate just me, or just New York (see [figure 57](#) in [chapter 24](#)), or just Earth and everyone on it, or just the Milky Way galaxy.

In the short run, local simulations should be easier to create. They require much less computational power. However, a local simulation has to interact with the rest of the world, and that can lead to trouble. In *The Thirteenth Floor*, the simulators simulated only Southern California. When the protagonist tried to drive to Nevada, he encountered signs saying "Road closed." He kept going, and the mountains morphed into thin green lines. That's not a good way to design a convincing simulation. If a local simulation is completely local, it cannot properly simulate interaction with the rest of the world.

To work well, local simulations will have to be flexible. To simulate me, simulators will have to simulate much of my environment. I talk to people elsewhere, see events taking place around the world on TV, and travel often. The people I meet interact with many others in turn. So a good simulation of my local environment will require a fairly detailed simulation of the rest of the world. The simulators may need to fill in more and more details as the simulation runs. For example, a simulation of the far side of the Moon will need revision once spacecraft can photograph it and send pictures back to Earth. There might be some natural stopping points: perhaps simulators

could render Earth in detail and the solar system as they need to, with just a rudimentary simulation of the universe beyond?

Philosophers revel in distinctions. There are many more distinctions we could make. We could distinguish between *temporary* and *permanent* simulations (Do people enter the simulation for a brief period or spend their whole lives there?), *perfect* and *imperfect* simulations (Do we faithfully simulate all the laws of physics, or do we allow approximations and exceptions?), and *pre-programmed* and *open-ended* simulations (Is there a single course of events programmed in advance, or can various things happen depending on initial conditions and what the sims choose?). You can probably think of other distinctions, but we already have enough to go on.

## **Can you prove you're not in a simulation?**

Can you prove you're not in a computer simulation?

You might think you have definitive evidence that you're not. I think that's impossible, because any such evidence could be simulated.

Maybe you think the glorious forest around you proves that your world isn't a simulation. But in principle, the forest could be simulated down to every last detail, and every last bit of light that reaches your eyes from the forest could be simulated, too. Your brain will react exactly as it would in the nonsimulated, ordinary world, so a simulated forest will look exactly like an ordinary one. Can you really prove that you aren't seeing a simulated forest?

Maybe you think your darling cat could never be simulated. But cats are biological systems, and it seems likely that biological mechanisms can be simulated. With good enough technology, a simulation of your cat might be indistinguishable from the original. Do you really know that your cat isn't a simulation?

Maybe you think the creative or loving behavior of the people around you could never be simulated. But what goes for cats goes for people. Human biology could well be simulated. Human behavior is caused by the human brain, and the brain seems to be a complicated machine. Do you really know that a full simulation of the brain could not reproduce all this behavior in detail?



Maybe you think your own body could never be simulated. You feel hunger and pain, you move around, you touch things with your hands, you eat and drink, you're aware of your own weight in a way that seems viscerally real. But as biological systems, bodies can be simulated. If your body is simulated so well that it sends exactly the same signals to your brain, your brain wouldn't be able to tell the difference.

Maybe you think your consciousness could never be simulated. You have subjective experience of the world from a first-person perspective: You experience colors, pains, thoughts, memories. It *feels like something* to be you. No mere simulation of a brain would experience this consciousness!

This issue—the issue of consciousness and whether a simulation could have it—is harder than the others. We'll grapple with it in detail later in the book. For now, we can set the issue of consciousness aside by focusing on impure simulations—that is, *Matrix*-style simulations in which you're a biosim connected to the simulation. Biosims are not themselves simulated. They have ordinary biological brains which will presumably be conscious like ours. Whether you're an ordinary person or a biosim whose brain is in the same state, things will look and feel the same to you.

Pure simulations, in which the people in the simulated world are all simulated themselves, raise the issue of whether simulated beings can be conscious. If we could prove that simulated beings couldn't be conscious, we could prove that we're not in a pure simulation (at least, given that we're sure we're conscious). In [chapter 15](#), I'll argue that simulated beings could be conscious. If a simulated brain precisely mirrors a biological brain, the conscious experience will be the same. If that's right, then just as we can never prove we're not in an impure simulation, we can also never prove that we're not in a pure simulation.

## **Can you prove you *are* in a simulation?**

I've argued that we can never prove we're not in a simulation. What about the other way around? Could we prove we *are* in a simulation?

In *The Matrix*, Neo realized he'd been living in a simulation when he took the red pill and woke up in a different reality. As I've noted, he shouldn't have been so sure. For all he knows, his old world was nonsimulated and the red pill plunged him into a simulation.

Still, we certainly could get very strong evidence that we're in a simulation. The simulators could lift the Sydney Harbour Bridge into the air and turn it upside down. They could show us the source code of the simulation. They could show us private episodes from our past, along with the simulation technology that produced them. They could show me a film of my brain hooked up to wires in the next reality up, with an associated readout of my thoughts and feelings. They could give me control of the simulation, so that I could move mountains in the world around me just by pressing some buttons.

Even this evidence would fall short of absolute proof that we're in a simulation. Maybe the world we're in is a nonsimulated magic world, like the Harry Potter world, in which all-powerful wizards are using their powers to convince us we're in a simulation. Maybe most of my life has been nonsimulated but simulators have put me into a temporary simulated duplicate to fool me. Or maybe I'm having a drug-induced hallucination. Still, I think that if I got evidence like this, I would probably be convinced that I am in a simulation.

## **Is the simulation hypothesis a scientific hypothesis?**

Sometimes people treat the simulation hypothesis as a scientific hypothesis—one that's testable in principle by observation or experiment. Might there be scientific evidence that we're in a simulation?

A 2012 article by physicists Silas Beane, Zohreh Davoudi, and Martin Savage argues that in principle we could someday get scientific evidence for the simulation hypothesis. The basic idea is that a simulation of our universe may well cut some corners by making approximations, and those approximations may show up in the evidence. The authors produce a mathematical analysis of how certain physical approximations using a "hypercubic spacetime" lattice would deviate from standard physics in a testable way. If our simulators used lattice-spacing of a certain size, a distinctive pattern among high-energy cosmic rays would result. The authors suggest that this provides a possible way of testing the simulation hypothesis in the future, though we do not have such evidence as things stand.

This potential evidence depends on the simulation's being *imperfect*. The same goes for the potential evidence discussed in the two preceding sections. Red pills, communication with simulators, and approximations are imperfections of a sort—that is, they're points at which the simulation deviates from the laws of the world it's simulating. In *The Matrix*, *déjà vu* experiences such as a black cat crossing one's path twice are said to arise from glitches in the program. A perfect simulation won't have glitches like this.

A perfect simulation can be defined as one that precisely mirrors the world it's simulating. If the world it's simulating obeys strict physical laws, a perfect simulation will simulate those laws precisely and will never deviate from them. Red pills, communication with simulators, and approximations are ruled out.

It's arguable that a digital computer could never perfectly simulate the continuous laws of physics, which involve precise quantities on a continuum. Still, a digital simulation should be able to approximate the known laws of physics to any degree of precision. And at least in principle there could be perfect simulation of known laws by an analog computer (perhaps an analog quantum computer) which deals in continuous quantities.

If we're in a perfect simulation, it's hard to see how we could ever get evidence of that fact. Our evidence in the simulation will always correspond precisely to evidence in the unsimulated world.

It's just as hard to get evidence that we're *not* in a perfect simulation. As before, any such evidence could in principle be simulated. In a perfect simulation, we would get simulations of the same evidence. At least if we assume that a simulated brain would have the same conscious experience as the brain it's simulating, then there will be no way from the inside to tell the difference between a nonsimulated universe and a perfect simulation of it.

Every now and then, an article appears in the popular press claiming that scientists have proved we're not in a simulation. One example from 2017 stemmed from publication of a research article in *Science Advances* arguing that classical computers cannot efficiently simulate quantum processes. The authors, physicists Zohar Ringel and Dmitry Kovrizhin, did not say that this rules out the simulation hypothesis, but some journalists used their article to draw that conclusion. Of course, the mere fact that classical computers cannot efficiently simulate our universe is no proof that

we're not in a simulation. As the computer scientist Scott Aaronson pointed out, to get around the problem we need only suppose that the simulation is using a quantum computer. We could even suppose that the simulation is simulating quantum processes using a classical computer running slowly and inefficiently. From the inside, we couldn't tell the difference.

Sometimes people say that no universe can contain a perfect simulation of itself, since the universe would need a simulation of the simulation, and a simulation of that, and so on, leading to an infinite stack of simulations. Now, such a stack is not obviously impossible. Perhaps an infinite universe could devote a small fraction of its resources to running a (still infinite) simulation of itself. The resulting stack of simulations would be no problem for an infinite universe. Even a finite but expanding universe could run an ongoing simulation of the past that lags a little behind reality.

Were it indeed the case that no universe could simulate itself, that still would not rule out the simulation hypothesis. There's no reason to suppose that the simulated universe and the simulating universe should be exactly the same. If we're in a simulation, the simulating universe may have an entirely different physics from ours and may be much larger than ours. If the simulating universe is infinite, and has infinite resources, simulating a finite universe will be easy.

To sum things up, I would say that in principle we can get evidence for and against various imperfect simulation hypotheses, which will presumably have empirical consequences we can test. So these imperfect simulation hypotheses count as scientific hypotheses. They may not yet be serious scientific hypotheses, since we don't yet have scientific evidence that gives them support, but at least they're testable in principle.

However, we can never get experimental evidence for or against perfect simulation hypotheses. A nonsimulated world and a perfect simulation of it will seem exactly the same. So, according to the testability criterion, the hypothesis that we're in a perfect simulation is not a scientific hypothesis. Instead, we can think of it as a philosophical hypothesis about the nature of our world.

Some hard-nosed scientists and philosophers may hold that because it's untestable, the perfect simulation hypothesis is meaningless. I'll argue in [chapter 4](#) that this is incorrect. In principle, we can construct perfect simulated worlds ourselves, with beings inside them. There will be no way for those beings ever to know that they're in a simulation. The simulation

hypothesis is demonstrably true of those beings. It follows that the hypothesis is meaningful. It may also be true of us, or it may not. Perhaps we will never know the answer to the question, but the hypothesis is either true or false all the same.

What about the original simulation hypothesis, saying that our world is a computer simulation? Is this a scientific hypothesis or a philosophical hypothesis?

The philosopher of science Karl Popper insisted that the hallmark of a scientific hypothesis is that it is *falsifiable*—capable of being proved false using scientific evidence. We've seen that the simulation hypothesis is not falsifiable because any evidence against it could be simulated. So Popper would say that it's not a scientific hypothesis.

Like many philosophers these days, I think Popper's criterion is too strong. There can be scientific hypotheses—for example, about the early universe—that could never be falsified. But I'm happy to say that the simulation hypothesis is not a squarely scientific hypothesis but one that is partly scientific and partly philosophical. Some versions of it are subject to test, while other versions of it are impossible to test. But whether testable or not, the simulation hypothesis is a perfectly meaningful hypothesis about our world.

## **The simulation hypothesis and the virtual-world hypothesis**

What is the relationship between computer simulations and virtual worlds? Recall that a virtual world is an interactive, computer-generated space. Is every virtual world a simulation? Is every simulation a virtual world?

Most virtual worlds found in video games can be regarded as simulations. This is most obvious in games that simulate some physical-world activity, like fishing or flying or playing basketball. These games are closest to Baudrillard's representations. They may not aim for perfect realism, but they try to reflect the real world. More exotic games, like *Space Invaders* and *World of Warcraft*, are closer to Baudrillard's simulacra. They don't purport to reflect the real world, but they're simulations of possible worlds. *Space Invaders* loosely simulates an alien invasion of Earth. *World*

of *Warcraft* simulates a physical environment with monsters, quests, and battles.

Even games like *Tetris* or *Pac-Man*, which don't obviously simulate physical environments, can be regarded as simulations if you squint at them in the right way. *Tetris* can be seen as a simulation of a two- or three-dimensional world with bricks falling from the sky. *Pac-Man* can be seen as a simulation of predators and prey running through a physical maze. Perhaps it's a stretch to see these as simulations; users may not see them this way, and simulation may have been no part of the designer's intentions. But as I'm understanding the simulation hypothesis, it doesn't matter whether users or the designer see the simulation as a simulation. So these virtual worlds still count as simulations for our purposes.

The same reasoning applies to any virtual world. Any virtual world involves a space, which we can in principle interpret as a simulation of a hypothetical physical space. In this broad sense, any virtual world involves a computer simulation.

What about the converse? Strictly speaking, not all computer simulations are virtual worlds. There are noninteractive simulations, like standard simulations of galaxy formation, that do not interact with users at all. Because they aren't interactive, they don't meet the definition of virtual worlds. But the hypothesis that I'm *in* a computer simulation requires that I'm interacting with a computer-generated world through my sensory inputs and motor outputs. This hypothesis is equivalent to the hypothesis that I'm in a virtual world.

As a result, the simulation hypothesis can equivalently be stated as the virtual-world hypothesis: I am in a virtual world.

To flesh out the picture, the simulation hypothesis suggests we're living in a *fully immersive* virtual world. A virtual world is immersive when you experience it all around you as if you were right there, as with today's standard virtual reality headsets. We defined VR in the introduction as an immersive virtual world. A virtual world is *fully* immersive when one is immersed in the virtual world with all of one's senses, experiencing it just as we experience the physical world. Our experience of the world we live in is fully immersive. So if we're in a virtual world at all, we're in fully immersive VR.

The simulation hypothesis is equivalent to the virtual-world hypothesis, but from now on I will mainly use the standard term "simulation

hypothesis.” In the same spirit, I’ll tend to use the word “simulation” for the sort of *Matrix*-style simulated universe relevant to the simulation hypothesis—that is, a lifelong and fully immersive simulated world in which users may not know they’re in a simulation. I’ll tend to use “virtual world” and “virtual reality” for the more down-to-earth virtual environments that users enter knowingly and for limited periods. This includes everything from video games and current VR headsets to extensions of that technology, such as the scenario of *Ready Player One*, in which people regularly hook themselves up to a fully immersive virtual world.

There’s a spectrum of worlds from current virtual worlds to full-scale simulations such as *The Matrix*. All of them count as virtual worlds in the strict sense, and both ends of the spectrum are relevant to my overarching claims, such as *virtual reality is genuine reality*. Still, down-to-earth virtual worlds and simulated universes raise somewhat different issues. In the next few chapters, simulated universes will take center stage.

Part 2

---

**KNOWLEDGE**



## Chapter 3

# Do we know things?

IN THE ANIMATED TV SERIES *BOJACK HORSEMAN*, THERE'S A TV show within the show called *Hollywood Stars and Celebrities: What Do They Know? Do They Know Things? Let's Find Out!* It's essentially a quiz show for the movie stars in the series, who inhabit an alternative reality in which the "Hollywood" sign has lost its final letter. Like much of the series, it's easily turned into philosophy. The Georgetown University philosopher Quill Kukla teaches a course called "BoJack Horseman and Philosophy," whose tag line is "What Do We Know? Do We Know Things? Let's Find Out!" That tag line pretty much sums up the history of epistemology—that is, the theory of knowledge—in Western philosophy.

*What do we know?* Most of us think we know a lot. We know what happened yesterday and what will probably happen tomorrow. We know about our families and our friends. We know some history, some science, and some philosophy. We even know a little about ourselves.

Philosophers have questioned each of these kinds of knowledge. The ancient Greek philosopher Sextus Empiricus (second or third century CE) questioned our knowledge of science. His Indian Buddhist contemporary Nāgārjuna questioned whether we gain knowledge from philosophy. The 11th-century Persian philosopher al-Ghazali questioned our knowledge of what we see and hear. The 18th-century Scottish philosopher David Hume questioned our knowledge of the future. The contemporary American philosophers Grace Helton and Eric Schwitzgebel have respectively questioned whether we know other people's minds and whether we know our own minds.

*Do we know things?* Some philosophers have questioned whether we know anything at all. The ancient skeptic Pyrrho and his followers said that we shouldn't trust any of our perceptions or our beliefs. Trusting them doesn't lead us to knowledge or to happiness, and if we refrain from believing anything, we can be free from worry. Most of us don't follow Pyrrho's advice; we believe things. But do we know those things?

*Let's find out!* To find out whether we know things, we have to figure out what knowledge is and whether we ever have it. And we have to assess the many challenges to our knowledge that philosophers have put forward over the ages.

A common view of knowledge, going back to Plato, is that knowledge is justified, true belief. To know something, you have to *think* it's true (that's belief), you have to be *right* about it (that's truth), and you have to have good *reasons* for believing it (that's justification).

If I falsely believe that Hillary Clinton ran a child sex ring, that's not knowledge. I'm wrong; it's just a false belief. If I guess someone's birthday and I'm right by pure chance, that's not knowledge. I don't have good reasons; it's an unjustified belief. There may be more to knowledge than justified true belief, but most philosophers think these three requirements are at the heart of the story.

Almost everyone agrees that knowledge is something we want. The 16th-century English philosopher of science Francis Bacon said, "Knowledge itself is power." The American president Thomas Jefferson added that knowledge is happiness and security. In her song "The Knowledge," Janet Jackson sang, "What you don't know can hurt you bad. . . . Get the knowledge."

At the same time, knowing things can be hard work. It's easy to go wrong. Our reasons for believing something are rarely as strong as we'd like them to be. As a result, many thinkers have been led to doubt whether we have any knowledge at all.

## **Skepticism about the external world**

In the opening of his 1983 book *Adventures in the Screen Trade*, the screenwriter William Goldman, who wrote *Butch Cassidy and the Sundance Kid* and *All the President's Men*, addresses the question of knowledge with

a declaration: “NOBODY KNOWS ANYTHING.” He’s talking about the movie business. But once again, the answer runs deeper.

Goldman’s famous line is an expression of *skepticism*, which is exactly the view that nobody knows anything. It’s a view with a long history.

In philosophy, a *skeptic* is someone who casts doubt on our beliefs about a certain domain. Goldman was a skeptic about the movie business. He thought our beliefs about how to make successful movies don’t amount to knowledge. A skeptic about the paranormal casts doubt on our beliefs about ghosts and telepathy. A skeptic about the news media casts doubt on beliefs acquired through the news media.

Skepticism about the news media and about the paranormal are examples of *local* skepticism: casting doubt on our beliefs in a specific domain. There are many forms of local skepticism. You could be a skeptic about the future (casting doubt on our beliefs about what will happen tomorrow), or about science (casting doubt on scientific findings), or about other people’s minds (casting doubt on whether we can ever know what other people are thinking).

The most virulent form of skepticism is *global skepticism*: casting doubt on all of our beliefs at once. The global skeptic says that we cannot know anything at all. We may have many beliefs about the world, but none of them amount to knowledge.

Perhaps the most well-known form of skepticism is skepticism about the external world: casting doubt on all of our beliefs about the world around us. This view is often called Cartesian skepticism, after René Descartes, who was its most famous proponent. Strictly speaking, Cartesian skepticism is not full-scale global skepticism, since it’s consistent with our knowing a few things—about logic or about our own minds, for example. But it’s so encompassing that I will count it as a form of global skepticism here.

Refuting Cartesian skepticism about the external world is one of the hardest problems in modern philosophy. Many philosophers have tried to refute it, but no refutation has commanded much of a consensus. In this book (especially [chapters 6, 9, 22, and 24](#)), I will lay out what I see as the best response to the Cartesian skeptic. Perhaps I too will fail, but I hope to gain some wisdom in the attempt.

My ambitions are limited. I’m going to argue that certain Cartesian arguments for global skepticism about the external world fail. I’m not trying

to refute local forms of skepticism, such as skepticism about the news media (though I'll come back to that issue in [chapter 13](#)). My target is the classical Cartesian skeptic who uses a radical hypothesis to cast doubt on all of our beliefs about the external world at once.

## How do you know your senses aren't deceiving you?

In 1641, Descartes published his *Meditations on First Philosophy*. He was trying to build a foundation for everything that we know. To build that foundation, he first had to tear everything down. His demolition crew included three classic arguments—concerning illusions, dreams, and demons—that cast doubt on our knowledge of the external world. These arguments weren't entirely new. Illusions and dreams were standard fare for skeptics in ancient times such as Sextus Empiricus and the Roman orator Cicero, as well as for medieval thinkers such as the 5th-century North African saint Augustine and the Persian philosopher al-Ghazali. We'll see that demons were used by Descartes's contemporaries as well. Nevertheless, Descartes gave these arguments their most influential formulation.

Descartes's first argument was based on illusions. *Our senses have deceived us before. How can we know they aren't deceiving us now?*

Most of us have experienced optical illusions in which appearances are different from reality. We've been fooled by smoke and mirrors. If our senses have deceived us in the past, they may be deceiving us now. So we can't be sure that whatever we observe in the external world is as it seems to be.

Descartes accepted that sensory illusions have limits. No sensory illusion could give people the sense of having an entirely different body or being in an entirely different environment. He wrote: "Yet although the senses occasionally deceive us with respect to objects which are very small or in the distance, there are many other beliefs about which doubt is quite impossible, even though they are derived from the senses—for example, that I am here, sitting by the fire, wearing a winter dressing-gown, holding this piece of paper in my hands, and so on."

A 21st-century reader will say, "Not so fast!" VR researchers regularly talk about "whole-body illusions"—the sort of thing Descartes thought was

impossible. I can see and control a body that isn't my biological body, and I'll sense that it's mine. Inside VR, Descartes could even have the sense that he's sitting by the fire in a dressing gown, holding a piece of paper. VR thereby strengthens Descartes's original argument based on illusions. Technology makes it harder to know that we're not experiencing an illusion right now.

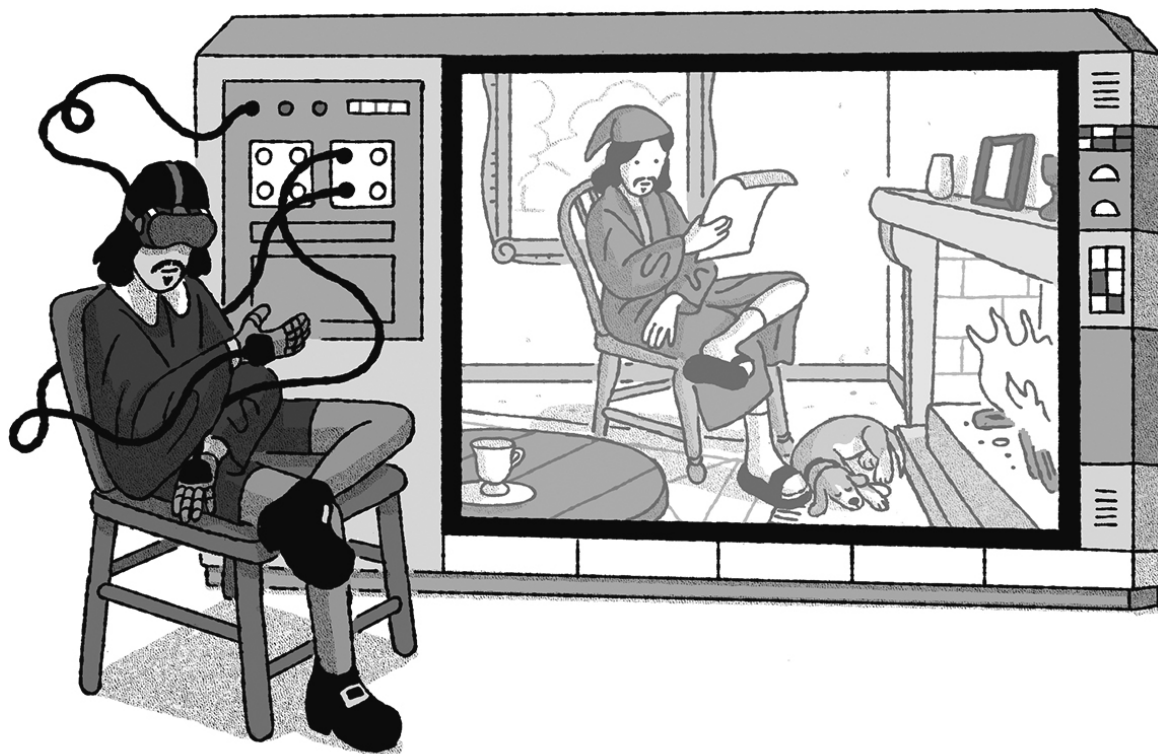


Figure 8 Inside VR, Descartes has the sense that he's sitting by the fire in a dressing gown, holding a piece of paper in his hands.

We can even mount a 21st-century version of Descartes's illusion argument: the argument from virtual reality. *Virtual reality devices have fooled people before. How do you know that a VR device isn't fooling you now?* In principle, everything you're seeing and hearing could be the product of a VR device. Can you really be sure you're not using such a device now?

Admittedly, to be fully convincing, today's technology would have to improve, but eventually we'll have VR contact lenses and other unnoticeable equipment that can handle all the senses. In principle, someone could put you into such an advanced VR device while you're

asleep. In the morning, you wake up in a virtual bed and go about your virtual day. If you've been put in a new virtual environment, like a virtual version of Mars, you'll presumably realize that something's wrong—unless your memories have also been tampered with, which goes a step further than I'm going here. So let's say that the VR environment is like your normal home, or like wherever you are right now. That way, you won't notice anything too strange.

Are you really sure you're not in such a VR device right now? If you're sure—how can you really rule out the VR hypothesis? If you're not sure—then how can you be sure about anything you perceive in the world around you? Can you really be sure that this is a genuine book you're reading, or a genuine chair you're sitting in? Can you really be sure about where you are or about whether what you're seeing is really there?

This VR argument casts doubt on your knowledge of what you're seeing and hearing now, and perhaps on your knowledge of the recent past. But it may not throw everything you know into question. VR per se won't tamper with memories, so your memories of growing up in your hometown won't be threatened. It also won't tamper with your general scientific or cultural knowledge, so your knowledge that Paris is in France will still be safe.

You could try to extend the VR argument to threaten these domains. Maybe a memory-tampering device could get into your brain and change your memories. Maybe a permanent VR device could ensure that all of your memories and your scientific knowledge would derive from VR. These extensions bring us beyond standard VR and into the domain of the simulation hypothesis, which I'll return to soon.

## **How do you know you're not dreaming?**

Descartes's second argument was about dreaming: *Dreams are like reality. How do we know we're not dreaming?*

We usually dream without being aware that we're dreaming. When we're dreaming, we typically think the world of the dream is real. There are occasionally lucid dreams, when people know they're dreaming, but these are the exception. Although most dreams are weirder and less stable than reality, in principle there could be dreams that are indistinguishable from

reality. How do you know you're not having such a dream right now? Perhaps you could pinch yourself or run some experiments. That wouldn't be too convincing, though, as any results you get could in principle come from a dream. And if you can't know you're not dreaming now, it seems that you can't know that anything around you is real.

In the 2010 movie *Inception* (spoiler alert!), the characters fall asleep and enter a dream world, and then enter dream worlds within dream worlds. For most of the movie, the main character, Dominick Cobb (played by Leonardo DiCaprio), knows he is in a dream world and that he's asleep in ordinary reality. But other characters, including Robert Fischer (played by Cillian Murphy), do not know this. Fischer is the one who is dreaming the dream world, and he treats the dream world like reality. At the end of the movie, when Cobb and the other characters seem to have returned to the ordinary world, the question arises: How do they know that this is not yet another dream world? It seems impossible to know for sure.

Descartes thought the dream argument was stronger than the illusion argument. Unlike an optical illusion, a dream could easily convince him that he was in his dressing gown by the fire when in fact he wasn't. Maybe a dream could even give him a different body. But for Descartes, this argument, too, had limits. Dreams never give you something absolutely new. If you dream of a head or a body, it must be based on your having perceived heads and bodies in the real world. Or at least the shapes and colors in the dream must be based on shapes and colors in the real world.

We haven't developed dream technology as thoroughly as we've developed VR technology, so Descartes's dream argument is less affected by technological change than his illusion argument. However, dream science has found some fairly good ways to know you're dreaming. Look at some writing on a page twice: In a dream it will usually change; outside a dream it usually won't. Another relevant piece of science is that we can indeed experience colors we've never experienced in reality. For example, residual images after we see certain colors can give us a shade of "dark yellow" that we can never get through ordinary perception.

Like the illusion argument and the VR argument, the dream argument casts doubt on our current and recent knowledge of the world around us: How do I know that what I'm seeing now, or what I saw a moment ago, is real? Preexisting knowledge is trickier. Dreams can sometimes alter our memories (in a dream, I can remember a different childhood) and our

cultural beliefs (I can dream that the Beatles are still performing together), but they don't usually alter our memories wholesale. One could postulate a lifelong dream in which every element of one's reality comes from a dream, but now we are once again in the science-fiction territory of simulations and the like.

Descartes's illusion argument and dream argument both work best in supporting local skepticism—casting doubt on some of our knowledge of the external world, but not all of it at once. Descartes was not satisfied with this. He was interested in global skepticism: that is, casting doubt on our knowledge of the entire external world all at once. For that, he needed a stronger argument.

## Descartes's evil demon

Descartes's third and most notorious argument is an argument about deception. *An all-powerful being could deceive me completely, by giving me experiences of a world that does not exist. How do I know this isn't happening to me?*

Descartes's original and central deceiver in the first meditation was an all-powerful and all-deceiving God. If God can do anything, surely God has the power to deceive us completely. The deceiver that everyone remembers, though, is Descartes's evil demon. In the original Latin, Descartes talked of a *genium malignum*, which might be translated as “bad genie” (French philosophers typically talk of a “malin génie”), but “evil demon” does the job in English. While a benevolent God might refuse to deceive us, an evil demon would have no such compunctions. Descartes introduces the demon this way:

I will suppose therefore that not God, who is supremely good and the source of all truth, but rather some evil demon of the utmost power and cunning has employed all his energies in order to deceive me. I shall think that the sky, the air, the earth, colours, shapes, sounds and all external things are merely the delusions of dreams which he has devised to ensnare my judgement.

The evil demon is devoted to deception. It feeds you sensations and perceptions, as of an external world, for your whole life. I remember my



growing up in Australia, and these days I seem to be living an enjoyable life in New York City as a professor of philosophy. But if Descartes's evil-demon hypothesis is correct, all of this was based on sensations and perceptions fed to me by the demon. In reality, I have spent my whole life in its lair, where it is manipulating my senses.

The evil-demon thought experiment was not entirely novel. The Columbia University historian of philosophy Christia Mercer has recently charted how the 16th-century Spanish theologian Teresa of Ávila wrote her own meditations in which deceiving demons played a central role. For Teresa, the issue was belief in God, and the demons were trying to deceive her to make her lose her faith. Teresa's book, *The Interior Castle*, was a huge seller in Descartes's time, and he almost certainly read it. Descartes's readers would also have encountered the illusion and dream arguments in the well-known writings of the 16th-century French essayist Michel de Montaigne. So while Descartes's *Meditations* were certainly an advance, he was building on the work of the women and men around him.

Some aspects of the evil demon story are portrayed in the 1998 movie *The Truman Show*. In the movie, Truman Burbank (played by Jim Carrey) is living in a bubble populated by actors. A television producer, Christof (played by Ed Harris), orchestrates the bubble to give him the sense that his is a normal life. Christof is playing the role of the evil demon. Some of Truman's world is real, though. He really has a body, he really lives on Earth, and he really interacts with people. Christof doesn't deceive him about that. The evil demon's victim is like a version of Truman who does not have a body and does not interact with people. Everything the victim experiences is produced by the evil demon.

How do you know that right now you're not being manipulated by an evil demon? It seems that you can't. Maybe there's some suggestion of the evil demon's handiwork—the fact that you're now reading about evil demons, for example; evil demons with a sense of humor might enjoy causing people to think about evil demons. Even without hints like this, it seems impossible to exclude the evil-demon hypothesis entirely. But if you can't know that you're not being manipulated by an evil demon, how can you know that anything is real?

The evil-demon argument calls into question everything you know about the external world. Therein lies its power. As we've seen, ordinary illusions and dreams don't threaten my knowledge of my childhood home

in Australia, or my knowledge that Einstein discovered relativity. The evil demon has been deceiving us our whole lives, so it threatens everything. My experiences of Australia may be a fiction. When I read about Einstein's discoveries, the stories may be made up. So if we cannot rule out the evil-demon hypothesis, we are threatened with global skepticism.

How does the evil demon do its work? Descartes is not clear on the details. Presumably the demon has to keep a complicated model of a fictional world in its head to make sure the subject has matching experiences over time. Every time I return to Australia or visit an old friend, my experience needs to be consistent with previous visits. The demon will also need models for places I've read about and places I'll eventually visit, as well as everything I read about in newspapers or watch on TV. The model will have to be constantly updated. This will be a lot of work, although perhaps the work is nothing for an all-powerful demon.

An especially insidious version of Descartes's evil demon gets inside people's minds and directly tampers with their thoughts. In the modern version, this demon could be an evil neuroscientist. Perhaps the demon manipulates your brain so that you believe that you're in Antarctica. Descartes says that a deceiver might even manipulate his thoughts so that "I too go wrong every time I add two and three." Perhaps the demon can make you believe  $2 + 3 = 6$ , and you will find this completely convincing.

The mind-tampering demon threatens to lead to a sort of *internal-world* skepticism in which you can't even trust your own rationality or reasoning anymore. This sort of evil demon is fascinating, but it's outside the scope of my discussion. I'm concerned here with scenarios in which my external world is manipulated, not scenarios in which my internal world is manipulated directly. I'll return to mind-tampering scenarios in the final chapter of this book.

## **From the evil demon to the simulation hypothesis**

If Descartes's evil demon lives in the computer age, its task is a lot easier. It can simply offload the modeling work into a computer. It can run a computer simulation of the world and connect subjects to the simulation so that they experience the world as it evolves. This is the setup in *The Matrix*,

where godlike machines play the role of the evil demon and a computer simulation takes care of the hard work.

In the 20th century, the American philosopher Hilary Putnam and others updated Descartes's idea with equipment from modern science. The evil demon was replaced by an evil scientist, and the person deceived by the evil demon was replaced by a *brain in a vat*. Like the brains that float in jars in the Steve Martin movie *The Man with Two Brains*, the brain is kept alive with a carefully balanced mix of nutrients. Putnam tells us that the brain's nerve endings are "connected to a super-scientific computer." The computer sends electronic impulses to the brain, bringing about the illusion that everything is normal. The brain experiences a richly detailed and well-populated world, but in fact it is alone in a laboratory.

Putnam's brain-in-a-vat scenario is very much like the scenario in *The Matrix*, except that in the movie full bodies in pods are connected to the computer. Putnam doesn't say much about what the computer is doing (neither does *The Matrix*), but clearly (as in *The Matrix*) it is running a simulation of the world that the brain is experiencing.

In the 21st century, philosophers' focus has gradually shifted from brains in vats to the simulation hypothesis. The simulation idea captures an element at the core of all of the great Cartesian scenarios: The evil demon must do its work by simulating a world. A lifelong dream can be seen as a sort of simulated world. The brain in a vat is connected to a simulation. And so on. Making the simulation a computer simulation helps us to pin down the scenario in more concrete terms without losing anything essential.

The brain-in-a-vat idea is one version of the simulation hypothesis. It involves an impure simulation, in which a brain is connected to the simulation from the outside. The simulation hypothesis also includes other versions, such as pure simulation versions in which the brain is internal to the simulation. Both of these scenarios can be used to mount a Cartesian argument for skepticism.

You might think that the switch from evil demons to brains in vats to simulations is a mere change in packaging, but there is one respect in which the use of modern technology makes the argument more powerful. Because the evil-demon hypothesis is so fanciful, Descartes was reluctant to put too much weight on it. It was important to him that his skeptical concerns be grounded in reasonable doubts that he should take seriously, given his beliefs. He gave more weight to the deceiving-God hypothesis because he

believed in an all-powerful God and thought it reasonable that God would have the power to deceive us. Because this was a realistic hypothesis, it gave him greater reason for doubt.

The simulation hypothesis may once have been a fanciful hypothesis, but it is rapidly becoming a serious hypothesis. Putnam put forward his brain-in-a-vat idea as a piece of science fiction. But since then, simulation and VR technologies have advanced fast, and it isn't hard to see a path to full-scale simulated worlds in which some people could spend a lifetime.

As a result, the simulation hypothesis is more realistic than the evil-demon hypothesis. As the British philosopher Barry Dainton has put it: "The threat posed by simulation scepticism is far more *real* than that posed by its predecessors." Descartes would doubtless have taken today's simulation hypothesis more seriously than his demon hypothesis, for just that reason. We should take it more seriously, too.

## **The master argument for skepticism**

Philosophers love arguments. This is not to say that they love disputes with each other, although many enjoy that, too. In philosophy, an argument is a chain of reasoning that supports a conclusion. I can argue that God exists by laying out some reasons for thinking that God exists and showing how they support my conclusion.

Sometimes arguments are informal. Maybe I try to convince you that we should go to a movie by giving some reasons: We both have spare time, it's a great movie, and it's only playing tonight. I can do the same in philosophy. I can try to convince you that you can't be certain of the world around you by giving some reasons: You've had sensory illusions before, so how do you know you're not having one now? If I do a good job, maybe it will convince you of the conclusion, or at least prompt you to take it seriously.

Sometimes arguments are formal. That may sound intimidating, but formal arguments are often simple. You lay out a number of claims that are *premises*, and then you lay out a *conclusion* that follows from them. Usually the idea is that the premises are plausible enough that people will have some inclination to accept them, and the conclusion drawn from these premises is bold enough to be interesting.

Here's a formal argument for skepticism about the external world.

1. You can't know you're not in a simulation.
  2. If you can't know you're not in a simulation, you can't know anything about the external world.
- 
3. So: You can't know anything about the external world.

Here, the first two claims are the premises, and the third claim is the conclusion. The conclusion follows logically from the premises: If the premises are true, the conclusion has to be true. When the conclusion follows from the premises, philosophers say the argument is *valid*. When in addition the premises are true, the argument is *sound*. Just because an argument is valid, this doesn't mean that the conclusion is true. After all, one or both of the premises could be false. But when an argument is sound, the conclusion has to be true. In the argument above, *if* you accept the two premises, you pretty much have to accept the conclusion.

Bertrand Russell once said, "The point of philosophy is to start with something so simple as not to seem worth stating, and to end with something so paradoxical that no one will believe it." The argument above at least has the potential to meet Russell's ideal. Both premises seem plausible, at least on a moment's reflection, and the conclusion seems surprising. That's one of the things that makes this argument so interesting.

In fact, this argument is so interesting that it, or something like it, is often regarded as the *master argument* for skepticism in recent philosophy. The details can change a bit. For example, we could replace simulations with evil demons or brains in vats, but the basic idea is intact.

Why believe the first premise? I've made an initial case in [chapter 2](#). In a good-enough simulation, the world would look and feel to you exactly as today's world looks and feels to you now. And if a simulation would look and feel the same as reality, it's hard to see how we could know we're in a simulation rather than reality.

Why believe the second premise? Pick anything you think you knew about the external world. You thought you knew that Paris is in France, or that there's a spoon in front of you. But if you're in a simulation, then your beliefs about Paris and the spoon come from the simulation, not from reality. Paris and the spoon are simulated. The world outside the simulation

may be entirely different. There may well be no Paris and no spoon in reality outside the simulation. So to know that Paris is in France or that there's truly a spoon in front of you, you have to rule out the possibility that you're in a simulation.

The reasoning here is a bit like this: If your phone is a knockoff, you don't really have an iPhone. So if you can't know that your phone isn't a knockoff, you can't know that you have an iPhone. In this case, we start from the plausible claim: If you're in a simulation, there's no spoon in front of you. By the same sort of reasoning as in the iPhone case, we get to: If you can't know you're not in a simulation, you can't know there's a spoon in front of you. The same applies to everything in the external world.

Our Reality Question about virtual reality was: *Is virtual reality real or an illusion?* If you answer by saying *Virtual reality is an illusion*, you'll probably accept the second premise. Here's why. Given this answer, you'll also accept *Simulations are illusions*, since simulations are a kind of virtual reality in the broad sense. In fact, you'll probably accept *If you're in a simulation, everything you experience in the external world is illusory*. So if you can't rule out the simulation hypothesis, you can't rule out that everything in the external world is illusory. It seems to follow that you can't know anything about the external world at all.

The conclusion is startling. If you're like most people, you thought you knew a lot of things. You thought you knew that Paris is in France, and you thought you knew what's physically in front of you. But it turns out you don't! The argument applies to more than just objects or cities. It applies to memories of your childhood. If you're in a simulation, so the argument goes, your memories of going to school aren't real, so you don't really know that you went to school. The same goes for pretty much everything you thought you knew about the external world and your life in it.

Strictly speaking, the argument doesn't stop you from knowing a *few* things about the external world. Some things are true as a matter of logic or mathematics. You can know that all dogs are dogs, for example. You can know that if there is one table here and a different table there, there are two tables. But these are all trivialities. To be strictly correct, we could adjust the conclusion to "We can't know anything substantial about the external world."

If we accept the premises, the argument leads us to global skepticism about the external world—that is, the view that we don't know anything

substantial about the external world. Maybe we can still know that two plus two is four, but that's not a huge consolation.

What can we do to avoid the shocking conclusion?

## I think, therefore I am

Descartes himself didn't want to be a skeptic. In fact, he wanted to establish a foundation for all knowledge. So after casting all our knowledge into doubt with his skeptical arguments, he tried to build it back up, piece by piece.

Descartes needed to start with a piece of knowledge he couldn't doubt. He needed to uncover something about reality that would be true even if he was having sensory illusions, even if he was dreaming, even if he was being fooled by an evil demon. He found a candidate: his own existence.

Descartes's famous argument for his own existence, presented most explicitly in his 1637 *Discourse on Method*, went like this: *Cogito, ergo sum*. I think, therefore I am.

Philosophers have interpreted Descartes's celebrated slogan in many different ways. But at least on the surface, it looks like an argument. The premise of the argument (to unpack it a little) is *I am thinking*. The conclusion is *I exist*. As with most arguments, the real work is done by the premise. Once you grant that, the conclusion *I exist* seems to follow as a matter of logic.

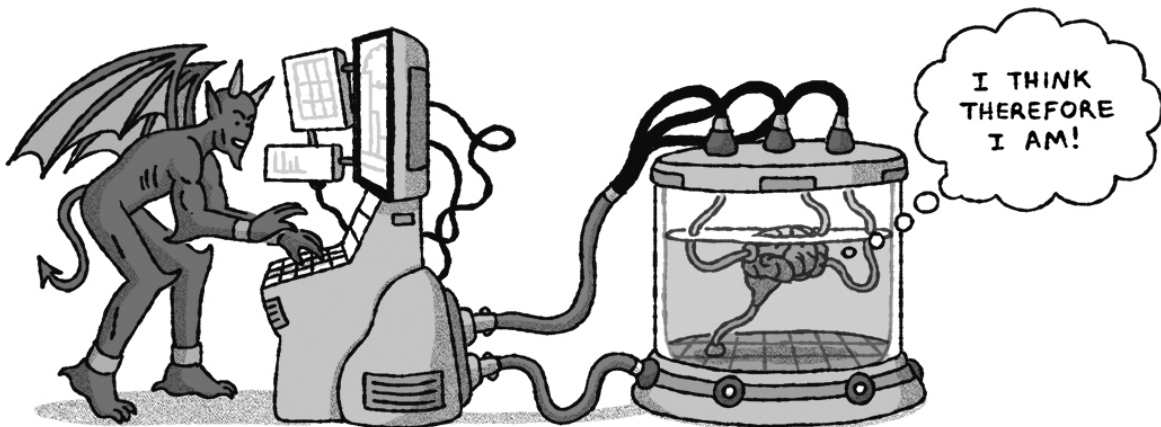


Figure 9 Even if you're a brain in a vat, receiving sensations from an evil demon, you can still reason, "I think, therefore I am."

How does Descartes know he's thinking? For a start, this knowledge does not seem to be undercut by the skeptical arguments. Even if you're in the grip of a sensory illusion, you're still thinking. Even if you're dreaming, you're still thinking. Even if you're being fooled by an evil demon, you're still thinking. Even if you're a brain in a vat, you're still thinking. Even if you're in a simulation, you're still thinking.

More deeply, Descartes reasoned that he could not doubt that he's thinking. Even if he doubted that he was thinking, his doubt was itself a sort of thinking. To doubt that one is thinking is internally inconsistent: The doubting itself shows that the doubt is wrong.

Once Descartes knew he was thinking, it was a small step to knowing his own existence. Where there is thinking, there must be a thinker. So Descartes concludes: *Sum!* I exist!

Plenty of philosophers have tried to poke holes in Descartes's *Cogito, ergo sum*. Some question the *cogito* part. How can Descartes be so sure that he's even managing to doubt? That is, how does he know that he's not a mindless automaton? Others question the step to *sum*. Is it so obvious that thinking requires a thinker? According to the 18th-century German philosopher Georg Lichtenberg, Descartes should have said, "There is thinking, therefore thought exists." That way, he could have known that thoughts exist, but he should not have been so sure about himself.

Still, a lot of people accept Descartes's *Cogito, ergo sum*. It's hard to doubt that I'm thinking. The evil-demon scenario doesn't really call my own mind into doubt, and it's not easy to generate scenarios that do. As a result, even some skeptical philosophers are prepared to say that we do know that we think, and that therefore we do know that we exist.

Speaking for myself, I don't think there's anything special about thinking per se. Descartes could have said, "I feel, therefore I am," or "I see, therefore I am," or "I worry, therefore I am." All of these are claims about his mind that he can be certain of and that aren't threatened by the evil demon. At least, he can be certain about these claims if they're understood as states of consciousness, or subjective experience. If we understand "see" as referring simply to the subjective experience of seeing, then Descartes can be certain he is seeing.

In my view, the best statement of the *cogito* is "I am conscious, therefore I am." Perhaps it's not surprising that I would say this, since thinking about consciousness is my day job. (A writer might say, "I write,



therefore I am.”) But it’s arguable that this is what Descartes really meant. He explicitly defines thought as including everything we’re conscious of, and says that it includes the senses and imagination as well as the intellect and the will.

Some theorists have tried to apply skepticism not only to the external world but also to consciousness itself, suggesting that consciousness could be an illusion. We’ll revisit this view in [chapter 15](#). It’s usually regarded as extreme, but it does demonstrate that in philosophy, everything is open to question.

If we grant *Cogito, ergo sum*, that gives Descartes a foundation. The hard part is what comes next. How do we get from knowledge of ourselves and our own minds to knowledge of the external world?